NATIONAL CENTER FOR EDUCATION STATISTICS

Working Paper Series

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.

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

NATIONAL CENTER FOR EDUCATION STATISTICS

Working Paper Series

Selected Papers on Education Surveys: Papers Presented at the 1996 Meeting of the American Statistical Association

Working Paper No. 97-01 February 1997

Contact: Dan Kasprzyk
Surveys and Cooperative Systems Group
(202) 219-1588
e-mail: daniel_kasprzyk@ed.gov

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

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

Office of Educational Research and Improvement
Marshall S. Smith
Acting Assistant Secretary

National Center for Education Statistics
Pascal D. Forgione, Jr.
Commissioner

Surveys and Cooperative Systems Group
Paul D. Planchon
Associate Commissioner

The National Center for Education Statistics (NCES) is the primary federal entity for collecting, analyzing, and reporting data related to education in the United States and other nations. It fulfills a congressional mandate to collect, collate, analyze, and report full and complete statistics on the condition of education in the United States; conduct and publish reports and specialized analyses of the meaning and significance of such statistics; assist state and local education agencies in improving their statistical systems; and review and report on education activities in foreign countries.

NCES activities are designed to address high priority education data needs; provide consistent, reliable, complete, and accurate indicators of education status and trends; and report timely, useful, and high quality data to the U.S. Department of Education, the Congress, the states, other education policymakers, practitioners, data users, and the general public.

We strive to make our products available in a variety of formats and in language that is appropriate to a variety of audiences. You, as our customer, are the best judge of our success in communicating information effectively. If you have any comments or suggestions about this or any other NCES product or report, we would like to hear from you. Please direct your comments to:

National Center for Education Statistics
Office of Educational Research and Improvement
U.S. Department of Education
555 New Jersey Avenue, NW
Washington, DC 20208

Suggested Citation

U.S. Department of Education. National Center for Education Statistics. *Selected Papers on Education Surveys: Papers Presented at the 1996 Meeting of the American Statistical Association*, Working Paper No. 97-01. Project Officer, Dan Kasprzyk. Washington, D.C.: 1997.

February 1997

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 Ruth R. Harris at (202) 219-1831 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. 20208-5654.

Susan Ahmed
Chief Mathematical Statistician
Statistical Standards and
Services Group

Samuel S. Peng
Director
Methodology, Training, and Customer
Service Program

This page intentionally left blank.

Table of Contents

Design and Estimation in School-Based Surveys

Policy Analysis with Education and Defense Manpower Survey Data

Preface

The 11 papers contained in this volume were presented at the 1996 American Statistical Association (ASA) meeting in Chicago, Illinois (August 4-8). This is the fourth collection of ASA papers of particular interest to users of NCES survey data published in the ***Working Papers Series***. The 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, Working Paper 95-01, which included papers from the 1994 ASA meeting, and Working Paper 96-02, which included papers from the 1995 ASA meeting.

The papers presented at the ASA meetings usually summarize ongoing research. In some cases, NCES publishes more complete reports on the topics. Readers who wish to learn more about the topics discussed in this working paper may wish to examine these technical reports:

NCES Reports

- *An Analysis of Response Rates in the 1993-94 Schools and Staffing Survey* (forthcoming)
- *An Exploratory Analysis of Response Rates in the 1990-91 Schools and Staffing Survey* (NCES 96-089)
- *Design Effects and Generalized Variance Functions for the 1990-91 Schools and Staffing Surveys (SASS) Volume I--User's Manual* (NCES 95-342-I)
- *Design Effects and Generalized Variance Functions for the 1990-91 Schools and Staffing Surveys (SASS) Volume II--Technical Report* (NCES 95-342-II)
- *Quality Profile for SASS: Aspects of the Quality of Data in the Schools and Staffing Surveys* (NCES 94-340)

- *Adjusting for Coverage Bias Using Telephone Service Interruption Data* (NCES 97-336)
- *Use of Cognitive Laboratories and Recorded Interviews in the National Household Education Survey* (NCES 96-332)

- *NCES Measurement Error Programs* (NCES 97-464; forthcoming)

Michael Garet's ASA paper, "Developing a National Survey of Private School Expenditures," could not be included in this collection, but for those interested in this topic, *Strategies for Collecting Finance Data from Private Schools* (Working Paper 96-16) is available. A complete list of ***Working Papers*** to date can be found at the end of this publication.

This page intentionally left blank.

TEACHER QUALITY AND EDUCATIONAL INEQUALITY

Richard M. Ingersoll, University of Georgia
Department of Sociology, UGA, Athens, GA 30602

KEY WORDS: NCES data; teacher quality, educational inequality

Introduction

Do all secondary students in the U.S. have equal access to qualified teachers? Are students in low-income schools more likely to be taught by teachers without basic qualifications, in their assigned teaching fields, than those in more affluent schools? Do schools serving predominantly minority student populations tend to have less qualified faculties? Moreover, are there differences in access to qualified teachers across different types of students and different types of classrooms within schools?

Over the past several decades, equality has been one the most fundamental concerns of education policy and research in the U.S. The focus of a vast amount of research and reform in education has been to uncover and address disparities in the resources and opportunities in education provided to students from different socio-economic backgrounds (e.g. Coleman et al. 1966). Among the most important of these educational resources is the teaching force. The largest single component of the cost of education in any country is teacher compensation. Moreover, teachers are, of course, a highly important part of the actual educational process, and student educational outcomes ultimately depend on the work of teachers. Indeed, it is precisely because the teaching force is a significant resource that equal access to qualified teachers and quality teaching has been a source of contention in the national debate over equality of educational opportunity.

Among those concerned with issues of educational equality, it is widely believed that students from disadvantaged backgrounds do not have equal access to qualified teachers. A number of critics have argued that, indeed, the most needy students in the U.S. -- those from poor, minority, and disadvantaged communities -- are taught by the least qualified teachers (e.g., Darling-Hammond 1987; Kozol 1991; Oakes 1990). These critics argue that low-income and high-minority schools are unable to offer competitive salaries, benefits or resources and, hence, simply cannot compete for the available supply of trained teachers. In this view, unequal access to qualified teachers and, hence, to quality teaching, is one of the key reasons for unequal results in student educational outcomes.

These critics argue, moreover, that patterns of unequal access to quality teachers also appear within schools. Not only do students in low-income and predominantly minority schools have less access to qualified staff, but, the critics add, low-income and minority students, when in affluent schools, also have less access to the best teachers. The latter is due to the practice of separating students and teachers by purported ability -- the system of tracking. In this view, minority and poor students are disproportionately placed in lower track and lower achievement courses, which these critics further claim, are taught by the least qualified teachers.

Despite the importance of this debate on educational equality and the widespread belief that schools, programs, and classes serving low-income and minority student populations have less access to quality teaching, there has actually not been much empirical research done on this issue, especially at the national level. One of the reasons for this dearth of research is the difficulty involved in obtaining data on the underlying issue of importance - the degree of actual exposure to quality teachers and quality teaching provided to students in classrooms. Assessing the caliber of teachers' classroom performance and the degree to which students have access to quality teaching in classrooms is a difficult empirical task because there is little consensus concerning both how to define and how to best measure quality teachers and teaching (e.g., Haney et al. 1987; Ingersoll 1996a). As a result, researchers typically turn to what is more easily assessed and more readily available - measures of teacher qualifications.

Although the qualifications of teachers - such as their education, training, and preparation - are only indirect measures of the quality of teaching that students receive, they provide useful information on this important educational resource.

Education and training are essential ingredients of quality teachers and quality teaching. There is almost universal agreement that one of the most important characteristics of a quality teacher is preparation in the subject or field in which the teacher is 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 reviews of this research see: Shavelson et al. 1989; Darling-Hammond and Hudson 1990; Murnane and Raizen 1988). Knowledge of subject matter and of pedagogical methods do not, of course, guarantee quality teachers nor quality teaching, but they are necessary prerequisites.

The argument for the necessity of education in subject knowledge is especially clear for the secondary school level. First, at the secondary level, 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 needed to teach different subjects is higher at the secondary school level, and therefore a clear case can be made that such teachers ought to have adequate substantive background in the subjects they teach.

In order to address fully the issue of access to qualified teachers, however, it is necessary to distinguish between teacher training and teaching assignment. These represent two distinct elements. Teacher training refers to the quantity and quality of teacher education and preparation. Assessments of training levels typically examine whether teachers have a basic college education, licensing and expertise in a specialty field. On the other hand, assessments of teacher qualifications also need to examine whether the fields of training and preparation of teachers match their teaching assignments. That is, such assessments need to assess the extent of out-of-field teaching - the phenomenon of trained teachers teaching subjects for which they have little training. It is important to distinguish between these two elements in assessments of teachers' qualifications because they have very different implications for policy. If underqualified teaching is due to inadequacies in the quantity or quality of teacher education and preparation, it is probable the source of the problem may lie with teacher education programs and standards. On the other hand, if underqualified teaching is due to high levels of mismatch between teachers' fields of training and their teaching assignments, then it is probable the source of the problem may lie with the supply of teachers or the management of schools.

The problem for research on teacher qualifications has been that there have not been the necessary data, especially at the national level, to adequately assess the extent to which teachers are assigned to teach out of their fields. Moreover, there has been little data on the numbers of students actually taught by out-of-field teachers -- information crucial to understanding disparities in student access to qualified teaching.

In order to address these and other data needs concerned with the staffing, occupational and organization aspects of schools, in the late 1980s the National Center for Education Statistics (NCES) designed and conducted the Schools and Staffing Survey (SASS), a major new survey of teachers and schools. NCES has since sponsored several projects designed to define and assess both the qualifications of the nation's teaching force and the extent of out-of-field teaching in the U.S. (McMillen and Bobbitt 1993; Bobbitt and McMillen 1995; Ingersoll 1995a, 1996b). These previous analyses have shown that, in fact, out-of-field teaching is extensive in U.S. schools. Moreover, these analyses have documented that this underqualified teaching was not due to a lack of basic education or training on the part of teachers. The source of out-of-field teaching lay in the lack of fit between teachers' fields of training and their teaching assignments. Most teachers have training, such as a college major, in their main field of assignment. But, many teachers, especially at the secondary level, are also assigned to teach additional courses in fields for which they have little or no formal background preparation.

This article expands on this earlier work by analyzing national data from the 1990-91 SASS to examine the issue of disparities in student access to qualified teachers. Rather than enter the debate as to what constitutes a qualified teacher, quality teaching or quality teacher training, this analysis adopts a minimal definition of adequate qualifications. The premise underlying this analysis is that adequately qualified staffing requires teachers, especially at the secondary school level and especially in the core academic fields, to hold, as a minimum prerequisite, at least a college minor in the fields taught. The analysis focuses on how many secondary level students enrolled in the core academic subjects (mathematics, English, social studies, science) are taught by teachers without at least a college minor in the field. In this view, even a moderate number of teachers lacking such minimal training prerequisites is a strong indication of inadequacies in the staffing of schools.

The analysis examines whether access to qualified teachers is equally distributed across different student populations. It begins by focusing on differences between high-poverty and low-poverty schools, and also between high-minority and low-minority schools. Many researchers assume that the high-poverty and minority populations are one and the same. It is important, however, to examine the data on out-of-field teaching by these two characteristics separately, because previous research has suggested that differences in the levels of teacher qualifications are not always the same across them (Pascal 1987).

The analysis also examines within-school differences in teacher qualifications across classes of different student ability groupings, and of different student races and ethnicities. Again, it is also important to examine out-of-field teaching separately by these sets of characteristics because it cannot be assumed that their relationships to teacher qualifications are the same.

Finally, this analysis examines within-school differences in teacher qualifications across different secondary school grade levels - specifically, grades 7 through 12. Although many may agree that basic education is an essential prerequisite of qualified teachers, there is probably less agreement whether out-of-field teaching has as serious consequences at the junior high level as it has for the senior high grades. Hence, it is important to distinguish among grades at the secondary level and to determine whether there are, in fact, differences in out-of-field teaching levels across these different grade levels.

Data and Methods

The data source for this study is the nationally representative 1990-91 Schools and Staffing Survey. The U.S. Census Bureau collected these data for NCES in early 1991 from a random sample stratified by state, sector and school level. SASS is particularly useful for addressing questions concerned with teachers' qualifications. It is the largest and most comprehensive dataset available on teachers and school staffing characteristics in the U.S. Indeed, as indicated earlier, this survey was conducted for the reason that there has been a paucity of nationally representative data on such issues. SASS, for example, includes a wide range of information on the training, education, qualifications and teaching assignments of teachers that can be disaggregated by field and also disaggregated by the characteristics of schools, students and classrooms (For more information on SASS, see Choy et al. 1993).

The sample utilized in the analysis consists of 25,427 public school teachers, including those employed both full-time and part-time. This analysis focuses solely on those teaching at the secondary-school level (grades 7 through 12), regardless of whether the school was actually a middle school, junior high school, a senior high school, a secondary school, or a combined school. Furthermore, it solely focuses on those who taught departmentalized courses in any of the core academic fields (English, mathematics, science, social studies). For example, secondary level teachers teaching multiple subjects in self-contained classes were excluded from the analysis. Likewise, the non-7-12th grade portions of the schedules of teachers in combined schools or middle schools were excluded.

For each class period in the school day of each of the sampled teachers, data were collected on the subject taught, grade level, class type or track, student achievement level, student race/ethnicity and the number of students enrolled. In addition, teachers reported their certification status and the major and minor fields of study for each of their degrees earned, at both the undergraduate and graduate levels. We have used these data in a series of projects to develop and compare a range of different measures of out-of-field teaching (see McMillen and Bobbitt 1993; Bobbitt and McMillen 1995; Ingersoll 1995a, 1996b). This analysis focuses on one measure drawn from this earlier research: *the percentage of public secondary school students enrolled in classes taught by teachers without at least a minor in the field.*

Fields are defined broadly in this analysis. The core academic subjects and college major/minors are broadly categorized into four fields parallel to conventional core academic departments in secondary schools: mathematics, science, social studies, and English. To be defined as in-field, English teachers must hold at least a minor in either English, English education, language arts, literature, reading, communication or journalism. Mathematics teachers must hold at least a minor in either mathematics, engineering, or mathematics education. Science teachers must hold a minor in any of the sciences. Social studies teachers (i.e., history, economics, civics, world civilization) must hold at least a minor in one of the social sciences, in history, or in social studies education.

The objective of this analysis is to examine differences in the levels of out-of-field teaching among different types of schools, based on the poverty level and race/ethnicity of the students enrolled and across different kinds of classrooms within schools. These measures are:
Poverty enrollment of school - percentage students in each school receiving federal reduced or free lunch program.
Low-poverty: less than 15%
Medium-poverty: 15% to 50%
High-poverty: 50% or more
Minority enrollment of classroom or school - percentage non-white students.
Low-minority: less than 15%
Medium-minority: 15% to 50%
High-minority: 50% or more
Type or track of class -
Low-track: general, remedial, vocational, special education
Medium-track: academic/college preparatory
High-track: honors, advanced placement, gifted.
Grade level of class - grades 7 through 12

Results

What proportion of the nation's public secondary students are taught core academic subjects by out-of-field teachers?

Overall, substantial proportions of students in public secondary schools in the U.S. were taught academic subjects by teachers without basic qualifications in those subjects. The proportions of public secondary school students taught each of the core academic fields by teachers without at least a minor in the field are presented in table 1.

For example, about one fifth of all public school students enrolled in English classes in grades 7-12, or about 4,310,000 of 20,700,000 students, were taught by teachers who did not have at least a minor in English, literature, communications, speech, journalism, English education or reading education. In addition, over one quarter of all public school students enrolled in mathematics classes in grades 7-12, or about 4,124,000 of 15,510,000 students, were taught by teachers without at least a minor in mathematics or in mathematics education. In science, 17 percent of all public school students enrolled in science classes in grades 7-12 were taught by teachers without at least a minor in any of the sciences or in science education.

Overall, a relatively low proportion of students were taught social studies out of field; thirteen percent of students enrolled in social studies were taught by teachers without at least a minor in any of the social sciences, in public affairs, in social studies education, or in history.

Table 1 - Percentage of public secondary school students enrolled in classes taught by teachers without at least a minor in the field, by field and selected *school* characteristics

	English	Math	Science	Social Studies
Total	20.8	26.6	16.5	13.4
Minority Enrollment of School:				
Low-minority	20.0	24.3	13.9	11.6
Medium-minority	19.1	23.1	16.6	15.6
High-minority	24.4	33.6	17.8	14.4
Poverty Enrollment of School:				
Low-poverty	15.6	20.6	11.9	11.6
Medium-poverty	21.7	30.1	16.0	14.5
High-poverty	33.2	32.6	29.3	15.3

Are students in schools serving predominantly poverty-level or minority student populations more likely to be taught by out-of-field teachers than students in schools serving predominantly not poor or white students?

There were also differences in the amount of out-of-field teaching across different types of schools, but this depended on the type of schools compared and the fields examined. Notably, although in some fields there appear to have been slight differences in levels of out-of-field teaching between high and low-minority schools, in no fields were these differences statistically significant.

In contrast, school poverty levels were clearly related to the amount of out-of-field teaching and the differences were in the direction predicted by the literature on educational inequality. That is, in no fields did high-poverty schools have less out-of-field teaching than did low-poverty schools, while in several fields, students in high-poverty schools received distinctly more out-of-field teaching then in low-poverty schools. For example, a third of English students in high-poverty schools, as opposed to 16 percent in low-poverty schools, were taught by teachers who did not have at least a minor in English, English education, language arts, literature, reading, communication or journalism. There was, however, little difference in out-of-field teaching in social studies between schools of different poverty levels. Regardless of the school poverty level, all had relatively low levels of out-of-field teaching in social studies.

Are students in low-track, or lower-grade level classes, or classes predominantly comprised of minority students, more likely to be taught by out-of-field teachers than students in high-track, higher-grade level or predominantly white classes?

The amount of out-of-field teaching was not equally distributed across different types of classes and groups in schools. These data are displayed in table 2.

Table 2 - Percentage of public secondary school students enrolled in classes taught by teachers without at least a minor in the field, by field and selected *classroom* characteristics

	English	Math	Science	Social Studies
Total	20.8	26.6	16.5	13.4
Type or Track of Class:				
Low-track	24.7	33.5	20.4	14.3
Medium-track	11.8	15.7	9.2	8.9
High-track	11.2	20.4	7.2	11.2
Minority Enrollment of Class:				
Low-minority	19.2	22.7	14.6	12.3
Medium-minority	19.9	24.2	17.7	15.0
High-minority	25.2	36.1	19.6	14.3
Grade Level of Class:				
7th grade	32.2	48.8	31.8	23.9
8th grade	32.9	37.1	23.8	19.7
9th grade	15.7	18.1	10.7	8.7
10th grade	11.1	16.8	8.9	8.8
11th grade	11.2	15.9	6.4	6.8
12th grade	13.9	24.2	13.1	11.3

In several fields, students in high-track classes had less out-of-field teaching than did those in the low-track classes. For instance, about one tenth of students in high-track English classes were taught by out-of-field teachers. But, about one quarter of those in low-track English classes received out-of-field teaching. There was, however, little difference in levels of out-of-field teaching between the two higher tracks - the honors/gifted/AP track and the college preparatory track.

In contrast to tracks, there was little difference in out-of-field teaching between predominantly white and predominantly minority classes. In none of the fields was there a statistically significant difference in out-of-field teaching between high-minority classes and low-minority classes.

There were, however, some distinct differences between the junior high school grade levels and the senior high school grade levels. Students in grade 7 were more likely to have received out-of-field teaching than were 12th grade students in all fields, with the exception of math. For example, about one third of science students in 7th grade were taught by teachers without at least a minor in any of the sciences or science education; while this was true for only about a tenth of the science students in 12th grade. In some fields, students in grade 8 were also more likely to have received out-of-field teaching than were 12th grade students. There were not, however, distinct differences among the senior high grade levels. Ninth grade students, for example, were not necessarily more likely to have been taught by an out-of-field teacher than were 12th grade students.

Discussion

The data clearly show that many students in public schools in grades 7-12, regardless of the type of school, were taught core academic subjects by teachers without at least a college minor in the field taught. They also show that there were some distinct inequities in the distribution of out-of-field teaching across schools and classrooms. This article does not, however, address the question of what are the reasons, causes or sources of out-of-field teaching, nor why some schools or classrooms have more it than others. Other analyses using SASS data to examine out-of-field teaching offer some insights and I will review these below.

Many have argued that out-of-field teaching is a problem of poorly trained teachers. As indicated earlier, this view is incorrect. Out-of-field teaching is not due to a lack of education on the part of teachers but is due to a lack of match between teachers' fields of training and their fields of assignment.

Other educational analysts have argued that out-of-field teaching is due to teacher shortages. There is some truth to this view. Some schools do report having difficulties finding qualified candidates for teaching job openings and school administrators commonly turn to the use of substitute teachers, in-school reassignments and hiring of the underqualified as coping strategies. Out-of-field teaching is the inevitable result of these kinds of coping strategies. But, contrary to conventional wisdom, neither out-of-field assignments nor teacher shortages are primarily due to increases in either student enrollments or teacher retirements.

The demand for new teachers is primarily from teacher turnover, not increases in student enrollments. Moreover, poor working conditions, not teacher retirements, create most turnover. Hence, shortages result most often from poor working conditions. Low teacher salaries, little faculty input into school policies, and rampant student discipline problems all contribute to teacher turnover. Improving these things would decrease turnover, which would quickly eliminate shortages. It would also remove much of the need for out-of-field assignments in the first place.

This points to an alternative explanation of out-of field teaching -- the low status of the occupation. Unlike in many of the other developed nations, teachers in the U.S. are largely treated as low and semi-skilled workers. The data suggest that out-of-field teaching is not an emergency condition, but a normal and ongoing practice in many schools. This prevalence attests to how widely accepted is the idea that teaching does not require any special expertise and that teachers are like interchangeable blocks that can be placed in any empty slot regardless of their type of training. Clearly, if teaching was treated as a highly valued profession and provided with commensurate rewards, respect and working conditions, there would be no problem attracting and retaining more than enough qualified teachers, and out-of-field teaching would neither be needed nor permitted.

Related to the question of the causes of out-of-field teaching, is a second question - why do some schools have more of it than others? In particular, why do low-income schools have higher levels of out-of-field teaching?

As mentioned earlier, one view, widely held among critics of educational inequality, is that low-income schools are not able to attract, or to retain, adequately trained teachers because they are unable to match the salaries, benefits and resources offered by more affluent schools. As a result, these critics hold, such schools have difficulties hiring adequately trained teacher candidates and suffer from high levels of teacher turnover (e.g., Kozol 1990; Oakes 1990).

There has, however, been little empirical verification of this view and, moreover, data from SASS suggest that this explanation may not be entirely correct. The data show, for example, that starting-level and advanced-level salaries in high-poverty schools are not appreciably lower than in other schools. In addition, teacher turnover rates are also not appreciably higher in low-income schools (Ingersoll

1995a). Low-income schools do appear to have slightly more difficulty in filling teaching openings. But, these differences appear to account for some, but not all, of the high levels of misqualified teachers in such schools. SASS data show, for example, that there are several factors besides the overall poverty or affluence of the student population that are related to the degree of out-of-field teaching in schools. For instance, school size and sector are both strongly related to out-of-field levels; small schools and private schools both have distinctly higher proportions of out-of-field teaching (Ingersoll 1995a, 1996c). These issues warrant further research.

An additional important issue concerns equalities in access to qualified teachers, according to the race/ethnicity of students. As noted above, it is commonly believed among education analysts that both poor and minority students do not have equal access to qualified teachers (e.g., Kozol 1990; Oakes 1990). In contrast, this analysis finds few distinct differences in levels of out-of-field teaching, according to the proportion of minority students in classrooms or in schools. This does not mean, of course, that there are no inequalities in access to quality teaching and quality teachers, according to the race/ethnicity of students. There may be other kinds of differences in access that are not revealed by the data and measures used in this analysis. Moreover, this analysis does not separately examine different minority groups and, hence, there may be differences in access between different minority groups not revealed here. What this analysis simply shows is that minority students, as a whole, were not more likely to have been taught by out-of-field teachers. Moreover, it also corroborates the importance of distinguishing between race/ethnicity and income/poverty characteristics of student populations.

References

Bobbitt, S. & McMillen, M. (1995). *Qualifications of the public school teacher workforce: 1988-1991.* (NCES Report No. 95-665). Washington, DC: U.S. Department of Education, National Center for Education Statistics.

Choy, S., Henke, R., Alt, M., Medrich, E., & Bobbitt, S. (1993). *Schools and staffing in the US: A statistical profile, 1990-91.* (NCES Report No. 93-146). Washington, DC: U.S. Department of Education, National Center for Education Statistics.

Coleman, J., Campbell, E., Hobson, C., McPartland, J., Mood, A., Weinfeld, F., & York, R. 1966. *Equality of Educational Opportunity*. Washington, D.C.: U.S. Government Printing Office.

Darling-Hammond, L. (1987). "Teacher Quality and Equality." In P. Keating and J.I. Goodlad, *Access to Knowledge*. New York: College Entrance Examination Board.

Darling-Hammond, L. and Hudson, L. (1990). "Pre-college Science and Mathematics Teachers: Supply, Demand and Quality. *Review of Research in Education.* Washington, D.C.: American Educational Research Association.

Haggstrom, G. W., Darling-Hammond, L., & Grissmer, D. (1988). *Assessing teacher supply and demand.* Santa Monica CA: Rand Corporation.

Haney, W., Madus, G., & Kreitzer, A. (1987). Charms talismanic: Testing teachers for the improvement of American education. *Review of Research in Education,* 13: 169-238. Washington, DC: American Educational Research Association.

Ingersoll, R. (1995a). *Teacher Supply, Teacher Quality and Teacher Turnover: 1990-91.* (NCES Report No. 95-744). Washington, DC: National Center for Education Statistics.

Ingersoll, R. (1995b) "Teacher Supply and Demand in the U.S." In *The Proceedings of the American Statistical Association: 1995.* Alexandria, Va: American Statistical Association.

Ingersoll, R. (1996a). *National Assessments of Teacher Quality.* Washington, DC: U.S. Department of Education, National Center for Education Statistics.

Ingersoll, R. (1996b). *Out-of-Field Teaching and Educational Equality.* (NCES Report No. 96-040). Washington, DC: National Center for Education Statistics.

Ingersoll, R. (1996c). *The Problem of Out-of-field Teaching in the U.S.* Paper presented at the annual meeting of the American Sociological Association.

Kozol, J. (1991). *Savage Inequalities*. New York: Harper-Collins.

McMillen, M. & Bobbitt, S. (1993). *Teacher certification, training and work assignments in public schools.* Paper presented at the annual meeting of the American Education Research Association.

Murnane, R. and Raizen, S. (1988). eds. *Improving indicators of the quality of science and mathematics education in grades k-12*. Washington, D.C.: National Academy Press.

Oakes, J. (1990). *Multiplying inequalities: The effects of race, social class, and tracking on opportunities to learn mathematics and science.* Santa Monica, CA: The RAND Corporation.

Pascal, A. (1987). *The qualifications of teachers in American high schools.* Santa Monica, CA: The RAND Corporation.

Shavelson, R., McDonnell, L. and Oakes, J. (1989). *Indicators for monitoring mathematics and science education*. Santa Monica, CA: Rand Corporation.

USING QUALITATIVE METHODS TO VALIDATE QUANTITATIVE SURVEY INSTRUMENTS

John E. Mullens, Policy Studies Associates
Daniel Kasprzyk, National Center for Education Statistics
John E. Mullens, PSA, 1718 Connecticut Avenue, N.W., #400, Washington, DC 20009

Key Words: validation, case study, focus group

Surveys are among the most cost-effective and least burdensome methods of collecting data on schools, classrooms, and teachers, but both researchers and respondents know that brief, self-report strategies may not portray a picture of instruction as sufficiently as needed to confidently assess instructional effectiveness. For a project at the National Center for Education Statistics (NCES) designed to investigate techniques and instruments to measure and understand the instructional processes used by eighth to tenth grade mathematics teachers, we used a multi-step pilot-study process to construct, refine, and validate survey instruments. In the process, we honed our knowledge about instruments and methods for collecting accurate, valid, and meaningful information that can be incorporated into future national data collection schemes.

In this paper, we outline the project scope, describe the data collection methods used, and assess their role in evaluating survey responses and improving instruments to provide portraits of relevant classroom processes. A complete description and assessment of this project is reported in Mullens and Leighton (1996).

The Classroom Instructional Processes Study

The NCES project, "Understanding Classroom Instructional Processes," was designed to (1) develop, pilot, and evaluate methods for collecting data on classroom instructional practices; (2) explore the combined use of questionnaires and related teacher log forms to portray classroom instructional processes; and (3) determine the feasibility of incorporating such methods into future NCES surveys or other data collection efforts. The project piloted focus groups and case studies (using classroom log forms, observations, and artifact collection) to assess the completeness and accuracy of data obtained from questionnaire responses. Through this process, we used qualitative methods to validate responses on quantitative survey instruments.

The results were intended to help NCES make decisions about data collection methods and instruments with which to develop an accurate portrait of eighth to tenth grade mathematics instruction. Having such data would expand NCES's ability to respond to Congress, other offices in the Department of Education; and other federal agencies, state departments of education, associations concerned with elementary and secondary education, and education research organizations. Data from previous surveys on similar topics have been used by all of these sectors, and in recent years there has been interest in expanding the scope of these data.

Context

Increased use of high stakes student testing as a measure of educational productivity has led to increased interest in determining the precise contribution of schooling to achievement, distinct from, for example, the contributions of prior learning or socioeconomic status. Experts in identifying the correlates of student achievement, such as Porter (1991) and Schmidt (1995) argue that many factors are at work: the content must be presented cogently, using subject-specific instructional techniques appropriate to both the material and the students' prior knowledge, and with emphasis matched to the topic's relative importance among desired outcomes. Valid and reliable assessments of instructional content and practices can contribute to descriptions of educational experiences, help explain achievement outcomes, and inform educational policy development at the local, state, and national levels (Burstein, Oakes, Guiton, 1992; Smith, 1988; Murnane, 1987). Despite this, minimal data on instructional practices are available from a nationally representative sample of U.S. classrooms.

This study builds on a prior review of existing measurement approaches (Mullens, 1995), and focuses on four major dimensions of classroom instruction: the conditions and context that direct or influence a teacher's selection of content and instructional methods; the course content and emphasis on those topics; patterns of classroom pedagogy and how teachers approach the process of teaching; and the resources available and used in the classroom.

Research Question

The study goal was to produce and evaluate instruments and methods that would provide data on how the instructional processes and content of eighth to tenth grade mathematics classes vary across the country. Within this overall goal, we also expected to advance our understanding about instruments and methods for

capturing accurate and meaningful information about classroom instructional processes: information that could be incorporated into national data collection schemes. Data from the field tests could provide evidence with which we could understand more about the items and instruments themselves. The full study explored three measurement questions, one of which is the focus of this paper:

> Do qualitative data collection instruments and techniques provide validating information with which we can better construct questionnaires?

Study Design

We selected mathematics as the focus of this study because it is a core subject of great interest to policy makers and thus one in which early exploratory work already provided a sound research foundation for further study. Within this content area, the study concentrated on eighth to tenth grade mathematics courses, covering topics in pre-algebra, algebra, and geometry: the math courses designed to serve as a "bridge" to more advanced math courses, and to offer students a conceptual understanding of mathematics with broad applications in life.

From preliminary research, we decided to validate the teacher questionnaire through focus groups and case studies. Focus groups are roundtable open discussions with small numbers of respondents who had already completed the questionnaire, to examine each item for unfamiliar or inexact terminology and how well the items and their responses represented their own teaching.

Case studies of classroom teachers included observing their teaching, having them maintain daily logs, and collecting artifacts of their instruction. We observed classroom instruction to evaluate the completeness of the instrument, looking especially for conceptual gaps in our understanding of how instruction occurred, and in how it was represented on the instruments. Daily logs, or diaries, were records documenting learning objectives, teachers' actions, students' activities, and the materials used during a single class. Four weeks of data enabled us to evaluate the consistency between teacher's questionnaire responses and her daily recordings of activities. Examining the instructional materials or artifacts used by teachers during that same period of time were intended to provide information on the same events from a different slant.

Elements of these processes to validate questionnaire items have been explored and improved in several recent studies in this field, including Reform Up Close (Porter, Kirst, Osthoff, Smithson, & Schneider, 1993), Third International Study of Mathematics and Science (1991), and Validating National Curriculum Indicators (Burstein, McDonnell, Van Winkle, Ormseth, Mirocha, & Guiton, 1995).

We piloted our instruments and process in two school districts, revised them, and obtained OMB clearance. We fieldtested the instruments and process in three school districts: one was a large, independent, urban district on the West Coast; the second was a large city/county urban district in the Southeast; and the third was a smaller, suburban/rural, county district in the Mid-Atlantic region. In all, 111 teachers completed questionnaires, and seven teachers, one or more from every field test site, participated in the case study.

Data Collection Methods

We used focus groups and case studies to validate responses on the teacher questionnaire.

Focus Groups

At each of the three sites, all teachers of eighth, ninth, and tenth grade mathematics received a letter of invitation that explained the study and requested their participation. Attendance at the seven voluntary focus group meetings ranged from one to 12 teachers and totaled 38. Teachers commented on their understanding of the item's intent and the appropriateness of the response format.

The greatest teacher concern across all sites was not (as we suspected) that the teacher questionnaire would not adequately portray their teaching, but that the particular class they were asked to describe (the first instructional period of the day) was not representative of their whole teaching load. Specific characteristics of the students in that class, according to most teachers, caused them to teach in some manner they felt was not representative of their overall efforts. Despite this concern, most focus groups came to the conclusion that while there was no single period that would catch each of them at their most representative, the combined results of all sampled teachers would indeed represent the overall collection of the activities of all teachers throughout the day.

Case Studies

Volunteer case-study teachers were observed by a project researcher, kept a daily log of classroom instructional activities and those of the students in their designated class during a four-week period, and collected instructional artifacts. The project had seven case studies.

Classroom logs. Information from classroom logs was used to assess the consistency of teachers' daily recordings of classroom practice with their one-time account of practice from the teacher questionnaire. The picture of classroom practice obtained from multiple weeks of log form data is a finer grained view of the enacted pedagogy than that provided by a teacher's questionnaire responses summarizing a semester of practice. For both practical and perceptive reasons, completing logs daily (or at most, weekly) is likely to result in more accurate data than a one-time retrospective survey. Because logs rely on teachers' short-term memory rather than their long-term memory and the summative ability needed for the questionnaire, the resulting data can be presumed to be more accurate. Furthermore, teachers may be more inclined toward honest accounts on a daily rendering since a single daily log, unlike teacher questionnaire responses, becomes one of many depictions of their practice. Classroom logs were also intended to be used by researchers to record events and activities as they observed in classrooms.

A weakness of prior research had been the inability to use data from classroom logs to estimate the reliability of questionnaire items (Porter, 1993). This difficulty stemmed from using different items of the log than were included on the questionnaire. We designed the log to be completed by case-study teachers as a record of the classroom instruction occurring during a single class period so that the data from four weeks of logs could be used to evaluate the validity of the teacher's responses on the questionnaire. To make this direct link possible and to build on the knowledge gained from prior research, we constructed the log by directly copying specific items and activities from those on the questionnaire; frequency response options covering a semester were replaced with time per use response options covering a single period. Sharing identical items between the two instruments was intended to facilitate the later comparison of the teacher's daily logs with her responses on the survey.

Classroom observations. Researchers observed case-study volunteers to help them understand the function of the classroom log and the process of using it. After observing the teacher instructing the targeted class, both the researcher and the case-study teacher completed a log form. Teacher and researcher then compared observations, discussing differences in coding. For all but two teachers, those differences were slight. Because these teachers had participated in the focus group discussions of the items, most already understood nuances of meanings that might make a difference in how they recorded their instruction. Researchers had enough concerns about the coding patterns of one teacher, however, to repeat the calibration process a second time.

Artifact collection. To provide further detail about their lessons (and reduce the need for written explanations), case-study teachers were asked to submit certain instructional items figuring prominently in lessons for the designated class. Such instructional items included copies of homework and in-class assignments; directions for papers, reports, or projects; copies of tests and quizzes; and any other written assignments. These artifacts were intended to provide another avenue through which researchers could interpret the teacher log data for each lesson.

Assessing the Methods

While each qualitative method helped validate the quantitative data obtained from the teacher survey, some contributed more information than others to our analysis.

Focus Groups

The purpose of focus groups was to provide respondent feedback on the survey instrument. They served that purpose well, and, unexpectedly, proved to be a major source of case-study volunteers. Especially at the beginning, focus groups allowed researchers to directly hear respondents' comments and probe their exact meanings. Such exchanges allowed both researchers and respondents to raise and explore many issues usefully, to validate the relevance of certain items across sites, to hone wording, and to generate additional ideas of emerging instructional practices. For example, an earlier version of the questionnaire included "calculator" on the list of instructional materials. Teachers were asked to indicate if calculators were available for use by students. In one focus group, teachers suggested that was not the issue. They had plenty of calculators available and even sufficient batteries. But the calculators were not sophisticated enough to do the kinds of operations the teachers wanted to teach their students. Because of that

discussion, the item was changed to "appropriate calculator".

Teachers at another site complained that the questionnaire afforded them no opportunity to indicate they structured their classes around cooperative learning strategies. They explained that cooperative learning was a major effort in the school district's instructional program, yet there was no place on the questionnaire to indicate the use of that strategy in the classroom. When that same comment was heard in another location, it was added to the list of teacher activities and student activities.

As the project continued and researchers held additional focus groups in new fieldtest sites, however, the utility of additional information new to researchers substantially decreased. We obtained little new information from the later focus groups.

Case Studies

The case studies provided substantial information with which to assess the construction of the teacher questionnaire.

Classroom logs. The project was not funded to design a process to validate the reliability of questionnaire items, but to fully understand the benefits and limitations of the information that we could obtain from logkeeping, we measured the consistency between teachers' daily reports of instructional activities and their semester report of the same activities. Assuming that the daily log reports had a higher level of teacher reporting accuracy than the questionnaire responses for reasons stated above, we used log responses to assess the reliability of the questionnaire responses. For each case-study teacher, we compared the sum of log-reported activities across a representative four week period with that teacher's questionnaire-reported activities over the semester. For example, if the questionnaire responses indicated that the teacher stimulated student discussions of multiple approaches more than once a week, we expected to see confirming entries of such discussions on the teacher's daily log. This provided a measure of the reporting reliability of individual questionnaire items.

Our sample of seven case-study teachers was purposive and too small to generalize to the larger sample of all survey respondents; nonetheless, Table 1 illustrates the type of information we might obtain using this process with a larger and appropriately random subsample of case-study teachers.

Table 1: Examples of consistency between (a) teachers' survey responses describing a semester of classes and (b) their class log entries over a four-week period, on the same teacher activity (nonrandom sample, N=7).

Teacher Activities	Percent direct agreement	Percent agreement within one survey response value category
Provide individual or small group tutoring as needed during individual seatwork or small group activities involving everyone	100	NA
Lecture, perhaps occasionally using the board or overhead projector to highlight a key term or present an outline	71	86
Demonstrate a concept, using two-dimensional graphics such as drawings on the board, overhead projector, or computer	71	86
Provide supplemental--remedial or enriching--instruction to a pull-out group while the rest of the class works in assignments	71	86
Administer a test or a quiz	57	86
Demonstrate a concept, using three-dimensional tools such as manipulatives, models, or other objects	57	71
Lead students in discussion, recitation, drills, or question-and-answer sessions	43	100
Observe or monitor student-led discussions	43	57
Work on administrative tasks while students work on assignments	29	57

Table reads: In a nonrandom sample of seven teachers over four weeks, teachers' responses on a survey item about tutoring were consistent in 100 percent of cases with their responses on the log item on the same topic. Teachers' responses on the survey item about lecturing were consistent with their log responses in 71 percent of cases and within one response value in 86 percent of cases.

These data, reporting consistencies between the teacher questionnaire and the log form for teacher activities from item 13 of the questionnaire, suggest that teachers' recollection of their instructional behavior varies according to the activity being reported. Examining the extreme cases, for example, there was one hundred percent consistency between teachers' reports on the questionnaire and on their daily logs about the frequency with which they provided individual or small group tutoring. Teachers apparently remember that type of instruction well. There was far less agreement (29 percent) between questionnaires and logs on how often the teacher works on administrative tasks while students work on assignments. Temporarily relaxing stringency to assess agreement within one survey response category increases the level of consistency between logs and questionnaires on this same item but only to 57 percent.

For four items, the rate of agreement between responses on the two instruments is above 70 percent, while in five it is 57 percent or less. To better understand these differences, we examined the direction of the mismatch of the five items with low agreement for possible evidence of socially desirable questionnaire responses. For three such items, data from the daily logs reported that the following activities actually occurred *more frequently* than teachers' questionnaire responses would indicate:

- administrative tasks
- drill and recitation
- student-led discussions

In one light, these responses may show evidence of social desirability, since the first two activities could be considered old fashioned or less than desirable in a climate of reform. Such an argument would suggest that subtle pressures may have influenced teachers' questionnaire responses. That student-led discussions appear to have actually happened more often than teachers indicated they did does not seem to follow that same explanation.

For two other items, data from the daily logs suggested that the following activities occurred *less frequently* than teachers' questionnaire responses would indicate:

- demonstrating a concept with three-dimensional tools
- administering a test or quiz

These differences suggest that teachers like to think they use three-dimensional manipulatives more than they actually do and that teachers administer fewer tests and quizzes than they might think.

Recalling again that this is only a demonstration analysis based on seven nonrandom sets of logs, we suggest no generalization of results beyond these seven teachers. The process, however, seems to show promise. Specific results from a larger and representative validation study might be different, but would likely be no less interesting.

Classroom observations. We designed the classroom observations of case-study teachers and the later discussion about completed log forms to provide those teachers with an experiential-based understanding of the meanings of the log form terms, and with practice in completing the form. Discussing specific events occurring within a particular class and how they translated into log form responses established common understandings of log form terms more directly than would have resulted from an abstract discussion only. Conducting multiple classroom observations across different sites and the resulting observation data also provided researchers with evidence with which to assess (1) the ability of the survey instrument to portray classroom processes accurately and (2) the match between actual classroom practice and survey scope, individual items, and response formats. In addition, the nonjudgmental research approach to discussing the observed instructional activities proved to be an unanticipated and effective method for cementing teacher cooperation and building confidence in the process.

During the fieldtests, researchers used their copy of the observation form to record classroom activities as they occurred, creating a real-time log of the instructional processes occurring during a single class session. Having the researcher use a more structured classroom observation instrument with which to initially record teacher and student activities and elapsed time may improve researchers' understanding of how teachers record their instructional processes on the log form, and may result in a more accurate recording of the duration of instructional elements occurring during instruction and the order in which they occurred.

Artifacts. We collected artifacts from case-study teachers to investigate the potential of such documents to more completely describe or illuminate classroom instructional processes. Although this process was

inexpensive, it afforded little analytical benefit to this study. The artifacts collected were primarily assignment sheets and examples of student work. We know in some cases, and suspect in others, that participating teachers sent incomplete records of the mathematic textbooks they used. Textbook pages or items used during instruction were the most notable void. Those artifacts we did have were difficult to assess. We can identify, for example, what was used during class (e.g., a practice sheet) and evaluate some elements of its content (e.g., estimation) but can tell little from the artifact itself about the instructional objective being addressed, how the artifact was used, or the amount of emphasis given to each element of the artifact. In further experiments with artifacts, we would investigate developing (1) a teacher checklist of contextual data surrounding the artifact's use within the lesson; and (2) a specific protocol for assessing important features of the artifact (such as instructional objective) and its use. Such protocols may be time consuming (and therefore expensive) to implement, effectively negating the original low cost of collecting the artifacts. So although artifact analysis may have great potential to add substance to self-reports, the process needs further attention.

Summary

With this task, we evaluated the usefulness of supplementary data collection in the form of focus groups and case studies in contributing to an understanding of how well our instrument measured the domains of interest, and how well the survey responses represented what teachers actually do. The focus group discussions provided excellent feedback on the survey, but are limited in the amount of new information provided by multiple focus groups. The case-study process, and classroom logs in particular, provide a valuable estimation of the consistency between responses on teacher questionnaires and on class logs. Classroom observation is beneficial to the researcher's understanding of the phenomenon under investigation and to the process of gaining trust for later segments of data collection, but we were disappointed in the results of our attempts to use artifacts to expand our understanding of classroom instructional processes, and see further experimentation as the key to greater benefits.

Based on the results of the research described above, we think certain qualitative methods can expand the ways in which we validate quantitative survey instruments. We are currently embarking on a project to survey a sample of 400 teachers of eighth to twelfth grade mathematics, engaging a subset of 60 in case studies. We will use the results reported here to expand our use of classroom observations and logs to validate the quantitative survey instruments.

References

Burstein, L., Oakes, J., & Guiton, G. (1992). Education indicators. In M.C. Alkin (Ed.), Encyclopedia of educational research (5th ed., pp. 409-418). New York: MacMillan.

Burstein, L., McDonnell, L., Van Winkle, J., Ormseth, T., Mirocha, J., & Guiton, G. (1995). Validating national curriculum indicators. Santa Monica, CA: RAND.

Mullens, J. (1995, April). Classroom instructional processes: A review of existing measurement approaches and their applicability for the Teacher Follow-up Survey (Working Paper No. 95-15). Washington, DC: National Center for Education Statistics.

Mullens, J., & Leighton, M. (1996). Understanding classroom instructional processes (draft). Washington, DC: Policy Studies Associates.

Murnane, R. (1987). Improving education indicators and economic indicators. Educational Evaluation and Policy Analysis, 9(2), 101-116.

Porter, A. (1991). Creating a system of school process indicators. Educational Evaluation and Policy Analysis, 13(1), 13-29.

Porter, A. (1993). Defining and measuring opportunity to learn. The debate on opportunity-to-learn standards: Supporting works. Washington, DC: National Governors' Association.

Porter, A., Kirst, M., Osthoff, E., Smithson, J., & Schneider, S. (1993). Reform up close: An analysis of high school mathematics and science classrooms. Madison, WI: Wisconsin Center for Education Research.

Schmidt, W. (1995, June). Presentation made at the Council of Chief State School Officers National Conference on Large Scale Assessment, Phoenix, AZ.

Smith, M. (1988, March). Educational indicators. Phi Delta Kappan, 487-491.

Third International Mathematics and Science Study. (1991). Project overview. East Lansing, MI: Author.

REVISING THE NCES PRIVATE SCHOOL SURVEY: A METHOD TO DESIGN A SYSTEMATIC CLASSIFICATION OF PRIVATE SCHOOLS IN THE UNITED STATES

Sylvia Kay Fisher, Bureau of Labor Statistics
Daniel Kasprzyk, National Center for Education Statistics
Sylvia Kay Fisher, BLS, Room 4915, 2 Massachusetts. Ave. N.E., Washington D.C. 20212

Key Words: Survey research methods, Cognitive techniques, Educational surveys

INTRODUCTION AND BACKGROUND

The Private School Survey (PSS) is a paper-and-pencil survey administered by mail to approximately 27,000 private schools throughout the United States under the auspices of the National Center for Education Statistics (NCES), U.S. Department of Education. The PSS form is generally completed at the private school site by the principal school administrator or office administrative assistant. The PSS is used for many purposes, including the identification and classification of private schools in the United States. Specifically, five PSS items are used to classify private schools into one of nine categories comprising a typology scheme designed to enhance federal-level statistical reporting about U.S. private school education.

The Bureau of Labor Statistics (BLS) was requested by NCES to conduct a series of tests on specific PSS questions to improve their ability to capture accurate information and improve the classification of U.S. private schools. Of particular concern was a question asking schools about their membership status relative to a series of school associations; this item has, in the past, led to respondent burden and possible underreporting. BLS developed a cognitive testing plan to address questions raised by NCES regarding the PSS. The purpose of the testing plan was to evaluate the language and visual construction of the five PSS questions, the sequencing between and within questions, and the manner in which respondents interpret the questions.

The cognitive testing plan is scheduled to be completed in two years of activities. The purpose of Year 1 activities was to conduct developmental questioning and directed probing to obtain information about the respondent's response process, and included a total of eighteen interviews with private schools and eleven educational experts. Year 1 activities were composed of three pretesting waves: Wave 1 and Wave 3 involved interviews with private schools, respectively. Wave 2 consisted of interviews with eleven educational experts

Feedback obtained from all Year 1 respondents were utilized to develop revision(s) of PSS questions, which are to be field-tested using a large national representative sample during Year 2 of the study. As part of the nation-wide field test, a validation study to estimate the degree of respondent error resulting from implementing the new PSS question(s) will be conducted. The results of Year 2 field-testing will yield a final set of items to be incorporated within NCES's Private School Survey.

This paper summarizes findings obtained through the first year of pretesting conducted during Waves 1-3. Final recommendations resulting from Year 1 testing are provided.

METHODOLOGY

Wave 1 Testing: The overall purpose of Wave 1 testing was to obtain preliminary information regarding how respondents answered PSS items, identify problems resulting from the response process, obtain information about how PSS information is collected at the private school site, and compile suggestions offered by participating school personnel. The original PSS items and an interview protocol was administered to nine private school respondents. Data collected from Wave 1 testing was used to develop preliminary revisions of the PSS items that could be tested in latter waves of testing.

Participating Schools: Private schools were categorized using grade level [kindergarten, elementary, and secondary] and type of affiliation [formal religious, informal religious, and non-sectarian] as major variables. The intersection of these two variables (and their three levels, respectively) resulted in the generation of a 3 X 3 matrix. One school was selected to represent each matrix cell. Wave 1 private schools were selected randomly from a computer-generated listing produced by NCES, which comprised all private schools in Miami, FL, Atlanta, GA, Boston, MA, and Minneapolis-St. Paul, MN. Private schools were selected from all four states, such that each of the nine cells in the 3 X 3 matrix would be represented.

Data Collection Procedures: Wave 1 respondents were private school administrators and/or administrative assistants responsible for completion of the PSS on behalf of the school. Participants were contacted by telephone and appointments were scheduled at a mutually convenient time. Interviews were conducted at the private school site and were approximately one hour in length. Respondents participated voluntarily and were not paid for their participation.

Cognitive Interviews: Interviews consisted of face-to-face cognitive tasks and the administration of an interview protocol. Respondents were queried about several aspects of the PSS, including: the definition of the terms "orientation," "purpose," or "affiliation;" the meaning of the term "school association;" and the process by which schools become members of a school association. Respondents completed a task in which they clustered 38 private school associations into groups and labeled them. They also provided a list of associations with which the school is affiliated. Their feedback and recommendations were used to develop preliminary revisions of the PSS items and examine their effectiveness in latter Waves of pretesting.

Wave 2 Testing: Wave 2 testing involved the conduct of cognitive interviews with eleven educational experts, representing national-level educational and religious school associations, based primarily in the greater metropolitan Washington D.C. area. Experts were consulted regarding the original PSS items, preliminary revisions of the PSS items, conceptual definitions, and their insights into private education. Their feedback and recommendations were used to modify the preliminary revisions of the PSS items and improve their efficacy and appropriateness for future testing.

Data Collection Procedures: Eleven association representatives were contacted by telephone and appointments were scheduled at a mutually convenient time. Interviews were conducted face-to-face at the association site and were approximately one hour in length. Respondents participated in the study voluntarily and were not compensated for their participation. All interviews were conducted within the greater metropolitan Washington, D.C. area (including Maryland and northern Virginia) and the state of Missouri.

Cognitive Interviews: Experts were queried about the PSS questions, conceptual definitions, and their insights into private education. Experts completed the original PSS items as though they were staff members in one of the schools belonging to their association. Many experts had been teachers and/or school principals and were able to accomplish this task easily. Feedback was obtained from the experts regarding the effectiveness of the original items. Experts also evaluated preliminary revisions of the PSS items, and provided their opinions about the effectiveness of the proposed revisions. Their responses and additional comments were used to generate a final version of the PSS items to be tested with private schools during Wave 3 testing.

Wave 3 Testing: The major objective of Wave 3 testing was to test the effectiveness of the newly revised PSS item revisions and compare their effectiveness to the original PSS items, as part of an interview conducted with private school respondents. Respondents were encouraged to offer comments and suggestions regarding the effectiveness of the revised versions of the PSS items.

Participating Schools: The 3 X 3 matrix used during Wave 1 testing was used to identify private schools participating in Wave 3 activities. Nine private schools were randomly selected from a computer-generated listing provided by NCES of all Washington, D.C. area private schools (including Maryland and Northern Virginia). Private schools were selected to ensure greater representation of private schools across the nine-part typology used by NCES; specifically, a home school, an Islamic school, an Orthodox Jewish school, a Montessori program school, and a school that only serves students with identified learning disabilities were identified to enhance the representativeness of sampled private schools.

Data Collection Procedures: Wave 3 respondents were private school administrators and/or administrative assistants responsible for completion of the PSS on behalf of the school. Participants were contacted by telephone and appointments were scheduled at a mutually convenient time at the private school site. Interviews lasted approximately forty-five to sixty minutes. Respondents participated in the study voluntarily and were not be paid for their participation.

Cognitive Interviews: The nine interviews consisted of a face-to-face cognitive interview and a brief, customized set of questions geared to the specialized needs of the participating private school. Respondents also completed the original and revised versions of Item 15(a)-15(d) and the original and two revised versions of PSS [16]. Respondents provided their evaluation of each version of the PSS items, and were encouraged to offer suggestions about ways to making the PSS more relevant to the specialized needs of their school. These informal interviews were designed to obtain additional information about the diverse and varied group of private schools who participated in the study.

Pretesting Results: Issues Associated with Completing Each PSS Item

15(a): ***Does this have a religious orientation, purpose, or affiliation?***

Response Issues: Pretesting identified the following potential sources of response error:

1. Respondents generally distinguished the terms religious "orientation" and "purpose" from "affiliation," which was perceived to refer to a formal relationship with an established religious group, institution, or denomination, with possible implications for school finances, curriculum, hiring practices, and expectations of students.
2. Many respondents had difficulty distinguishing religious "purpose" from "orientation," which were perceived to be virtually synonymous, and generally agreed that a school with a religious orientation or purpose is not necessarily beholden in any way to a religious institution or group. Those respondents who distinguished "purpose" from "orientation" reported that a private school's purpose stems from its orientation or vice versa.
3. Some respondents reported that some private schools have a formal affiliation with a religious organization, yet have no actual religious orientation or purpose. These schools were often originally founded by an established church, but no longer maintained a religious orientation, purpose, or programming. These schools differ from schools that maintain a religious affiliation with a religious organization or body, and/or a religious orientation or purpose that is expressed in the school's daily life.
4. Experts also identified problems in distinguishing the three terms "orientation," "purpose," and "affiliation;" however, they were unable to achieve consensus regarding the best means to improve 15(a).
5. Definitional problems affect responses to 15(a), but can also affect responses to 15(b) and 15(c). Respondents who endorse "yes" to 15(a) are directed via an accompanying arrow to respond to 15(b) and 15(c), while respondents who select "no" are directed via another arrow to *"GO to item 16, page 10."* Thus, errors in responding to 15(a) can result in response errors for Items 15 (b) and 15(c).

Final Modification of Item 15(a): Pretesting results verify that the use of all three terms "orientation," "purpose," and "affiliation" is necessary to account for the subtle complexities associated with the many types of relationships religiously-oriented private schools have with religious institutions and/or groups. As presently worded, 15(a) appears to most effectively capture the greatest number of correct endorsements. Therefore, no modifications were made to 15(a), and its originally form was retained for the PSS field test.

15(b): ***Is this school/program formally affiliated with a national religious denomination?***

Response Issues: Pretesting identified the following potential sources of response error:

1. The term "formally affiliated" was conceptually difficult for many respondents. Because some private schools are operated independently by a single church or temple that sponsors the school, many respondents (including Jewish and non-denominational Christian respondents) do not perceive themselves to have a "formal" affiliation with a national religious denomination, and fail to endorse "yes" to 15(b).
2. The terms "national" and "denomination" in the phrase "national religious denomination" provided sources of cognitive confusion for some respondents. Roman Catholic, Jewish, and Islamic respondents were unclear whether they should endorse "yes" to 15(b), because they perceived their religions to be international rather than "national" in scope. Similarly, the term "denomination" had little meaning for these respondents; many favored the use of a more inclusive term less associated with Protestant mainstream religious groups, such as "organization" or "institution." These issues were also identified by the experts during Wave 2 testing.

Final Modification of Item 15(b): Item 15(b) was revised to address the potential sources of response error identified through pretesting. The following modifications were made to 15(b) to facilitate the respondent's ability to endorse the appropriate response:

- Item 15(b) was revised to drop the word "national," which posed difficulties for many religiously-oriented respondents This modification was approved by all nine respondents during Wave 3 testing (including Jewish, Roman Catholic, and Islamic).
- The term "formally" was dropped and replaced with "affiliated in some way" to account for all possible types of affiliations private schools may have with a religious group and/or a single religious institution (e.g., church, temple, etc.).
- The phrase "religious denomination" was amended to read "religious denomination or organization" to include all respondents from religions that do not recognize the term "denomination;" this should facilitate the respondent's ability to endorse "yes" appropriately. The term "organization" was added because it subsumes both religious groups

and individual and independent religious institutions.

- 15(b) was amended for the field test as follows:

15(b): ***Is this school/program affiliated in some way with a religious denomination or organization?***

15(c): ***What is this school's religious orientation or affiliation?***

Response Issues: Item 15(b) contains a list of 20 response options representing major religious groups that exist in the U.S (e.g., Roman Catholic, Amish, Calvinist, Islamic, Jewish, etc.). "Roman Catholic" is the first option provided, because the majority of U.S. private schools are Roman Catholic. Four Lutheran options, representing the four major U.S. Lutheran synods., are also provided within the list. These options are labeled by their acronyms, and are not separated from the other available options. The final option provided is "other;" a blank line is also provided for respondents who choose to write in a religious group not included within the list of options. Pretesting identified the following potential sources of response error for 15(c):

1. Some respondents were concerned that "Roman Catholic" was not placed alphabetically in the list, and hypothesized that its placement as the first option might offend some respondents. A few experts also favored alphabetizing the list of religions for this reason.
2. Some respondents were concerned that the inclusion of four Lutheran groups within the list of religious groups may cause the list to appear to be weighted too heavily in favor of one group. Therefore, for pretesting Waves 2 and 3, the four Lutheran groups were collapsed into a single option "Lutheran." However, four Wave 2 experts indicated that all four sub-groupings of "Lutheran" should be retained in the final version of 15(c), because the four Lutheran Schools groups varied substantively in management and philosophy, and these distinctions should be duly noted. All four experts agreed it would be acceptable to place the four Lutheran groups under the greater category heading of "Lutheran."
3. Respondents reported no problems with the stem of 15(c) as presented in the current PSS.

Final Modification of Item 15(c): Item 15(c) was revised as specified below to address the potential sources of response error identified through pretesting:

- Because more than one-half of U.S. private schools are Roman Catholic, it is reasonable, logical, and convenient to place the most frequently selected option first in a list of options. Three experts representing religious private school associations (only one of which was Catholic) agreed it was acceptable to retain the current placement of "Roman Catholic," because of large number of Roman Catholic private school respondents. Therefore, the current placement of the option "Roman Catholic" was retained for the field test.
- All four types of Lutheran synods have been retained within the final version of 15(c) to ensure appropriate representation. However, these four options have been subsumed beneath the greater term "Lutheran" and an arrow has been provided to direct respondents who select "Lutheran" to the list of four Lutheran synods.
- Based on respondents' recommendations, the option "Gospel/Full Gospel" was added to the list of religious groups provided as options. The option "Greek Orthodox" was also replaced with the more inclusive term "Eastern Orthodox."

15(d): ***Which of the following categories best describes this school?***

- ❑ **Parochial (or inter-parochial)**
- ❑ **Diocesan**
- ❑ **Private**

Answer this question only if you marked "Roman Catholic" for question c above. Mark (X) only one box.

Response Issues: This item is designed exclusively for Roman Catholic respondents. Response options represent Roman Catholic organizational structures that frequently maintain, support, or sponsor the Roman Catholic private school. The respondent is directed to respond to 15(d) by an arrow stemming from the option "Roman Catholic" in 15(c) and extending along the left-hand margin to 15(d). Pretesting identified the following potential sources of response error:

1. Responses to 15(c) can affect responses to 15(d). Specifically, the arrow stemming from the option "Roman Catholic" in 15(c), which leads the respondent to 15(d) is frequently missed by respondents. Some respondents do not read the directions supplied beneath the stem in 15(d) which indicate 15(d) is designed exclusively for Roman Catholic respondents. Thus, some non-Catholic religiously-oriented and non-sectarian respondents selected "parochial" and/or "private" under 15(d), because they simply scanned the page, saw the word "private," interpreted the item as including all private school respondents, and endorsed either :parochial" or "private."

Final Modifications of Item 15(d): Item 15(d) was revised as specified below to address the potential sources of response error identified through pretesting:

- Item 15(d) will be eliminated altogether, and the response options beneath 15(d) will be moved to a new location beneath the option "Roman Catholic" within 15(c). By collapsing these options, Roman Catholic respondents will be able to proceed directly (by means of an arrow) to the additional set of options "parochial," "diocesan," and "private," after identifying themselves as Roman Catholic in 15(c). All nine private schools participating in Wave 3 testing agreed that this revision was an appropriate and more advantageous solution.

16: ***To which of the following associations or organizations does this school belong?***

Mark (X) all that apply.

Response Issues: PSS [16] requires respondents to reply to the question above using a list of 38 private schools associations from which to identify associations they belong to; respondents also have the option to select "None of the Above." This list is contains some informal grouping of types of related associations that are predominately in alphabetical order. There are also eleven "Other ... " options, such as "Other Montessori association(s)" and "Other religious school association(s)." All 38 association names are presented in the same type, consecutively, and without breaks. Pretesting identified the following potential sources of response error:

1. The list of associations was haphazardly organized and difficult to read. Some respondents had difficulty finding their associations.
2. "None of the Above" was on the bottom of the list and respondents who wanted to endorse this option were required to scan the entire list prior to finding this option.
3. Many respondents only knew the acronym for an association rather than the association's full name (e.g., AMS - American Montessori Society). This made completing [16] more difficult for some respondents, who struggled to remember the full name of their association.
4. One respondent recommended a line be added for each of the "other" options provided within the list of associations so that schools could write in the name of their school association. This recommendation was predicated on the assumption that private school respondents would like to feel they are a part of a recognized school association. Three educational experts favored this modification.
5. Some private school respondents identified additional associations to be included within the list. These additional associations were examined to evaluate whether a sufficient number of schools pertained to the association. Associations meeting this criteria were added to final list of associations provided in [16].

Sorting Task: A sorting task was used to evaluate whether the list could be divided into major section headings that would facilitate responding to [16]. The primary objective of the sorting task was to provide a categorization schema for the private school associations listed in [16] and simplify the task of completing [16]. The introduction of association category headings was hypothesized to simplify the task of completing [16] by allowing respondents to quickly scan the list of associations, and more readily discern where the associations they belong to are located. Category headings would also minimize reading required of respondents, decrease respondents' boredom and inattention, and decrease the likelihood that association(s) the school belongs to is/are missed by the respondent.

A set of sorting cards was created, each made of durable cardboard and labeled with the name of an association from the PSS list. Sorting cards were numbered from 1 to 38, and provided to respondents in the order of presentation currently used in the PSS form (e.g., 1, 2, 3,..., 38) to more closely approximate the actual task of completing [16].

Respondents were asked to sort the cards into clusters of associations representing logical and reasonable groupings. Respondents identified a mean number of 9.1 cluster names (range 4-14). Respondents exhibited several patterns when clustering associations. A common pattern was the identification of very small clusters of only two or three associations (e.g., bilingual, international, and alternative associations). Four respondents identified ten or more such cluster names.

Another clustering pattern that occurred frequently was the grouping of religiously-oriented private school associations by religion (e.g., Christian, Jewish, and Friends); this pattern emerged frequently from respondents from religiously-oriented private schools. Conversely, non-sectarian schools were more likely to group all religious associations together under a neutral and inclusive category name such as "Religious Orientation."

Some respondents classified a small number of specialized associations into clusters such as "Early Childhood," "Special Needs" or "Exceptional Education," or clusters with an educational or programmatic focus such as "Special Interest" and "Special Emphasis." Finally, several respondents

distinguished religiously-oriented associations from non-sectarian associations, which were labeled "Independent School Orientation" and "Other/Independent non-Religious."

The task of developing category headings requires that headings be sufficiently inclusive to allow several associations to be grouped within the association heading name. However, it is essential that the new categorization system be parsimonious, because the inclusion of too many headings can be cumbersome and detract from the primary purpose of simplifying the list of associations. Therefore, a maximum of three category headings were identified to simplify the task of completing [16].

Final Modification of Item 16: Item 16 was revised as specified below to address the potential sources of response error identified through pretesting:

- Three category headings were identified to simplify the task of completing [16]: 1) Religious; 2) Special Emphasis; and 3) Other School Associations or Organizations. Ten educational experts unanimously favored this revision, because it was easier for private school respondents to grasp. Category headings provided the respondent with a structure conducive to ease in responding, and facilitated visual acuity so that respondents would be less likely to miss their school association. Several experts reported respondents would save time by simply going to the association heading that contained the association(s) where they were members. Although the term "Other Private School Associations or Organizations" might potentially be offensive to some respondents, the majority of experts did not anticipate this would be a serious difficulty.
- Four additional associations were added to the list as recommended by respondents and experts.
- Acronyms for each association was placed in parentheses in capital boldface letters next to each association name, so that respondents that only recognized the association acronym would be able to quickly identify the association from the list. This feature allows respondents to scan the list of acronyms to find their association, and should improve the likelihood that respondents can find their association.
- A screening item was added at the top of the list to screen out private schools that were not members of any school association. This category would include private schools that 1) have not yet applied to any associations for membership; 2) cannot afford association membership dues; and 3) may not join any private school associations for religious or other reasons. The screening item was worded "This school/program does NOT belong to any association or organization [Mark X]."
- An "other" category was provided beneath each of the major categories, such as "Other Special Emphasis associations". This feature would allow private schools with memberships in educational associations not included in the greater list to provide a response so they would not feel unimportant. "Other" options were printed in italics to distinguish them from association names; this feature makes these options more readily visible.
- Include an "other" option at the bottom of the list with a line available for respondents to write-in their school association(s) to alleviate any possible feelings of inferiority that would result from failing to check any options provided in the list; this feature would also allow NCES to consult with the association and evaluate whether it should be included in latter versions of the PSS.

CONCLUSION

The Private School Survey is used to identify and categorize private schools in the United States. The first year of testing of the PSS items resulted in the implementation of several modifications to the original PSS items. These modifications were identified through suggestions made by participating private school respondents and educational experts during the early Waves of testing. Revisions were refined based upon feedback received during latter waves of testing.

Currently, a field test of the revised PSS items is underway with a proportional representative sample of approximately 1800 private schools throughout the United States. The results of the field test will be compared to responses generated by respondents who completed the PSS form during the 1996 administration. The comparison group will be drawn in a manner to ensure its statistical comparability with the field test sample. The results of field testing are expected to yield a final set of items that will be incorporated into the Private School Survey and used to categorize private schools throughout the United States.

The opinions expressed in this paper are those of the authors and do not necessarily represent those of the Bureau of Labor Statistics or the National Center for Education Statistics.

AN ANALYSIS OF RESPONSE RATES OF SASS 1993-94

Sameena M. Salvucci, Fan Zhang, Mingxiu Hu, David Monaco, Synectics for Management Decisions, Inc., Kerry Gruber, National Center for Education Statistics
Fan Zhang, Synectics for Management Decisions, Inc., 3030 Clarendon Blvd., Arlington, VA 22201

Key Words: Complex sample surveys, Unit nonresponse, Hierarchical nonresponse, Multivariate modeling, Cooperation rates, Sampling frame.

1. Introduction

This paper addresses the most pervasive and challenging source of nonsampling error in estimates from sample surveys which is the error associated with incomplete data. Incomplete data resulting from three sources are of particular importance in sample surveys: item nonresponse, unit nonresponse, and undercoverage.[1] The concern for nonresponse, whether item or unit, is twofold. Nonresponse reduces the sample size and thus increases the sampling variance. Respondents may also differ significantly from nonrespondents, thus, the estimate obtained from respondents can be biased and the magnitude of this bias may be unknown. Concerns about bias are generally greater as the rate of nonresponse increases.

The particular focus of this paper is to quantify the extent of unit nonresponse in the 1993-94 Schools and Staffing Survey (SASS) conducted by the National Center for Education Statistics (NCES) and to assess the impact of differences in the known characteristics of respondents and nonrespondents for different subgroups of the survey populations in order to provide some indication of the *potential* effects of nonresponse bias. The results of this study can be used to further control and adjust survey estimates for bias, and improve survey operations. While the scope of this paper is chiefly descriptive, inferential modeling of the response rates is also provided as an example for future SASS research.

2. 1993-94 SASS

The 1993-94 SASS is the third study of public and private elementary and secondary schools in a series of surveys begun in 1987-88 by NCES. Survey data from schools, local education agencies (LEAs), administrators, and teachers in the United States were collected by mail with telephone follow-up of nonrespondents first during the 1987-88 school year and again during the 1990-91 and the 1993-94 school years. The series provides data on school and teacher characteristics, school operations, programs and policies, teacher demand and supply, and the opinions and attitudes of teachers and school administrators about policies and working conditions. The analytic power of the data is enhanced by the ability to link survey data for individual LEAs, schools, administrators, and teachers. In 1993-94 new library, librarian and student SASS components were initiated that could also be linked. In addition, computer assisted telephone interviewing (CATI) facilities were introduced for the first time during the 1993-94 SASS and were used extensively for nonresponse follow-up.

The 1993-94 SASS consists of thirteen components: the School Surveys, the School Administrator Surveys, the Teacher Surveys, the Teacher Demand and Shortage Survey, the Library Surveys, the Librarian Surveys, and the Student Record Surveys. Some 13,000 schools and administrators, and 67,000 teachers were selected. In addition, 5,500 local education agencies associated with the selected schools and 100 districts not associated with schools were selected in the 1993-94 SASS. Some 7,600 libraries and librarians, and 6,900 student records were also selected. Details pertaining to the frame, stratification, and sample selection for each of the survey components are presented in Abramson *et al.* (1996).

3. Weighted Unit Response Rates

For each survey of SASS, weighted unit response rates were calculated. The weighted response rates were derived by dividing the sum of the basic weights for the interview cases by the sum of the basic weights for the eligible cases (the number of sampled cases minus the out-of-scope cases). The basic weight for each sample case was assigned at the time of sampling as the inverse of the probability of selection.

In the first stage of this study we tested whether there is a significant difference between respondents and nonrespondents for a range of characteristics. To make this kind of inference to the underlying population based on SASS data, the conventional Pearson chi-squared statistic is not appropriate anymore. Tests were performed using a modified Pearson test statistic called Rao-Scott3 (RS3)[2]. A

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

[2] Rao, J. and Scott, A. (1981). "The Analysis of Categorical Data from Complex Sample Surveys: Chi-squared Tests for Goodness of Fit and Independence in Two-way Tables." *Journal of the American Statistical*

statistical software package called the WesVarPC® 2.0 provides a convenient procedure for this purpose. WesVarPC® not only calculates the weighted response rates, the standard error, the sample size and design effect, but also provides the Rao-Scott3 statistic that reflects the complex sample design.

Within the public and private sectors, the results of the significance tests are fairly uniform. That is, if respondents are significantly (or not significantly) different for a variable for a public sector survey component then they are likely significant (or not significant) for other public sector components as well. There are some interesting differences in results of the significance tests when compared across the public and private survey components. The most striking contrast in results exists for the variable "school sampled in 1990-91 SASS". Whether or not a school was surveyed in the 1990-91 SASS proved not to be significant for all of the public school components, while it was significant for all of the private school components.

Some interesting patterns arise when response rates are looked at across surveys. Tables 3.1 (public components) and 3.2 (private components) show the tests results and rankings of response rates for different levels of a selected set of variables that are common across all surveys. For public components the response rates for minority enrollment and urbanicity show some very strong patterns. Schools with a minority enrollment greater than 50.5 percent had the lowest response rates for all public components except the Student Record Component. Furthermore, minority enrollment showed a significant association with response status for all public components except the Student Record Component. Urbanicity showed a very strong pattern, with rural/small towns having the highest response rate, followed by urban fringe/large towns, and then central cities with the lowest response rates. This pattern was the same for all components except the Student Record Component. But, as with minority enrollment, urbanicity showed a significant association with response status for all public components except the Student Record Component. For private components the response rates for region and school size show some patterns. The Midwest region had the highest response rates for all private components except the Student Record Component. While response rates for schools with 1 to 149 students were always the lowest for all private components except the Student Record Component. Similar to the public side region and school size showed a significant association with response for all private components except the Student Record Component.

The rankings, when viewed across the public and private components, show two variables with similarities - school level and urbanicity. For school level, eight of the 12 public and private components have secondary schools with the highest response rate followed by elementary schools, and then combined schools. School urbanicity also showed a fairly strong ranking pattern, where eight of the 12 public and private components have schools in rural/small towns with the highest response rate followed by those in urban fringe/large towns, and then those in central cities.

4. Hierarchical Response Patterns

In the second stage of this study we examined the hierarchical nature of the nonresponse in the 1993-94 SASS. The aim was to find out about the *jointness* of nonresponse; for example to learn whether administrators in responding schools are more or less likely to respond than administrators in nonresponding schools. Specifically, we tested to see if there is a significant difference in response rates of each of the following types of respondents: (1) public and private school administrators, (2) public and private schools, (3) public and private school teachers, (4) public and private school libraries, (5) public and private school librarians, and (6) local education agencies (LEAs) when "linked" with the response status of other SASS components.

The results indicated that all units in the 1993-94 SASS (e.g. administrators that are "linked" within other units such as schools) are more likely to respond when the "linked" unit responds and in a large number of cases the difference in response is significant

5. Components of Nonresponse/Cooperation Rate

To measure the ability of a survey to establish contact with sampled units, the reasons for nonresponse are important In the 1993-94 SASS, three categories of reasons were recorded: 1) **refusal**, the nonrespondent refuses to take part in the survey; 2) **unable-to-contact**, contact with the nonrespondent was not able to be made through the nonresponse followup procedures; and 3) **other**, for example the questionnaire was not returned or the questionnaire was returned but it was incomplete.

Two teacher components had very high unable-to-contact rates (17.9%). For those instances with high unable-to-contact rates it is sensible to look at a **cooperation rate**, which is the response rate given the cases can be contacted. The cooperation rate is the number of interviews divided by the number of eligible cases contacted: Cooperation Rate = Interview /

Association, 76: 221-230. Rao, J. and Scott, A. (1984). "On Chi-squared Tests for Multiway Contingency Tables with Cell Proportions Estimated from Survey Data." *The Annals of Statistics*, 12: 46-60.

(Interview + Refusal + Other). Compare with the Response Rate = Interview / (Interview + Refusal + Unable to Contact + Other). The advantage of using the cooperation rate is that it controls for differences due to the unable-to-contact cases. Using the cooperation rate will eliminate the confounding effect associated with unable-to-contact cases.

To illustrate this confounding effect the significance tests for the teacher components were calculated using the cooperation rates since their unable-to-contact rates are the highest of all the components and the difference between their response and cooperation rates were among the highest (see Tables 5.1 and 5.2 below)

Table 5.1 -- Weighted response and cooperation rates: Public School Teacher Component (Rates in percent)

Variable	Response Rate	Cooperation Rate
School Type		
Regular	88.26	89.25
Non-regular	86.25	88.42

Table 5.2 -- Weighted response and cooperation rates: Private School Teacher Component (Rates in percent).

Variable	Response Rate	Cooperation Rate
Urbanicity		
Rural/Small town	83.10	85.38
Urban/large town	80.41	82.48
Central City	78.79	82.64
New Teacher		
Yes	81.02	85.16
No	80.05	82.76

The tests results for the public teacher component indicated that using cooperation rates the variable school type was not significant anymore (see table 5.3). The reason for this is that the low response rates for the non-regular schools is due to a higher unable-to-contact rate than regular schools.

Table 5.3 -- P-value of the Independence Test: Public School Teacher Component

Variable	P-value based on Response Rate	P-value based on Cooperation Rate
School Type	0.0172	0.1092

For the private school component, the lower response rate in central cities is due to a high unable-to-contact rate. After adjusting for this, by removing the unable-to-contact cases, urbanicity is not significant (see table 5.4). On the another hand, there is a high unable-to-contact rate for new teachers and that caused a low response rate for the new teachers. After the unable-to-contact cases are removed, the new teachers have a significantly higher cooperation rate than the others and the variable new teacher becomes significant.

Table 5.4 -- P-value of the Independence Test: Private School Teacher Component.

Variable	P-value base on Response Rate	P-value based on Cooperation Rate
Urbanicity	0.0094	0.0923
New Teacher	0.3206	0.0107

6. Multivariate Model (Public School Component)

In the last stage of our study we assessed the multivariate adjusted effects (on the response rates) of the significant variables which were identified in the first stage of our study (see section 3). We fitted multivariate logistic regression models. In this section three issues will be discussed: model selection, model interpretation, and comparisons of univariate unadjusted results with multivariate adjusted results.

In our model selection, we began with the following potential variables which were considered in the univariate analysis: *urbanicity*, *region*, *school level*, *school size*, *school type*, *minority enrollment*, *sampled with certainty*, *submitted a teacher list*, *source*, and *sampled in the 1990-91 SASS*. Variable *submitted a teacher list*, which has the most significant effect on the school nonresponse, is eliminated from the multivariate model due to interpretation difficulties. This variable is more like a questionnaire variable rather than a design variable in terms of the time we observe the variable. We can not use it to predict the probability of school response. The variables *sampled in the 1990-91 SASS* and *sampled with certainty* are dropped from the model due to their ignorable contribution to the model.

The software package WesVarPC® was used to fit the multivariate logistic regression model with the seven selected independent variables as well as seven separate univariate logistic regression models for those variables. Table 6.1 presents a comparison of the p-values for the Rao-Scott, univariate logistic regression model, and multivariate logistic regression model tests.

It is noted that p-values for the Rao-Scott test (Rao and Scott, 1984, or RS3 in WesVarPC output) and the univariate logistic regression are pretty close. The only significant difference between these two tests is for variables *source* and *school type*, but their p-values are still comparable. If we test the hypotheses at 0.01 level, both tests will reach the same conclusion of significance.

However, the multivariate logistic model test results are very different from Rao-Scott test results and the univariate logistic model test results, especially for variables *minority enrollment*, *urbanicity*, *region*, and *school type*. Variables *minority enrollment*, *urbanicity*, and *region* are highly significant in the Rao-Scott tests and the univariate logistic regression model, but they are not significant at all, with high p-values of 0.3936,

0.1016 and 0.1115, respectively, in the multivariate logistic regression model when we adjust for other variables simultaneously. This happens because there exists an antagonism effect on these variables. We must take into account of this antagonism effect when we interpret the effects of minority enrollment, urbanicity and region. The significant effects of these three variables shown in the univariate analysis are simply caused by imbalance of the other significant variables among these three variables. On the other hand, there exists an synergysm effect on variable *school type*. In the multivariate logistic regression model, *school type* is significant (at 0.01 level) with a p-value of 0.0047, but it is not significant in the univariate logistic regression model or by Rao-Scott test with p-values of 0.0397 and 0.0719, respectively. That means that some information about school nonresponse provided by school type is covered by the noise of other factors. We must retrieve that part of information by adjusting to other factors simultaneously through a multivariate model.

For the other three variables, *school size*, *school level* and *source*, it seems that there is neither an antagonism effect nor a synergism effect. The univariate results are almost identical to the multivariate results for those variables. The four-level variable *school size* is the most significant variable for explaining the variation of school nonresponse.

The entropy for the multivariate model is 2.13%. As pointed out by the documentation for WesVarPC®, this entropy may not be appropriate to measure the strength of the association.

Table 6.2 presents the parameter estimates, standard errors, odds ratios and p-values of the tests for the dummy variables which represent the independent variables in the multivariate logistic regression model. The parameter estimates and odds ratios describe the nature of the association between the school nonresponse and the selected independent factors.

We found that the response rate of a rural/small town school is barely significantly higher than a central city school with an odds ratio of 1.377, although the overall factor urbanicity is not significant with a p-value of 0.1016. *School level*, *school size*, and *school type* are all significant factors for the school nonresponse. A combined school and an elementary school are both less likely to respond to the survey than a secondary school; a smaller school is more likely to respond than a larger school. The odds ratio comparing a school with enrollment between 1 and 149 students and a school with an enrollment of 750 or more students is 2.316; non-regular school is less likely to respond than a regular school with an odds ratio about one-half. However, *Minority enrollment*, *region* and *source* has no significant effect on the school nonresponse.

We also fit a reduced model which eliminates all dummy variables that are not significant at 0.1 level in the full model. The results of the reduced model for school level, school size, school type and source are almost identical to those in the full model presented in Table 8.2. However, in the reduced model, urbanicity is highly significant with P-value of 0.0063, but this p-value is for the testing the difference of the school response rate between rural/small town schools and all other schools. Similarly, the test to compare the Midwest vs the other three regions is barely significant at 0.05 level.

In summary, we find that school size, school level, and school type are the only three factors which have a significant effect on school nonresponse. Neither the three-level variable urbanicity or the four-level region variable have significant overall effects, but a rural/small town school has a significantly higher probability to respond than an urban fringe/large town or a central city school, and the Midwest has significantly higher response rate than other regions. Minority enrollment, which is highly significant in the univariate model, is not significant at all in the multivariate model. The sample frame source "CCD update" is a little better than other three sources (close to significant), but the other three sources are not significantly different at all.

7. Conclusions

Results of assessing the differences in known characteristics of respondents and nonrespondents for different subgroups of the sampled populations indicated that patterns of nonresponse among characteristics such as region, urbanicity, school level, and school size persisted from the 1990-91 survey round to the current round. For example, response rates for rural/small town public schools were the highest and response rates for central city schools were the lowest in both 1990-91 and 1993-94. In addition, a set of characteristics including some of those mentioned above, whether a school submitted a teacher list, and minority enrollment in a school were shown to have significant differences between respondents and nonrespondents.

One of the more striking results of our analysis pertain to the examination of whether response patterns in a survey component are hierarchically associated with response patterns of linked components. Our analysis showed response rates were higher among linked responding units versus linked nonresponding units. For example, response rates for LEAs were higher for those LEAs linked with responding schools versus those linked with nonresponding schools.

Response rate components (e.g. out-of-scope rates, refusals, non-locatables, etc.) were examined in an attempt to provide tools to monitor the quality of the SASS frame and the corresponding 1993-94 SASS survey statistics. Out-of-scopes rates for the public components were lower than their private counterparts. In addition, out-of-scope rates for the private school library component, and both the public and private school librarian components were quite high. Cooperation rates for components with high non-locatable rates, such as the teacher component, were calculated and tested. In most cases, for the teacher survey, the significance results were different for cooperation rates versus response rates indicating that the unable to contact cases had a confounding effect on the results of these significance tests.

Finally, the results of fitting a multivariate logistic regression nonresponse model for the public school component were compared to the univariate level significance results. School size, school level, and school type were the three factors shown to jointly have a significant effect on school nonresponse. These results show that the effects of some variables on the response status can be explained by the other variables hence a reduced model is preferable. The model results can be used to adjust weights for nonresponse.

References

Abramson, R., Cole, C., Jackson, B., Parmer, R., and Kaufman, S. (1996), "1993-94 Schools and Staffing Survey: Sample Design and Estimation," technical report, National Center for Education Statistics, Washington, DC.

Jabine, Thomas B. (1994), "Quality Profile for SASS", NCES 94-340, National Center for Education Statistics, Washington, DC.

Rao, J. and Scott, A. (1984), "On chi-squared tests for multiway contingency tables with cell proportions estimated from survey data," *The Annals of Statistics*, 12, 46-60.

Scheuren, F., Monaco, D., Zhang F., Ikosi, G., Chang, M., and Gruber, K. (1996), "An Exploratory Analysis of Response Rates in the 1990-91 Schools and Staffing Survey," forthcoming report, National Center for Education Statistics, Washington, DC.

Westat, Inc. (1996), "A User's Guide to WesVar PC®," Westat Inc., 1650 Research Boulevard, Rockville, MD 20850.

Table 3.1 -- Public component response rate ranks: Schools and Staffing Survey 1993-94, Public Administrator, School, Teacher, Library, Librarian, and Student Components.

Component	Administrator	School	Teacher	Library	Librarian	Student
Variable						
Minority Enrollment (test result)	S	S	S	S	S	NS
Less than 5.5%	1	1	1	2	3	1
5.5 - 20.5%	2	2	2	1	1	4
20.5 - 50.5%	3	3	3	3	2	2
Greater than 50.5%	4	4	4	4	4	3
Region (test result)	S	S	S	NS	NS	S
Midwest	1	1	1	2	2	1
Northeast	3	4	4	3	1	3
South	2	2	2	1	3	2
West	4	3	3	4	4	4
School Level (test result)	NS	S	NS	S	S	NS
Elementary	3	2	2	2	2	2
Secondary	2	1	1	1	1	3
Combined	1	3	3	3	3	1
School Size (test result)	S	S	S	S	NS	NS
1 to 149	1	1	1	4	4	1
150 to 499	2	2	2	3	2	3
500 to 749	3	3	3	1	3	2
750 or more	4	4	4	2	1	4
Urbanicity (test result)	S	S	S	S	S	NS
Rural/small town	1	1	1	1	1	2
Urban fringe/large town	2	2	2	2	2	3
Central City	3	3	3	3	3	1

"S" indicates a significant association between respondents and nonrespondents for the different levels of the variable.
"NS" indicates that there is not a significant association between respondents and nonrespondents for the different levels of the variable.

Table 3.2 -- Private component response rate ranks: Schools and Staffing Survey 1993-94, Private Administrator, School, Teacher, Library, Librarian, and Student Components.

Component	Administrator	School	Teacher	Library	Librarian	Student
Variable						
Region (test result)	S	S	S	S	S	NS
Midwest	1	1	1	1	1	2
Northeast	3	3	3	4	4	3
South	4	2	2	2	3	4
West	2	4	4	3	2	1
School Level (test result)	S	S	S	S	S	NS
Elementary	2	1	2	2	2	1
Secondary	1	1	1	1	1	2
Combined	3	3	3	3	3	3
School Size (test result)	S	S	S	S	S	NS
1 to 149	4	4	4	4	4	3
150 to 499	3	1	2	3	3	2
500 to 749	2	3	1	2	1	4
750 or more	1	2	3	1	2	1
Urbanicity (test result)	NS	NS	S	S	NS	NS
Rural/small town	3	1	1	1	3	2
Urban fringe/large town	2	2	2	2	2	3
Central City	1	3	3	3	1	1

"S" indicates a significant association between respondents and nonrespondents for the different levels of the variable.
"NS" indicates that there is not a significant association between respondents and nonrespondents for the different levels of the variable.

Table 6.1 -- P-values for Rao-Scott, univariate logistic regression model, and multivariate logistic regression model tests (Public School)

Variable	Rao-Scott (RS3)	Univariate Model	Multivariate Model
Urbanicity	0.0001	0.0001	0.1016
Region	0.0030	0.0109	0.1115
Minority Enrollment	0.0002	0.0002	0.3936
Source	0.0175	0.0605	0.0746
School Level	0.0100	0.0119	0.0116
School Size	0.0000	0.0001	0.0001
School Type	0.0719	0.0397	0.0047

Table 6.2 -- Parameter Estimate, Odds Ratio and P-value: Public School.

Pairwise Comparison	Parameter Estimate	Standard Error	Odds Ratio	P-value
Urbanicity				
Rural/small town vs Central City	0.32	0.154	1.377	0.0410
Urban fringe/large town vs Central city	0.07	0.159	1.073	0.6600
Region				
Midwest vs West	0.29	0.162	1.336	0.0844
Northeast vs West	-0.12	0.179	0.887	0.4928
South vs West	0.18	0.136	1.197	0.2019
Minority Enrollment				
Less than 5.5% vs Greater than 20.5%	0.16	0.136	1.174	0.2458
5.5-20.5% vs Greater than 20.5%	0.11	0.119	1.116	0.3615
Source				
CCD update vs others	0.46	0.254	1.584	0.0746
School Level				
Combined vs Secondary	-0.36	0.156	0.698	0.0268
Elementary vs Secondary	-0.23	0.089	0.795	0.0140
School Size				
1 to 149 vs 750 or more	0.84	0.156	2.316	0.0000
150 to 499 vs 750 or more	0.40	0.120	1.492	0.0015
500 to 749 vs 750 or more	0.30	0.152	1.350	0.0543
School Type				
Non-regular vs Regular	-0.68	0.229	0.507	0.0047

AN OVERVIEW OF NCES SURVEYS REINTERVIEW PROGRAMS

Valerie Conley, Steven Fink, Mehrdad Saba, Synectics for Management Decisions Inc.,
Steven Kaufman, National Center for Education Statistics
Valerie Conley, Synectics for Management Decisions, Inc., 3030 Clarendon Blvd #305, Arlington VA 22201

KEY WORDS: Reinterview Programs, Simple Response Variance, Gross Difference Rate, Index of Inconsistency, Response Bias

1. Introduction

The National Center for Education Statistics (NCES) conducts a variety of programs to assess the quality of the data it collects in its surveys. Traditionally, the emphasis has been to estimate nonsampling error components in a survey model: reinterview programs and validity evaluations are part of the overall survey design for most of its complex sample surveys. This paper highlights the NCES application of the reinterview while serving also as an overview of the techniques and methods that quantify measurement error which are used in NCES data quality assessment.

2. Programs Purpose and Background

A reinterview -- replicated measurement on the same unit -- is a new interview which repeats all or a subset of the original interview questions. At the 1991 American Statistical Association (ASA) meeting, the National Agricultural Statistics Service (NASS) presented a paper which traced the history of reinterview studies at NASS. The authors concluded that "An important product of reinterview surveys has been the identification of reasons for reporting errors. These include definitional problems, misinterpretation of questions and survey concepts, and simple reporting errors. Such cognitive information obtained from reinterviews has been valuable in survey instrument development, training, and the interpretation of survey results" (Hanuschak et al., 1991).

The purpose of NCES reinterview programs is:
1) to determine how good questions are with respect to a measurement of the response error and 2) to assess the quality of the data collected. The extent of the research effort varies across surveys from small sample reinterview programs conducted as part of a survey questionnaire field test to larger samples that range between 1 and 11 percent of a full-scale study. These programs have been used for three major purposes:

- *Identifying specific questions that may be problematic for respondents and result in high variability*
- *Quantifying the magnitude of the measurement error*
- *Providing feedback on the design of questionnaire items for future surveys*

Specifically, the purpose of the reinterview is to gain insight into the adequacy of questions. This gain can be achieved analytically by measuring two components of survey response -- response variance and response bias. These two measures are explained in more detail in section 5.

Another common purpose of reinterview programs is to verify that the original interviews were genuine. NCES often uses a combination of mail and Computer Assisted Telephone Interview (CATI) for its surveys. Most of the NCES reinterview programs were done using Computer Assisted Telephone Interview (CATI) in a centralized setting. Since the CATI interviews were closely monitored, it is highly unlikely that a telephone interviewer could invent or falsify interviews. Therefore, this aspect of reinterview is not typically a focus of NCES reinterview programs.

3. Surveys and Reinterview Design

Several NCES surveys have conducted reinterview programs for more than one round or cycle of the survey, specifically Baccalaureate and Beyond (B&B), Beginning Postsecondary Students (BPS), National Household Education Survey (NHES), Schools and Staffing Survey (SASS), and Teacher Follow-up Survey (TFS). Most of the programs do not include the same items on subsequent rounds of the reinterview, however. The BPS reinterview programs, for example, are designed "to build on previous analyses by targeting revised or new items not previously evaluated" (Burkheimer et al. 1992). Most of the NCES reinterview programs were developed to estimate response variance, but some, such as the Adult Education component of NHES:95, included a response bias study as well. However, the surveys that involved

testing, such as the National Adult Literacy Survey (NALS), the National Assessment of Educational Progress (NAEP), and the National Educational Longitudinal Study (NELS) never considered retesting or reinterviewing because NCES considered such activities too much of a burden on the respondents.

The following issues are considered in NCES reinterview designs:

- *Time Lag*

Time lag between the original and the reinterview for most of NCES surveys is usually stated as a range of days or weeks following the original survey such as "the reinterviews were conducted in October and November, about 4 to 6 weeks after the original interview" (Brick, Cahalan et al., 1994, p. 3-3). Early Childhood Education (ECE) reinterviews for NHES were designed for 14 days after the completion of the original ECE interview, but they were actually completed between 14 and 20 days after the original interview. BPS reinterviews were conducted up to 8 weeks after completing the original interview, and reinterviews in NPSAS were conducted between one and three months after the original interview.

- *Reinterview Instrument*

The reinterview instrument is a subset of the original questionnaire, but the question wording is almost always identical between the original and the reinterview instrument. In some cases, however, adjustments were made to the question wording in an attempt to gain more reliable data such as National Survey of Postsecondary Faculty (NSOPF). This is most often the case if the reinterview is conducted as part of the field test and not as part of the full-scale study. Other case examples are National Postsecondary Student Aid Study (NPSAS) and B&B.

- *Mode*

The mode of reinterviews is usually telephone regardless of the original interview mode. Conducting all the reinterviews by telephone violates survey error model assumptions that require the reinterview to be an independent replication of the original interview in order to estimate response variance accurately. Therefore, SASS included research in its 1991 reinterview program to determine the impact mode change might have on data quality. Most of the mail respondents were reinterviewed by mail and the telephone follow-up cases were reinterviewed by telephone. Generally reinterviews conducted by mail showed relatively lower response variance than the telephone reinterviews (Royce, 1994).

4. Sample Size and Response Rates

The design of NCES reinterview programs typically includes a target number or percentage of completed reinterviews. The reinterview sample size for RCG:91, for example, was 583 with a goal of 500 completed reinterviews. SASS reinterviews 10 percent of the School and Administrator samples and one percent of the Teacher sample to have a reinterview sample of just over 1,000 for each of its components. The reinterview sample sizes are considerably smaller when the reinterview program is conducted as part of the field test. The reinterview sample size for the NPSAS 1992-93 field test, for example, was 237.

NCES reinterview response rates vary from a low of 51 percent to a high of 94 percent with most reinterview programs having response rates in the mid-eighties to low nineties.

5. Measurement Error Estimation

Reinterview programs at NCES tend to measure the response variance and response bias using simple measures of consistency. Response variance is a component of measurement error which examines how consistently respondents answer questions in a survey. Response bias, on the other hand, measures the systematic nonsampling errors. In order to estimate response variance and response bias, it is necessary to define a general measurement error model:

Let $Y_{t1} = 1$ if Yes is recorded for unit t in the original interview and $Y_{t1} = 0$ otherwise.

Let $Y_{t2} = 1$ if Yes is recorded for unit t in the reinterview and $Y_{t2} = 0$ otherwise.

For unit t (t = 1,2,...,n) and the ith measurement (i = 1,2), the assumed model is:

$$Y_{ti} = X_t + B_{ti} + e_{ti}$$

Where Y_{ti} is a Bernoulli random variable, X_t is the "true" value of unit t, assumed unchanged between measurements, B_{ti} is the response bias, and e_{ti} is a random measurement error.

To illustrate the model consider the crossclassification of two measurements of an individual population characteristic (for example, whether a person who participates is a college graduate) obtained from an original interview and a reinterview of the same sample of individuals. Table 1 shows the crossclassification (Brick, Cahalan et al., 1994):

Table1: Two measurement crossclassification

Reinterview	**Original Interview**		
Response	Yes	No	Total
Yes	a	b	a + b
No	c	d	c + d
Total	a + c	b + d	n = a + b + c + d

In their simplest form, reinterview results are analyzed using measurements derived from this crossclassification. These measures include the overall deviation between the interview and the reinterview, deviations on individual responses, and the index of crude agreement. The three specific measures most commonly used by NCES are:

- Gross Difference Rate (GDR)
  -Measures the weighted percentage of cases reported differently in the original and the reinterview as
  $$(b + c) / 2n$$

- Index of Inconsistency (IOI)
  -Estimates the proportion of total survey variance due to simple response variance as
  $$n(b + c) / 2(a + c)(b + d)$$
  -Assumes simple random sampling with replacement

- Net Difference Rate (NDR)
  -Computed after reconciliation for each answer category of a question
  -Weighted difference of the false positive and false negative rates calculated as
  $$(c - b) / n$$

Typically, the first two measures, GDR and IOI, are used to estimate simple response variance. The third measure, NDR, is used to estimate response bias, which most measurement models at NCES assume to be constant in the repeated measurement. NDR is also used to test the independence of the reinterview.

6. NCES Reinterview Results

This section will present a summary of selected results from reinterview programs conducted by NCES. Only results from reinterview programs conducted as part of full-scale studies are provided here. These studies include NHES, RCG, and SASS.

National Household Education Survey (NHES)

The National Household Education Survey (NHES) is designed to collect education data from U.S. households through telephone interviews, using random digit dialing (RDD) and computer-assisted telephone interviewing (CATI) procedures. The sample is drawn from the noninstitutionalized civilian population in households having a telephone in the 50 states and the District of Columbia.

NCES has conducted four comprehensive reinterview programs for the full-scale NHES surveys. The reinterview program for NHES:91 was administered only on the early childhood (ECE) component. In NHES:93 both components underwent reinterviews, while only the adult education component was reinterviewed for NHES:95.

All NHES reinterview programs have used gross difference rate (GDR) and net difference rate (NDR), and all except NHES:95 used index of inconsistency (IOI), as measures of response variability and response bias for critical items in the surveys.

The NHES:91 reinterview results suggested that the ECE interview measured some variables with relative success, but it also revealed some items needed to be handled carefully when tabulating findings and for which alternative methods of collection should be considered (Brick et al., 1991).

The early childhood component of NHES:91 reinterview program included questions on current enrollment (whether the child was attending school and, if so, what grade) and home environment (reading and television habits). All of the seven enrollment items had low GDRS and IOIs. Of the four home environment variables there were two worth noting:

- P19/E36: How often do you or other family members read stories to (child)?
- P22/E40: How many hours each day does (child) watch television or videotapes?

Brick et al. (1991) suggested the relatively large IOI (42.0) for the television question might be due to the "general ambiguity in the item, the crude measurement scale (whole hours) relative to the internal variability in the item, and differing circumstances" (p. D-20). The reading question also had a relatively large GDR (23.3) and IOI (33.5).

The two topical components of NHES:93 were: 1) the School Readiness (SR) interview of parents of children (ages 3-7 and 8-10) enrolled in second grade or below and 2) the School Safety and Discipline (SS&D) interview of parents of students enrolled in grades 3-12 and youths enrolled in grades 6 through 12. The subset of the original SR and SS&D questionnaire items chosen were selected because they were substantively important, not highly time dependent, and not examined in the NHES:91 reinterview. The reinterview sample sizes were substantially increased from the 604 of the NHES:91 reinterview program to 977 for SR and 1,131 for SS&D in order to obtain more reliable estimates of the response variance for key questions. The reinterview did not reveal any items with response problems that were severe enough to cause researchers to question analysis based on the item.

The NHES:95 reinterview program examined and estimated measurement errors as components of nonsampling error in the Adult Education (AE) survey. A subset of items from the original interview were selected and the original and reinterview responses were then compared to estimate the consistency of reporting. Interviews were sampled at different rates for participants and nonparticipants (i.e., people who did not participate in adult education activities), with a total of 1,289 cases selected for reinterview.

The GDRs for the NHES:95 reinterview programs were low for the adult education participation and the education background items, indicating that responses to those questions were consistent. The GDRs for barrier to participation items (such as obstacles that prevented respondents from adult education activities) were much higher than for the other subject areas, indicating that responses were not consistent. Only four (out of 15) barrier items had GDRs of less than 10 percent, and the highest GDR approached 50 percent. This inconsistency may have been related to factors like recoding the questions, additional eligibility criteria, and small sample sizes. Nonetheless, barrier items had some response problems and did not appear to be reliable.

NCES also conducted a separate study to measure bias for NHES:95. The methodology used for this survey appeared to have potential for detecting biases; however, this method -- intensive interviews -- was not as successful as the standard NHES reinterviews for estimating consistency of reporting.

Recent College Graduates (RCG)

The 1991 Survey of Recent College Graduates (RCG:91) provides data on the occupational and educational outcomes of bachelor's degree and master's degree recipients one year after graduation. Telephone interviews were conducted between July 1991 and December 1991 using computer-assisted telephone interviewing (CATI).

Two measurement error models were estimated from the reinterview data. The first model (the simple response variance model) assumed the errors were all from random sources. This model was then expanded to allow for systematic errors or biases. Both models assumed that the interviewers were not a source of systematic error in the data collection process, but the first assumed that the measurement errors were the same across sampled graduates and from one trial to the next. Thus, if the reinterview was uncorrelated with the original interview, then the number of original and reinterview errors should be roughly equal.

Of the 16 reinterview items in these categories, only two had an IOI greater than fifty percent. One item was related to employment experience, while the other was a question dealing specifically with teacher certification and employment:

> Q24: Were you looking for work during the week of April 22, 1991?
>
> Q62: Prior to completing the requirements for your 1989-90 degree, were you at any time employed as a school teacher at any grade level, from prekindergarten through grade 12? Please exclude student or practice teaching and work as a teacher's aide.

Question 24 was asked only of a subset of the sample of graduates -- those who were unemployed. The reduced sample size may have contributed to a larger GDR and IOI. There were also potential recall problems since the question referred to a specific

period of time. For question 62, no explanation was offered for the possibly high random measurement error.

The overall conclusions of Brick, Cahalan et al. were that even though measurement errors were an important source of error in RCG:91, the estimates from the survey were not greatly distorted by these errors. The relatively small GDRs indicated responses were consistent; however, the IOIs being generally moderate implied that improvements in questionnaire wording and construction might help to reduce measurement errors in future surveys.

Schools and Staffing Survey (SASS)

The Schools and Staffing Survey (SASS) is a periodic, integrated system of surveys designed to collect data on characteristics of public and private school teachers, administrators, and their workplaces. The first two rounds of SASS (1987-88, and 1990-91) included the School Survey, the School Administrator Survey, the Teacher Demand and Shortage Survey (TDS), the Teacher Survey, and, one year later each time, the Teacher Follow-up Survey. SASS includes reinterview programs as part of its survey design, although it has used other methodologies for measuring response variance and response bias as well, including validation studies, such as the Teacher Listing Validation Study (TLVS), and follow-up cognitive research.

The following part of this paper summarizes overall results and comparison of the 1987-88 and 1990-91 reinterviews. There was no difference in response variance between public and private administrators, schools, or teachers.

Thirty-nine percent of the 1990-91 SASS reinterview questions showed low response variance. This was significantly better than the 11 percent of reinterview questions for SASS 1987-88 with low response variance (see table 2). Moreover, there was a 23 percentage point difference between 1990-91 and 1987-88 SASS items with a high response variance (26 percent versus 49 percent) (Royce, 1994).

It is important to note that the results across 1987-88 and 1990-91 are not strictly comparable. Different sets of questions were used for the two interviews. Among the 15 factual questions common to both years, 11 showed significant revisions in 1991. Four of these items displayed reduced response variance, which indicates improvement in these questions (Bushery, Royce and Kasprzyk, 1992, p. 459).

Table 2. Summary of 1987-88 and 1990-91 SASS reinterview response variance results*

	Low		Moderate		High	
	Number	**Percent**	**Number**	**Percent**	**Number**	**Percent**
All three components:						
1988	4	11%	14	40%	17	49%
1991	43	39%	38	35%	28	26%
Administrator Survey:						
1988	1	11%	4	44%	4	44%
1991	5	20%	10	40%	10	40%
Teacher Survey:						
1988	3	25%	4	33%	5	42%
1991	21	44%	16	33%	11	23%
School Survey:						
1988	0	0%	6	43%	8	57%
1991	17	47%	12	33%	7	19%

*Questions for which index could be reliably estimated.

SOURCE: Royce, D. (1994), *1991 Schools and Staffing Survey (SASS) Reinterview Response Variance Report,* (working paper 94-03), U.S. Department of Education, Office of Educational Research and Improvement, National Center for Education Statistics.

7. Conclusion

This study is part of the adjudicated report on "NCES MEASUREMENT ERROR PROGRAMS" which is scheduled for publication in the spring of 1997. This report synthesizes results from a sample of NCES reinterview programs, validity evaluations, and cognitive research studies.

Results indicate that the measurement error programs have helped NCES to improve the quality of its data. Over different rounds of surveys, the reinterview sample sizes have increased and response variance in most of the surveys in different areas have been improved. In some surveys it was found that inconsistencies between responses were attributed to factors like recoding the questions, lack of knowledge about the questions, eligibility criteria and small sample sizes. More recently, NCES applied alternative methods separately from reinterview programs to measure response bias. These methods were effective but costly. It was suggested that these methods be used when there is an indication of reporting errors.

Although, overall studies indicated that questions had low to moderate GDRs and low to moderate IOIs in several different NCES surveys, improvements in questionnaire wording and construction might help to reduce measurement errors in future surveys.

References

Brick, J., Cahalan, M., Gray, L., Severynse, J. and Stowe, P. 1994. *A Study of Selected Nonsampling Errors in the 1991 Survey of Recent College Graduates.* (NCES 95-640). Washington, DC: U.S. Department of Education. Office of Educational Research and Improvement. National Center for Education Statistics.

Brick, J., Collins, M., Celebuski, C. Nolin, M., Squadere, T., Ha, P., Wernimont, J., West, J., Chandler, K., Hausken, E. and Owings, J. 1991. *NHES:91 1991 National Household Education Survey Methodology Report.* (NCES 92-019). Washington, DC: U.S. Department of Education. Office of Educational Research and Improvement. National Center for Education Statistics.

Brick, J., Collins, M., Nolin, M., Ha, P., Levinsohn, M. and Chandler, K. 1994a. *National Household Education Survey of 1993: School Safety and Discipline Data File User's Manual.* (NCES 94-218). Washington, D.C.: U.S. Department of Education. Office of Educational Research and Improvement. National Center for Education Statistics.

Brick, J., Collins, M., Nolin, M., Ha, P., Levinsohn, M. and Chandler, K. 1994b. *National Household Education Survey of 1993: School Readiness Data File User's Manual.* (NCES 94-193). Washington, D.C.: U.S. Department of Education. Office of Educational Research and Improvement. National Center for Education Statistics.

Burkheimer, G., Forsyth, B., Wheeless, S., Mowbray, K., Boehnlein, L., Knight, S., Veith, K. and Knepper, P. 1992. *Beginning Postsecondary Students Longitudinal Study Field Test Methodology Report. BPS:90/92.* (NCES 92-160). Washington, DC: U.S. Department of Education. Office of Educational Research and Improvement. National Center for Education Statistics.

Bushery, J., Royce, D. and Kasprzyk, D. 1992. *The Schools and Staffing Survey: How Reinterview Measures Data Quality.* Proceedings of the Section on Survey Research Methods, American Statistical Association. Alexandria, VA: ASA, 458-463.

Forsman, G. and Schreiner, I. 1991. The design and analysis of reinterview: An overview. In P. Biemer, R. Groves, L. Lyberg, N. Mathiewetz, and S. Sudman (eds.), *Measurement Error in Surveys.* 279-302. New York: John Wiley & Sons, Inc.

Hanuschak, G., Atkinson D., Iwig, W. and Tolemeo, V. 1991. *History of Reinterview Studies at NASS.* Proceedings of the section on Survey Research Methods. American Statistical Association.

Kasprzyk, D. 1992. *The Schools and Staffing Survey: Research Issues.* Proceedings of the section on Survey Research Methods. American Statistical Association.

Royce, D. 1994. *1991 Schools and Staffing Survey (SASS) Reinterview Response Variance Report.* (Working Paper 94-03). Washington, D.C.: U.S. Department of Education. Office of Educational Research and Improvement. National Center for Education Statistics.

ESTIMATING RESPONSE BIAS IN AN ADULT EDUCATION SURVEY

J. Michael Brick and David Morganstein, Westat, Inc.
David Morganstein, Westat, Inc., 1650 Research Boulevard, Rockville, Maryland 20850

KEY WORDS: Reinterview, telephone survey, measurement errors

1. Introduction

Estimates from surveys are subject to both variable and systematic nonsampling errors. Variable nonsampling errors, or response variance, are those that might vary across repeated surveys administered to the same sample, assuming that the conditions of the interview could be controlled so that the surveys were independent. For example, the same respondent might report annual income differently when asked in repetitions of the same survey because the method used by the respondent to estimate income might vary (records might be used, recall might be used, or the value might be estimated using different schemes). These circumstances would lead to variable errors for estimates of income.

Systematic nonsampling errors, on the other hand, are those that have a particular direction. For example, if respondents tend to omit certain types of income, say interest income from savings, then the estimated income would be expected to be lower than the true income. In repetitions of the same survey, the estimated income would always be less than the true income. These types of systematic errors are called response bias. Survey estimates can be subject to both response variance and response bias.

Measuring response bias is typically very difficult. This study examines an intensive reinterview as a particular approach to estimating response bias. Other approaches for measuring response bias, the reasons for using an intensive reinterview, and the goals of the study are presented in the next section, after describing the source of the data. Section 3 outlines the methods used to collect the intensive reinterview data. Section 4 gives the estimates of the response bias and possible explanations of the findings. The last section summarizes the highlights of the study and the applicability of the method to other surveys. A more complete analysis of this study is given in Brick *et al.* (1996a).

2. Study Design

The source of the data for this analysis is a special methodological study undertaken as part of the 1995 National Household Education Survey (NHES:95). The NHES is an ongoing data collection system of the National Center for Education Statistics (NCES) conducted by Westat, Inc. designed to address a wide range of education-related issues. It is a telephone survey of the noninstitutionalized civilian population of the US that has been conducted in 1991, 1993, 1995 and 1996.

In the NHES, households are selected for the survey using random digit dialing (RDD) methods and data are collected using computer-assisted telephone interviewing (CATI) procedures. Approximately 60,000 households are screened for each administration. The NHES survey for a given year typically consists of a Screener, which collects household composition and demographic data, and extended interviews on two substantive components addressing education-related topics. This study is based on the Adult Education (AE) component of the NHES:95. It was designed to estimate the percentage of adults participating in adult education activities and the characteristics of both participants and nonparticipants.

As noted above, it is often difficult to measure response bias. A frequently used method of doing this is to compare the results of the survey against answers from a more definitive source, such as an administrative record file. However, record checks have their own limitations, *e.g.*, , record checks can only be used if records exist on the survey topic and those records can be accessed. Brick *et al.* (1994) found that, even for the well-defined topic of teacher certification, records were not complete and accurate and could not be matched to the survey respondents without error.

Another way of measuring response bias is through the use of reinterviews. Reinterviews are ordinarily undertaken to measure response variance rather than response bias. However, sometimes a process called reconciliation is used in reinterviewing to measure bias. If the original and reinterview responses are different, then the respondent is asked to reconcile the differences and the resulting response is called the reconciled response. The reconciliation is often conducted by a supervisor rather than a regular interviewer, assuming this will make the reconciled response less subject to error. Under these assumptions, the difference between the original and reconciled response has been used to estimate response bias (Forsman and Schreiner 1991).

Reconciliation has been used in earlier NHES studies to estimate response bias (Brick *et al.* 1996b) and a reinterview was also conducted for the NHES:95

AE interviews (Brick *et al.* 1996c). However, there is little evidence that reinterviews, even the reconciled responses, actually measure response bias. As a result of this, the NHES:95 reinterview study was designed to estimate the response variance.

It should not be too surprising that this approach does not provide reliable estimates of response bias. The methods used in the reinterviews, such as selecting interviewers from the original interviewer pool, asking the questions in much the same way as asked in the original interview, not informing the interviewer or the respondent of the answers from the original interview, and waiting at least 14 days between interviews so that the respondent will not remember the details of the original interview are all designed to support the measurement of response variance rather than response bias.

The intensive reinterview was designed to be an alternative method of estimating response bias that did not suffer from some of the shortcomings of the record check or regular reinterviews. The intensive reinterview method was pioneered by Belson (1986) who focused on difficult or sensitive topics primarily in opinion and marketing research.

The intensive reinterview differed from the regular reinterview in a number of ways. The interviewers were not selected from the regular pool of telephone interviewers, but were persons with previous experience in interviewing using less structured methods. The interviewers were trained to use a protocol and to conduct the reinterviews in a conversational mode, using probes and other devices to trigger recall and comprehension. The reinterview was focused on a few topics and ample time was allowed for discussing these few points. The respondents were encouraged to voice their opinions and understanding of the topics. Furthermore, attempts were made to engage the respondents in the interview by explicitly asking for their advice on ways to improve the interview. The hope was that these methods would lead to more complete and accurate reporting in the intensive reinterview.

Although there were four major research objectives of the study, only two of them are discussed in this paper. The first goal of the study was to examine the potential bias in the estimates of the percentage of adults who participated in adult education activities. The bias could be due to either underreporting participation or overreporting activities that took place outside of the time frame of the survey (i.e., the past 12 months prior to the original interview). Respondents might underreport participation either because they might not recall a qualifying activity during the 12 months before they were interviewed or because they might not comprehend the range of activities that were included as adult education. These types of underreporting would lead to downward bias in participation rates. However, an upward bias could occur if respondents "telescoped" some activities. Telescoping is reporting activities that took place outside of the time frame of the survey as having taken place within that time frame. As described below, underreporting was expected to be minimal in all types of adult education (the six types of activities were: ESL, adult basic education/GED preparation classes, credential programs, apprenticeships, work-related courses, and personal development courses), except work-related and personal development courses. As a result, the intensive reinterview focused on these two types of participation in order to assess the bias in the overall participation rate.

The second goal was to obtain more accurate estimates of participation in work-related and personal development courses, separately. As a result of the differences between the estimates of participation from earlier AE surveys and cognitive laboratory work, it was suspected that work-related courses and personal development courses were susceptible to underreporting. One of the major concerns for reporting these types of courses is that respondents might not comprehend the full range of activities that are included as work-related and personal development courses. These types of comprehension problems could combine with recall problems and result in underreporting of work-related and personal development courses.

3. Intensive Reinterview Method

In an attempt to more closely determine the respondent's actual status or opinions, the intensive interview was more of a directed conversation between the respondent and the interviewer rather than a formally scripted interview. Respondents were reminded of their answers in the original survey and asked if the answers were still true for them. They were asked to recall other details related to their responses. Interviewers were fully knowledgeable about the original answers given by the respondent. Tactics similar to those used in cognitive laboratory work, such as asking open-ended questions and using probes to encourage the respondent to elaborate on his or her answer, were used. The goal was to obtain more detailed and accurate information by understanding the respondent's perspective and the reasons for his or her answers.

The intensive reinterview was a new undertaking and presented several challenges. For example, the interviewers who conducted the NHES interviews were thoroughly trained to read the questions verbatim and to avoid affective behavior that might influence the

respondent. Adopting the conversational and unstructured interviewing method called for in the intensive reinterview required major changes in their behavior. The interviewers were also called upon to implement some methods used in cognitive research, but they were not previously trained in these methods. The respondents also faced a challenge because the intensive reinterview differed significantly from the type of interview they had already done. They were called upon to give reasons for their responses and provide details rather than choose among response alternatives.

To address these challenges, a protocol and data collection methods were developed especially for this study. The full details of the protocol development, the methods used to select and train the interviewers, the sampling of respondents to the original interview, and data collection methods are provided by Brick *et al.* (1996a). These issues are very important but space limitations prevent giving but a few important features of the final sample in this paper.

Although the goal of the study was to develop estimates of bias, only a very limited sample size could be fielded. Because of the small sample sizes, it was decided that the typical design-based estimates gathered from the original interview would be subject to very large sampling errors and relationships would be obscured by these sampling errors. Thus, the results from this relatively small sample were analyzed assuming the observations were from independent, identically distributed random variables and sampling weights were not used. The sample was randomly selected from both participants and nonparticipants who completed the AE extended interview. In order for a case to be eligible for the study, certain conditions had to be met. For example, the original interview had to be conducted in English.

A sample of 230 adults was selected to meet specific targets by participation status and educational level. Of the 206 sampled adults who completed the intensive reinterviews (90 percent), 115 were nonparticipants in the original interview and 91 were participants.

4. Findings

4.1 Bias in Participation Rate Estimates

The first goal of the study was to estimate the response bias associated with estimates of the rate of participation in AE from the NHES:95. Adults were classified as participants in AE if they had participated in one or more of six different types of adult education activities during the past 12 months. Based on the responses in the NHES:95, 40 percent of all adults had participated in one or more of these activities in the last 12 months (Kim *et al.* 1995).

As noted earlier, responding to the items about participation in work-related and personal development courses was identified as being problematic during the design phase of the NHES:95. After the survey was completed, the results from the reinterview confirmed that these two types of participation were much more likely to be reported inconsistently than any of the other types (Brick *et al.* 1996c). These results support the decision to restrict this study to an in-depth examination of reporting work-related and personal development courses.

Adults were classified as nonparticipants in the original interview if they said they had not taken any courses in the last 12 months. Of the 115 nonparticipants who responded to the intensive reinterview, 41 percent indicated that they had taken one or more work-related or personal development courses (in the intensive reinterview, respondents were not asked about other types of courses). Since none of the participants sampled for the intensive reinterview denied having taken courses, the response bias in the overall participation rate is one-directional and substantial.

Assuming the responding nonparticipants in the Bias Study are a simple random sample of all adults classified as nonparticipants in the NHES:95 (the analysis is thus unweighted), the bias in the NHES:95 estimate is 24 percent. The bias is estimated by multiplying the percent of all adults who were nonparticipants as reported in the NHES:95 by the percent of the nonparticipants who reported participating in the intensive reinterview. In general, the estimated bias is

$$\hat{b}(p_0) = p_0\hat{y}^{p} - (100 - p_0)\hat{y}^{np} \qquad (1)$$

where p_0 is the estimate of the percentage of adults classified as participants in the initial interview, $\hat{y}^{p}$ is the estimate of the proportion of participants in the initial interview who reported not participating in the intensive reinterview, and $\hat{y}^{np}$ is the estimate of the proportion of nonparticipants in the initial interview who reported participating in the intensive reinterview. In this case, the last term of the estimated bias is zero ($\hat{y}^{p}$=0), because no initial participants said they had not taken any courses during the intensive reinterview. If the bias in the estimated percentage of adults who participated in AE is 24 percent, then the bias-corrected estimate is that 64 percent of adults participated in AE in 1995. This is substantially larger than the 40 percent reported in the NHES:95. Both the bias and the percentage participating from the NHES:95 are subject to sampling error and because of the sample size the sampling error of the bias is very large relative to that for the estimate from the NHES:95. Taking advantage

of the fact that $\hat{y}^{p}$=0, the estimated bias can be written as

$$\hat{b}(p_0) = (p_0 - 100)\hat{y}^{np}.$$

Thus, the estimated variance of the bias is a product of random variables and the approximate variance for a product of independent random variables is (Hansen *et al.* 1953)

$$Var(\hat{b}) = \left(\hat{y}^{np}\right)^2 Var(p_0) + (p_0 - 100)^2 Var(\hat{y}^{np}) \quad (2)$$

Substituting the estimated values into (2) and taking the square root, the standard error of the estimated bias of 24 percent is 2.7 percent. Thus, a 95 percent confidence interval for the estimated bias is from 19 percent to 29 percent, and for the percent of adults participating, the confidence interval is from 59 to 69 percent.

Considering the nonparticipating respondents to the intensive reinterview a simple random sample of all nonparticipants in the original survey is a key assumption in estimating the response bias. Usually, the sampling procedures would ensure that this assumption holds, but the sampling methods described earlier were primarily concerned with making sure the sample sizes for specific groups were large enough to provide some nonparticipants with various characteristics. In addition, the small sample size does not allow for broad generalizations. Thus, these estimates are exploratory and should not be used to make bias corrections to the NHES:95 estimates.

To evaluate the reasonability of this assumption, the characteristics of the responding nonparticipants from the intensive reinterview were compared to the characteristics of all nonparticipants from the NHES:95. While the age and sex distributions are similar, the educational attainment distributions are different, with a much larger percentage of intensive reinterview nonparticipants having less than a high school education. This is a consequence of the sampling methods used for the intensive reinterview. This difference highlights the fact that the point estimates and confidence intervals from the study are subject to specification errors that cannot be measured. Despite this shortcoming, the findings clearly show that a relatively large fraction of the adults classified as nonparticipants in the original survey did identify AE activities in the intensive reinterview.

In addition to nonparticipants, participants in the original survey who were sampled were asked if there were any courses they had not reported in the initial interview by using the same types of probes described above. All of the work-related and personal development courses participants reported in the NHES:95 were verified as being within the eligible 12 month time period. In addition, about one-third of the sampled participants identified additional courses that were not reported in the original interview.

The reporting of additional work-related and personal development courses by adults classified as participants in the original survey is a further indication that the respondents may have had a more restrictive understanding of the scope of activities than was intended. Drawing on the work of Schwarz (1995), one interpretation of this finding is that respondents might have reacted to the context of the original interview in determining what was an eligible activity. The NHES:95 interview began by asking about more formal types of participation and some respondents may have created a response paradigm before the questions about the less formal activities were asked. In the intensive reinterview the context was different because the only types of courses discussed were work-related and personal development courses.

4.2 Bias in Work-Related and Personal Development Participation Estimates

Overall, about half the adults who named additional courses reported work-related courses and half reported personal development courses. Participants were more likely to add personal development courses and the nonparticipants were more likely to add work-related courses, but these differences are not statistically significant.

Based on the NHES:95 responses, 21 percent of adults were estimated to have participated in work-related courses during the previous 12 months and 20 percent were estimated to have participated in personal development courses (Kim *et al.* 1995). The extent of the bias in these estimates can be estimated using equation (1). The bias for the work-related participation rate is

$$\hat{b}(p_{0,wr}) = (1 - p_{0,wr})\hat{y}^{np}_{wr} \quad (3)$$

where $p_{0,wr}$ is the estimate of the percentage of adults classified as work-related participants in the initial interview, and $\hat{y}^{np}_{wr}$ is the estimate of the proportion of adults who did not report participating in work-related activities in the initial interview but reported participating in the intensive reinterview. Because we are now dealing with participation in a particular type of adult education, the value of $\hat{y}^{np}_{wr}$ has two components: those classified as nonparticipants who reported taking work-related courses in the intensive reinterview, and participants in the initial survey who reported taking work-related courses for the first time in the intensive

reinterview. In the intensive reinterview, 23 percent of the nonparticipants reported taking work-related courses and 8 percent of the participants reported taking work-related courses for the first time.

Substituting the values into (3), the estimated bias for the percent of adults participating in work-related courses is 16 percent. The standard error can be computed using (2), where $\hat{y}_{wr}^{np}$ is treated as a sum of the two components described above. Using this approach, the standard error of the estimated bias is 3 percent and the 95 percent confidence interval for the estimated bias is from 10 to 22 percent.

The same calculations can be performed for personal development courses to compute the estimated bias and its standard error. The estimated biases are summarized in Table 1 below.

These estimates show that the underreporting bias is approximately the same for both work-related and personal development courses. While these estimates are subject to the same caveats as the overall estimates of participation rates, they also have an interesting implication because of the difference in the wording of the questions about the two types of participation. The introduction to the question about work-related courses does not include specific examples, but does mention courses taken at work, taken somewhere else but related to work or career, and courses taken to obtain a license or certificate related to work or career. On the other hand, the introduction to the question about personal development specifically mentions courses including "arts and crafts, sports or recreation, first aid or childbirth, Bible study, or any other course."

One way of interpreting the equal biases for the two types of participation is that adding examples does not improve the quality of reporting in this situation. This interpretation is consistent with the hypothesis that respondents develop a response paradigm in the original interview that includes only more formal courses. If this is true, then the addition of the specific examples may not cause respondents to change their paradigm and a different approach might be needed to address the exclusion of less formal courses in reporting.

5. Summary

The estimated bias in the overall participation rate of adults was 24 percent and the bias-corrected estimate is that 64 percent of adults participated in AE in 1995. This is substantially larger than the 40 percent reported in the NHES:95. The underreporting of participation for work-related and personal development courses was also substantial and of about the same magnitude for each of these types of participation.

A reason for the underreporting interpretation may be related to how respondents react to the context of the interview. Some respondents may have created a response paradigm that restricted their answers to more formal courses before the questions about the work-related and personal development courses were asked. Despite the fact that more examples were used for the personal development courses than for the work-related course question, the estimated biases were approximately the same for the two types of participation. This suggests that simply adding examples to the wordings of the questions may not improve the quality of reporting and that other approaches to the underreporting problem may be needed.

If the adults have developed a response paradigm that focuses on formal types of participation (*i.e.*, traditional schooling or formal programs), then a relatively drastic intervention may be needed to modify this behavior. For example, a modification in which the respondents are asked to actively cooperate in changing the focus, for example by giving examples of less formal courses, might be more effective.

The intensive reinterview methodology appears to have good potential as a method for detecting biases, especially if more traditional methods like record check studies are not feasible. The alternative approach of using reconciled reinterviews, on the other hand, has not proven to be successful for estimating bias. However, from an operational perspective, it is important to understand that this method is more costly than a regular reinterview. As a result, this method should be used primarily when there is an indication of reporting errors and the estimates subject to the biases are important to the survey objectives.

References

Belson, W.A. (1986). *Validity in survey research.* Cambridge, England: University Press.

Brick, J.M., Cahalan, M., Gray, L., Severynse, J., and Stowe, P. (1994). "A study of selected sampling errors in the 1991 Survey of Recent College Graduates," Technical Report. U.S. Department of Education, Office of Educational Research and Improvement, NCES 95-640.

Brick, J.M., Kim, K., Nolin, M.J., and Collins, M. (1996a). "Estimation of response bias in the NHES:95 Adult Education Survey." Working Paper ,96-13. U.S. Department of Education, Office of Educational Research and Improvement, National Center for Education Statistics.

Brick, J.M., Rizzo, L., and Wernimont, J. (1996b). "The 1993 National Household Education Survey: Reinterview results for the School Readiness and School Safety and Discipline components." Technical report. U.S. Department of Education, Office of Educational Research and Improvement, NCES. (in progress).

Brick, J.M., Wernimont, J., and Montes, M. (1996c). "The 1995 National Household Education Survey: Reinterview results for the Adult Education component." Working Paper, 96-14. U.S. Department of Education, Office of Educational Research and Improvement, National Center for Education Statistics.

Forsman, G., and Schreiner, I. (1991). "The design and analysis of reinterview: An overview." In P. Biemer, R. Groves, L. Lyberg, N. Mathiowetz, and S. Sudman (eds.), *Measurement error in surveys*, 279-302. New York: John Wiley & Sons.

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

Kim, K., Collins, M., Stowe, P., and Chandler, K. (1995). "Forty percent of adults participate in adult education activities: 1994-95." Statistics in Brief. U.S. Department of Education, Office of Educational Research and Improvement, NCES 95-823.

Schwarz, N. (1995) "What respondents learn from questionnaires: the survey interview and the logic of conversation." *International Statistical Review*, 62, 2, 153-177.

Table 1. Estimates of bias in overall, work-related, and personal development participation rates.

Type of participation	NHES:95 estimate	Estimated bias	Sampling error	Bias-corrected estimate
Overall	40%	24%	2.7%	64%
Work-related	21%	16%	3.0%	37%
Personal development	20%	14%	3.0%	34%

Source: U.S. Department of Education, National Center for Education Statistics, Bias Study of the National Household Education Survey, 1995.

OPTIMAL PERIODICITY OF A SURVEY: EXTENSIONS OF PROBABLE-ERROR MODELS

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

KEY WORDS: Absolute error modeling, Data deterioration, Fixed-and-variable costs, Repeated surveys, Sampling designs

This paper extends prior work on the problem of choosing optimal periodicity (and associated sample sizes) for repeated surveys of public and private schools with joint consideration of data deterioration (resulting from unobserved year-to-year changes in the underlying process variables), sampling error, and cost. The family of "probable-error models" that was first described in Ghosh *et al.* (1994) has been extended and empirical results obtained for state-level as well as national-level estimates using data from three rounds of the Schools and Staffing Survey (SASS). As noted in the 1994 paper, the models provide "a direct approximate method for characterizing the decision problem of making a joint choice of inter-survey intervals and sample sizes under a fixed cost constraint." The extensions reviewed in the present paper assume, for the most part, that conventional direct estimation methods will be used by the data user. In the case of a proposed alternative sampling design suggested by the modeling results, the data user may wish to consider the use of an indirect estimation (time series modeling) approach along the lines discussed in Smith *et al.* (1995).

SASS was conducted at three-year intervals for school years 1987-88, 1990-91, and 1993-94. Future rounds may be conducted at intersurvey intervals of 4, 5, or 6 years. The modeling extensions are illustrated here in a review of two of several new models that were formulated as modifications of the earlier models. The two models provide alternative formulations to account for the approximate average errors incurred by a data user within successive 12-month periods following a SASS data collection and up to the time of the next data collection. Projected absolute errors have been estimated for future national-level and typical state-level data collections for selected policy variables and a range of fixed-to-variable cost ratios for each possible periodicity.

The two illustrative models, denoted as Model 3A and Model 4M, are modifications of Model 3 and Model 4 of the 1994 paper. They combine a sampling absolute error *(s.a.e.)* and a process shift D over time in different ways to obtain, for different periodicities, estimates of the year-by-year projected absolute errors that would be incurred by a data user as well as average projected absolute errors for each multi-year periodicity. Annual dollar resources for SASS are assumed to be fixed. For each of several scenarios this assumption constrains the total annualized cost to a fixed amount and hence determines the sample size for each combination of a periodicity (4, 5, or 6 years) and a fixed-to-variable cost ratio.

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 could affect the quality of decisions made by users. On the other hand, if the survey is undertaken very frequently, the costs of conducting the survey and analyzing the data and the indirect costs of the response burden may be judged to have costs that exceed the benefits achieved in using fresh data. In the context of repeated surveys, it is useful to distinguish both opportunities and problems presented by different designs.

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. In fact, the "truly optimal" formulation may be precluded by external constraints.

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 depend on the periodicity. If so we represent the start-up cost as $C_{0,k}$ (where k is the periodicity), which may be regarded as increasing with increasing periodicity; that is, the start-up cost may be more if the periodicity is five years compared to the start-up cost for a periodicity of four years and so on. On the other hand, the start-up cost may be considered to be constant; that is, it may not depend on the periodicity of the survey. Further details are given in Ghosh *et al.* (1994).

We assume that the true value of a variable remains constant for a year after the survey date. This is an appropriate assumption for the SASS survey system since nearly all of the observed variables under the various SASS questionnaires have an annual

accounting period and the SASS data user is interested in changes in variables which are specified to change as of some conventional time point. For example, the official figure for enrollment and number of teachers in a public school is the enrollment "on or about October 1" of the school year. The corresponding number of teachers or the full-time equivalent (FTE) number of teachers are counted at about the same point in time. The student enrollment and the teacher count may fluctuate during the academic year, but SASS and the Common Core of Data (CCD) are, in effect, taking snapshots at the same time over a sequence of years. The error committed in using a survey estimate is exactly equal to the difference between the survey estimate and the true value. Within the first twelve-month interval from the survey date any user incurs an error which equals the difference between the true value and the survey estimate. The estimated standard error of the survey estimate provides an indication of this difference.

If one were interested in estimating from SASS data for a survey year the mean of some characteristic for a specified group of schools, such as the average "number of K-12 teachers that are new to the school this year" for all regular public elementary schools in the state of California, then the estimate would be constructed by applying the school weight for each school to the reported number of new teachers for that school, summing the products and dividing by the sum of the school weights. For some of the SASS-based public school statistics published by NCES, such as those in the *Statistical Profiles* for each round of SASS, the NCES publications include tables of state-by-state estimates of the statistics and, for a selected subset of of the state-by-state statistics, they also include tables of the estimated standard errors for those statistics. For example, the publication *Schools and Staffing in the United States: A Statistical Profile, 1990-91* includes for public schools both estimates of the statistics and estimated standard errors for these statistics on a state-by-state basis for (1) Number of public schools and students and average number of students per full-time-equivalent (FTE) teacher, (2) Percentage distribution of public school teachers by sex and race-ethnicity, percent minority teachers, and average teacher age, and (3) Average base salary for full-time public school teachers and average public school principal salary. As stated in the technical notes to that publication, "Standard errors were estimated using a balanced repeated replications procedure that incorporates the design features of this complex survey."

The difference between the true value and the survey estimate is the deviation from the mean m in the normal distribution of the survey estimate x considered as random variables. We denote the average of the absolute deviations as the "sampling absolute error" or *(s.a.e.)*. Assuming a normal distribution, the projected absolute error incurred by a user during the first year after the survey is $0.8\ s\ /\ \mathrm{sqrt}(n)$ where $s\ /\ \mathrm{sqrt}(n)$ is the standard error of the estimate, assuming simple random sampling. At the end of each year we assume that the true value undergoes a change. The magnitude of this change at the end of each year is denoted $|\ D\ |$. The sampling error component is $0.8\ s\ /\ \mathrm{sqrt}(n)$. Thus the expected value of the total error committed by a data user is dependent on *(s.a.e.)* and on $|\ D\ |$. The magnitude of the change at the end of the second year is also $|\ D\ |$, and so on.

In Model 3A, which is a variant of the Model 3 described in Ghosh *et al.* (1994), we assume the year-to-year process disturbance (process error) to be a normal variable with a zero mean. (If needed, a process error with a nonzero mean could be incorporated into the analysis framework.) Since the process error and the sampling error are both assumed to be normally distributed, they can be readily combined. The projected absolute error is then a linear combination of the process absolute error and the sampling absolute error.

In Model 4M, we explicitly assume that the process change which occurs each year (for example, every October) occurs in accordance with a Random Walk process in discrete time. That is,

$$x_t = x_{t-1} + w_t$$

where w_t has mean zero. We then calculate the average error for different possible periodicities of the repeated survey. The optimal intersurvey interval can be determined if the process variance and the sampling variance are known. In a Random Walk model, the current level of the process is the best current forecast for any future year. One assumes that any known trend component has already been subtracted out. In general, data users will typically use the last available survey value as long as no new survey has been conducted. This assumption concerning user behavior is consistent with our assumption of an underlying Random Walk process.

As noted above, Model 3A is a variant of Model 3 of Ghosh *et al.* (1994). The new Model 4M is a modification and replacement for Model 4 of that 1994 paper. In the original Model 4 we introduced the concept of a loss parameter that converted the sampling error together with the unobserved process shift in non-survey years to a loss expressed in

monetary units. The combination of average cost and average error over a period of years was minimized to determine the optimum periodicity. This was a variation on an approach in Smith (1980) and also on an analysis suggested by S. Kaufman. Model 4M, however, does away with the need for a separate loss parameter, thus avoiding the introduction of a subjective judgment on the part of the survey administrator.

The following table sets forth the year-by-year evolution of the projected absolute errors for the two models. In Model 3A the evolution is based on $|D|$, the magnitude of the annual change in the true value, and the sampling absolute error, *(s.a.e.)*. For Model 4M, the evolution is based on D^2, which is proportional to the variance of the process disturbance, and the sampling absolute error, *(s.a.e.)*. The *(s.a.e.)* depends on the sample size which, in turn, depends on the chosen periodicity under the constraint of fixed annualized cost.

Projected Absolute Errors for Selected Models

Year	Model 3A	Model 4M
1	$(s.a.e.)$	$(s.a.e.)$
2	$0.8\,\lvert D\rvert+(s.a.e.)$	$0.8\sqrt{D^2+\left[\frac{(s.a.e.)}{0.8}\right]^2}$
3	$0.8\sqrt{2}\,\lvert D\rvert+(s.a.e.)$	$0.8\sqrt{2D^2+\left[\frac{(s.a.e.)}{0.8}\right]^2}$
4	$0.8\sqrt{3}\,\lvert D\rvert+(s.a.e.)$	$0.8\sqrt{3D^2+\left[\frac{(s.a.e.)}{0.8}\right]^2}$
5	$0.8\sqrt{4}\,\lvert D\rvert+(s.a.e.)$	$0.8\sqrt{4D^2+\left[\frac{(s.a.e.)}{0.8}\right]^2}$
6	$0.8\sqrt{5}\,\lvert D\rvert+(s.a.e.)$	$0.8\sqrt{5D^2+\left[\frac{(s.a.e.)}{0.8}\right]^2}$
Avg p.a.e. (6 yrs)	$\frac{0.8}{6}\sum_{i=1}^{6}\sqrt{i-1}\,\lvert D\rvert+(s.a.e.)$	$\frac{0.8}{6}\sum_{i=1}^{6}\sqrt{(i-1)D^2+\left[\frac{(s.a.e.)}{0.8}\right]^2}$

We applied the models described above using three rounds of SASS data at the national level (U.S.) and at the state level for three selected states (California, Iowa, and New York). The following twelve variables were selected from the School, Administrator, and Teacher questionnaires:

Item 1. Number of students served by chapter 1 services (Schools--public).
Item 4. Number of K-12 teachers that are new to the school this year (Schools--public).
Item 6. Percentage of all schools with minority principals (Adminr--public and private).
Item 7A. Number of students per FTE teacher, by sector (Schools--public).
Item 7B. Number of students per FTE teacher, by sector (Schools--private).
Item 8A. Percentage of schools in which various programs and services were available (Schools--public).
Item 8B. Percentage of schools in which various programs and services were available (Schools--private).
Item 9. Percentage of principals having master's degree (Administrator--public).
Item 10A. Percentage of full time teachers who received various types of compensation (Teacher--public).
Item 10B. Percentage of full time teachers who received various types of compensation (Teacher--private).
Item 11A. Percentage of full time teachers newly hired and were first time teachers (Teacher--public).
Item 11B. Percentage of full time teachers newly hired and who were first time teachers (Teacher--private).

Private school items 7B, 8B, 10B, and 11B were omitted from the state-level computer runs since state-level estimates are not published by NCES for private schools. Item 6, which is based on pooled data for public and private schools combined, was retained in all runs.

We obtained approximate estimates for the fixed cost and variable cost elements of SASS. We applied the two models for each variable listed above, and computed the projected absolute error for periodicities of four, five, and six years and for specified scenarios of fixed-to-variable cost. The accompanying graphs for Model 3A and Model 4M, respectively, show the average *rel p.a.e.* (where the *rel p.a.e.* for each variable is its *p.a.e.* divided by its mean) for Iowa, New York, California, and the U.S. for a set of eight policy variables with a fixed total cost and a fixed-to-variable cost ratio of 50:50. We see that for the U.S. as a whole, shorter periodicities (even with their smaller sample sizes) result in smaller relative projected absolute errors. For California and Iowa, under both Model 3A and Model 4M, the averages of the relative projected absolute errors are larger for short periodicities and smaller for longer periodicities. For New York, the mean values of the *rel p.a.e.* are essentially flat under Model 3A over the periodicity range from 2 to 6 years but under Model 4M the values decline initially, with a minimum at a periodicity of 4 years, and then rise slightly for periodicities of 5 and 6 years.

Under the probable-error models the data users who are primarily interested in carrying out analyses for individual States will generally incur smaller errors if they are provided with datasets from longer periodicities and hence larger sample sizes. Data users who are primarily interested in carrying out analyses for the U.S. as a whole will incur smaller errors if they are provided with datasets from shorter periodicities and correspondingly smaller sample sizes.

These observations have led to an alternating large-and-small-sample scenario which was formulated as follows: Assume the same fixed annualized resource budget that would otherwise support the large-sample scenarios with a periodicity of five years over a range of cost ratios (0.3, 0.4, 0.5, 0.6, or 0.7). Then the assumed sample sizes for the U.S. at the mid-point cost ratio p=0.5 were 9,000 public schools and 48,000 teachers. These sizes were proportionally smaller or larger for smaller or larger cost ratios. Assign these sample sizes to a periodicity of six years instead of to a periodicity of five years. This results in a "cost dividend" of 20 per cent which can be invested in a one-fifth U.S. sample of 1,800 public schools and 9,600 teachers for a data collection which can be conducted at the halfway point between two full-sample data collections; namely, three years after the previous large data collection. Assume that there is no processing delay. For simplicity, assume that the schools in the one-fifth sample are nonoverlapping with the schools in the full sample. Further assume that for the U.S. as whole only direct estimates will be of interest and, hence, the two independent samples (the full sample and the one-fifth sample) will be treated as independent cross-section surveys three years apart.

Now consider the projected absolute errors that will be incurred by a data user over a six-year period. During the first, second, and third years after a full-sample data collection a data user who is interested in national data will continue to use that sample. Under

either Model 3A or Model 4M the projected absolute errors will increase each year. In the fourth year after the large sample data collection a new dataset from the one-fifth size national sample will become available. The data user then disregards the data in the old large sample and begins to use the data from the new one-fifth sample and continues to use it until data from the next full sample becomes available in the seventh year. For the U.S., the sample sizes in the one-fifth national sample are large enough that quite good estimates may be made. That is the user is not heavily penalized in shifting every three years between the full national sample and the one-fifth national sample.

For the cost ratio p=0.5 the average rel p.a.e. values for an alternating large-and-small sample design with large sample periodicity of six years are less than or equal to the average rel p.a.e. values for the single large-sample scenario with periodicity of five years. Furthermore, the user of U.S.-level data will be receiving the benefits associated with the receipt of fresh data every three years instead of every five years. Related numerical results will be found in Smith, Ghosh, and Chang (1996).

Our main conclusion from the present study is that the National Center for Education Statistics should consider adopting an alternating large-and-small-sample design for SASS with an appropriate full-sample periodicity together with a mid-period fractional-sample to provide a periodic update at the national level and for larger States.

References

Abramson, R. *et al.* (1996), "1993-94 Schools and Staffing Survey: Sample Design and Estimation," Technical Report NCES 96-089, Washington, DC: U.S. Department of Education, Office of Educational Research and Improvement, National Center for Education Statistics (forthcoming).

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

Kaufman, S. (1991), "1988 Schools and Staffing Survey Sample Design and Estimation," Technical Report NCES 91-127, Washington, DC: U.S. Department of Education, Office of Educational Research and Improvement, National Center for Education Statistics.

Kaufman, S. and Huang, H. (1993), "1990-91 Schools and Staffing Survey: Sample Design and Estimation," Technical Report NCES 93-449, Washington, DC: U.S. Department of Education, Office of Educational Research and Improvement, National Center for Education Statistics.

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.

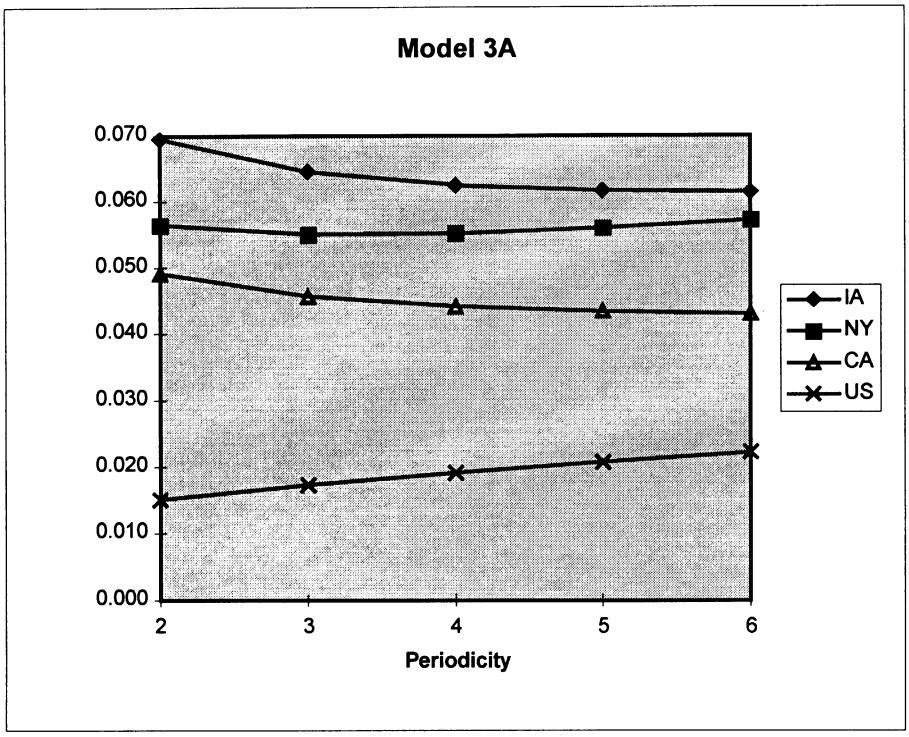
Smith, W. (1980), "Sample Size and Timing Decisions for Repeated Socioeconomic Surveys," unpublished D.Sc. dissertation, The George Washington University, School of Engineering and Applied Science.

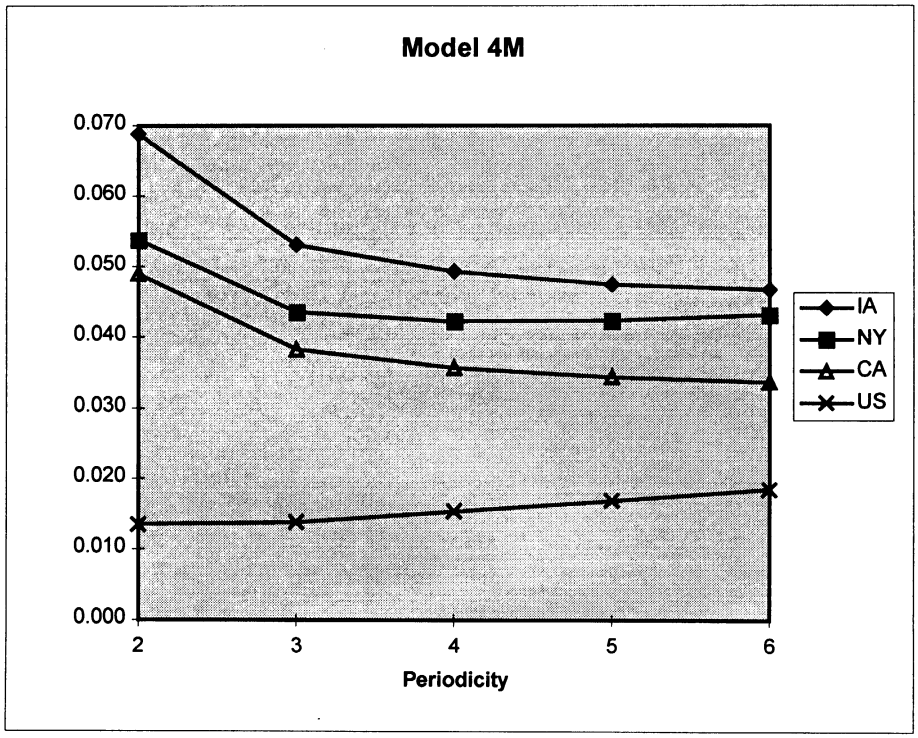
Smith, W. and Barzily, Z. (1982), "Kalman Filter Techniques for Control of Repeated Economic Surveys," *Journal of Economic Dynamics and Control,* 4, 261-279.

Smith, W., Ghosh, D. and Chang, M. (1995), "Optimal Periodicity of a Survey: Alternatives Under Cost and Policy Constraints," *1995 Proceedings of the ASA Section on Survey Research Methods.*

Smith, W., Ghosh, D. and Chang, M. (1996), "Optimizing the Periodicity of the Schools and Staffing Survey: An Updated Assessment Based on Three Rounds of SASS Data," Technical Report, Synectics for Management Decisions, Inc., Arlington, VA (forthcoming).

Average *rel p.a.e.* for Iowa, New York, California, and U.S. for Eight Policy Variables with Fixed Total Cost and p=0.5





Projected Absolute Errors for Selected Models

Year	Model 3A	Model 4M
1	$(s.a.e.)$	$(s.a.e.)$
2	$0.8\,\lvert D\rvert+(s.a.e.)$	$0.8\sqrt{D^2+\left[\frac{(s.a.e.)}{0.8}\right]^2}$
3	$0.8\sqrt{2}\,\lvert D\rvert+(s.a.e.)$	$0.8\sqrt{2D^2+\left[\frac{(s.a.e.)}{0.8}\right]^2}$
4	$0.8\sqrt{3}\,\lvert D\rvert+(s.a.e.)$	$0.8\sqrt{3D^2+\left[\frac{(s.a.e.)}{0.8}\right]^2}$
5	$0.8\sqrt{4}\,\lvert D\rvert+(s.a.e.)$	$0.8\sqrt{4D^2+\left[\frac{(s.a.e.)}{0.8}\right]^2}$
6	$0.8\sqrt{5}\,\lvert D\rvert+(s.a.e.)$	$0.8\sqrt{5D^2+\left[\frac{(s.a.e.)}{0.8}\right]^2}$
Avg p.a.e. (6 yrs)	$\frac{0.8}{6}\sum_{i=1}^{6}\sqrt{i-1}\,\lvert D\rvert+(s.a.e.)$	$\frac{0.8}{6}\sum_{i=1}^{6}\sqrt{(i-1)D^2+\left[\frac{(s.a.e.)}{0.8}\right]^2}$

This page intentionally left blank.

ESTIMATING THE VARIANCE IN THE PRESENCE OF IMPUTATION USING A RESIDUAL

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

Key Words: Imputation, Variance, Simulation, Half-Sample Replication

1. Introduction

With most surveys there is always some item nonresponse. One way of dealing with item nonresponse is to include imputations for the missing data. This makes it easy to compute appropriate population estimates. However, the variance will be underestimated in the presence of imputation. This follows from the main assumption with design based estimation; the only random variable is the variable specifying which units are in sample. When an imputation is generated from a sample, the imputation is not known given the unit requiring imputation. However, since it is assumed that the imputed value is known (i.e., it has no variability), standard variance estimation can not properly reflect the imputation variance.

A number of methodologies have been proposed to estimate imputation variance. Multiple imputation (Rubin,1978), a modified jacknife (Rao, 1992) and a model assisted methodology (Särndal, 1990) are some of the approaches that have been proposed. Rao's and Särndal's methodologies require special software to estimate the variance. Rubin's methodology requires the computation of the variance estimate associated with the average of the multiple imputation estimates. With respect to additional software requirements, the multiple imputation variance estimate, although not as complicated as the other methods, is still more complicated then standard variance estimation.

The methodology proposed in this paper is a mixture of the methodologies described above. Like multiple imputation, two potential imputations are assigned to each nonrespondent. The difference between these values is a residual that is added to appropriate data elements to reflect the imputation variance. The residual term differs from the one used in the modified jacknife, in that the modified jacknife residual introduces the variability replicate by replicate, while here the variability is introduced to the actual data elements. Once the residuals have been added, standard replication programs can compute the total variance. The questions this paper addresses are: 1) when will this process appropriately measure the total variance; and 2) when the process is not appropriate, what must be added to provide appropriate estimates.

A simulation study, modeled after NCES's Schools and Staffing Survey (SASS), which has a complex sample design, will demonstrate the proposed procedures. A nearest neighbor imputation will be used. The nonresponse will be modeled assuming unequal nonresponse rates per cell and that larger units are more likely to be nonrespondents.

2. Imputations

The nearest neighbor imputations used in this paper are done within imputation cells, after the schools have been sorted by the number of student per school. The imputation cells are state/school level/urbanicity. There are three school levels - elementary, secondary and combined schools. There are three levels of urbanicity - central city, urban fringe/large town and rural/small town. After the file is sorted, it is accessed sequentially using the nearest responding school as the donor for a nonresponding school. Two imputations will be determined for each nonrespondent. One where the file is sorted in ascending order and another where the file is sorted in descending order. A random selection is used to determine which imputation is used in the estimate of interest ($\hat{y}_{\bullet}$).

3. Definitions

r is the set of responding units

nr is the set of nonresponding units

3.1 Terms defining the imputation and residual

$$\tilde{y}_k = \begin{cases} y_k & \text{if } k \in \text{r} \\ y_{k1} I_k + y_{k2}(1 - I_k) & \text{if } k \in \text{nr} \end{cases}$$

I_k represents an independent selection within each unit k with probability of .5 for a value of 1, and 0 otherwise

y_k is the response from unit k

y_{k1} is the nearest neighbor imputation in ascending order

y_{k2} is the nearest neighbor imputation in descending order

$\hat{y}_{\bullet} = \sum_{k \in s} w_k \tilde{y}_k$, the main estimate of interest

w_k is the sampling weight for unit k

s is a probability sample of units

$\hat{y} = \sum_{k \in s} w_k y_k$, the estimate with complete response

$$\tilde{d}_k^R = \begin{cases} 0 \text{ if } k \in \text{r} \\ \hat{y}_{j_k 2} - \hat{y}_{j_k 1} \text{ if } k \in \text{nr and } y_{k1} \text{ is used in } \hat{y}_\bullet \\ \hat{y}_{j_k 1} - \hat{y}_{j_k 2} \text{ if } k \in \text{nr and } y_{k2} \text{ is used in } \hat{y}_\bullet \end{cases}$$

where:

j_k is a unit independently and randomly selected from $\{j \mid j \in \text{nr and within } k\text{'s imputation cell}\}$. The selection is done proportional to the w_j's.

In addition, R will be used to represent the selection of unit j_k.

$\hat{d}^R = \sum_{k \in s} w_k \tilde{d}_k^R$

3.2 Terms used in section 4

$$\tilde{d}_{ki}^T = \begin{cases} 0 \text{ if } k \in \text{r} \\ y_{m_k 2} - y_{m_k 1} \text{ if } k \in \text{nr and } i = 1 \\ y_{m_k 1} - y_{m_k 2} \text{ if } k \in \text{nr and } i = 2 \end{cases}$$

m_k is defined like j_k above, independently for $i = 1$ and 2.

T is used to represent the m_k selection.

$\hat{d}_i^T = \sum_{k \in s} w_k \tilde{d}_{ki}^T = \hat{y}_{2i}^T - \hat{y}_{1i}^T$

$$\tilde{y}_{ki} = \begin{cases} y_k & \text{if } k \in \text{r} \\ y_{k1} & \text{if } k \in \text{nr and } i = 1 \\ y_{k2} & \text{if } k \in \text{nr and } i = 2 \end{cases}$$

$\hat{y}_i = \sum_{k \in s} w_k \tilde{y}_{ki}$,

$\bar{y}_\bullet = 1/2(\hat{y}_1 + \hat{y}_2)$

3.3 Terms used in section 5

$$\tilde{d}_k = \begin{cases} 0 \text{ if } k \in \text{r} \\ y_{k2} - y_{k1} \text{ if } k \in \text{nr and } y_{k1} \text{ is used in } \hat{y}_\bullet \\ y_{k1} - y_{k2} \text{ if } k \in \text{nr and } y_{k2} \text{ is used in } \hat{y}_\bullet \end{cases}$$

y_k' is the term subtracted in $\tilde{d}_k$

$\hat{y}' = \sum_{k \in s} w_k y_k'$

$\hat{d} = \sum_{k \in s} w_k \tilde{d}_k$

4. Measuring the Imputation Variance

The goal of this paper is to present a methodology which measures the imputation variance using standard replication variance software packages, in a simple manner. One way of doing this is adding an independent residual term ($\tilde{d}_k^R$) to the data elements ($\tilde{y}_k$), so that an appropriate amount of imputation variance is added. The estimate ($\hat{y}_\bullet$) is transformed into $\hat{Y} = \hat{y}_\bullet + \hat{d}^R$. An appropriate constant is added to the $\tilde{d}_k^R$'s, so that $\hat{Y} = \hat{y}_\bullet$. Now, the question is when does $V(\hat{Y})$ appropriately measure the true variance. When $V(\hat{Y})$ is not appropriate, what must be done to make it appropriate. The first step is to compute $V(\hat{Y})$.

4.1 Computation of Variance of $\hat{Y}$

Let $\hat{Y} = \sum_{k \in s} w_k (\tilde{y}_k + \tilde{d}_k^R) = \hat{y}_\bullet + \hat{d}^R$.

$$V(\hat{Y}) = \underset{1}{V}\,\underset{2}{E}(\hat{Y}) + \underset{1}{E}\,\underset{2}{V}(\hat{Y}), \quad (1)$$

where: 1 represents the selection of the sample s and 2 represents the imputation selection of I_k and residual selection R.

4.2 Computation of $\underset{1}{V}\,\underset{2}{E}(\hat{Y})$

$$\underset{2}{E}(\hat{Y}) = 1/2(\hat{y}_1 + \hat{d}_1^T) + 1/2(\hat{y}_2 + \hat{d}_2^T)$$

d_i^T has been normalized to equal zero

Looking at the right hand term first:

$$1/4\,\underset{1}{V}(\hat{y}_1 + \hat{d}_1^T) = 1/4\,\underset{1}{V}(\hat{y}_1 + \hat{y}_{21}^T - \hat{y} + \hat{y} - \hat{y}_{11}^T)$$

$$= 1/4[\underset{1}{V}(\hat{y}_1) + \underset{1}{V}(\hat{y}_{21}^T - \hat{y}) + \underset{1}{V}(\hat{y} - \hat{y}_{11}^T)$$

$$+ 2\,\underset{1}{\text{cov}}(\hat{y}_{21}^T - \hat{y}, \hat{y} - \hat{y}_{11}^T)], \text{ since } \text{cov}(\hat{y}_1, \hat{d}_1^T) = 0$$

$$= 1/4[\underset{1}{V}(\hat{y}_1 - (\hat{y}_{21}^T - \hat{y})) + \underset{1}{V}(\hat{y} - \hat{y}_{11}^T) \quad (2)$$

$$+ 2\,\underset{1}{\text{cov}}(\hat{y}_{21}^T - \hat{y}, \hat{y})], \text{ assuming } \hat{y}_{11}^T \text{ is distributed as } \hat{y}_1$$

Likewise,

$$1/4\,\underset{1}{V}(\hat{y}_2 + \hat{d}_2^T) = 1/4[\underset{1}{V}(\hat{y}_2 - (\hat{y}_{22}^T - \hat{y})) + \underset{1}{V}(\hat{y} - \hat{y}_{12}^T)$$

$$+ 2\,\underset{1}{\text{cov}}(\hat{y}_{22}^T - \hat{y}, \hat{y})] \quad (3)$$

Combining 2 and 3 gives:

$$\underset{1}{V}\,\underset{2}{E}(\hat{Y}) = \underset{1}{V}(\hat{y}) + \underset{1}{V}(\hat{y} - \bar{y}_\bullet) + 2\,\underset{1}{\text{cov}}(\bar{y}_\bullet - \hat{y}, \hat{y}), \quad (4)$$

assuming that $\hat{y}_{21}^T$ and $\hat{y}_{12}^T$ are distributed as $\hat{y}_2$; and $\hat{y}_{11}^T$ and $\hat{y}_{22}^T$ are distributed as $\hat{y}_1$;.

$\text{cov}(\hat{y} + \hat{y}_1 - \hat{y}_{21}^T, \hat{y} + \hat{y}_2 - \hat{y}_{22}^T)$ and $\text{cov}(\hat{y} - \hat{y}_{11}^T, \hat{y} - \hat{y}_{12}^T)$ are zero by the independence in the T selection

4.3 Computation of $V_2(\hat{Y})$

$$V_2(\hat{Y}) = V_2(\hat{y}_\bullet) + V_2(\hat{d}^R)$$
$$= V_2(\hat{y}_\bullet) + V_2^{ppz}(\hat{d}^R) \quad (5)$$

V_2^{ppz} is the variance associated with probability proportionate to size sampling with replacement. The size is the w_k's for $k \in$ nr within each imputation cell.

4.4 Computation of $V(\hat{Y})$

Putting 1, 4 and 5 together gives:

$$V(\hat{Y}) = V_1(\hat{y}) + V_1(\hat{y} - \bar{y}_\bullet) + 2\,\mathrm{cov}_1(\bar{y}_\bullet - \hat{y}, \hat{y})$$
$$+ E_1 V_2(\hat{y}_\bullet) + E_1 V_2^{ppz}(\hat{d}^R)$$
$$= V(\hat{y}) + V(\hat{y} - \hat{y}_\bullet) + 2\,\mathrm{cov}(\bar{y}_\bullet - \hat{y}, \hat{y})$$
$$+ E_1 V_2^{ppz}(\hat{d}^R)$$

As formulated by Särndal (1994), $V(\hat{y})$ is the variance assuming no nonresponse and $V(\hat{y} - \hat{y}_\bullet)$ is the imputation variance. The sum of these components equals the total variance, assuming the $\mathrm{cov}(\hat{y}, \hat{y} - \hat{y}_\bullet) = 0$. The simulation done in Särndal (1994), using nearest neighbor imputation, indicates that the covariance term is small and can be approximated by zero.

$2\,\mathrm{cov}(\bar{y}_\bullet - \hat{y}, \hat{y})$ and $E_1 V_2^{ppz}(\hat{d}^R)$ are terms not included in the estimate of the true variance. There are two ways of handling these terms. The first is to estimate the terms and subtract them from $V(\hat{Y})$. The second way is realizing that the only non-zero values in these terms comes from nonrespondents and if the item response rate is relatively high, these terms should be small and can be ignored.

5. **Computation of** $\hat{\mathrm{cov}}(\bar{y}_\bullet - \hat{y}, \hat{y})$

Since $\mathrm{cov}(\bar{y}_\bullet - \hat{y}, \hat{y})$ requires y_k for $k \in$ nr, it can not be computed with the given sample. However, it can be approximated using $\tilde{d}_k$. If $\tilde{d}_k = y_{k2} - y_{k1}$, and given how $\tilde{d}_k$ is formed, y_{k2} is the nearest neighbor imputation for y_{k1}. Therefore, $\tilde{d}_k$ can be viewed as the difference between the nearest neighbor imputation for a known value and the known value. Since $\tilde{d}_k$, in terms of the nearest neighbor imputation, is close to unit k, the covariance might be approximated by $\mathrm{cov}(\hat{d}, \hat{y}')$.

6. **Computation of** $V_2^{ppz}(\hat{d}^R)$

$E_1 V_2^{ppz}(\hat{d}^R)$ is estimated from a sample by computing $V_2^{ppz}(\hat{d}^R)$ for that sample. This is done by either using an exact formula (see Cochran, 1977) or by replication given a set of replicate weights designed for the conditional variance. For this report the exact formula will be used.

7. **Simulations**

In order to determine how well the proposed imputations work, a simulation study is performed for two variations of the proposed variance methodology, using four states within the SASS sample design and performing 5,000 simulations.

For each simulation sample, a nearest neighbor imputation ($\tilde{y}_k$) is performed on the selected sample of nonrespondents. Three populations of nonrespondents with 5%, 15% and 30% nonresponse will be generated and compared.

The simulation estimates are based on the design and collection variables that can be found on the frame. In this way, estimates for any selected sample, as well as, estimates of the true variance, can be computed.

7.1 Population of Nonrespondents

In order to do the analysis, a population of nonrespondents must be defined. A school k is chosen to be a member of the nonrespondent population by independently selecting k, proportional to P_k^{NR}. P_k^{NR} is determined to obtain an expected X% unweighted nonresponse rate in s.

$$P_k^{NR} = \left(p_k / \sum_{k \in c} p_k^2 \right) \times N_c^{NR}$$

p_k is the probability of selecting k in s

c is an imputation cell

N_c^{NR} is the expected number of nonrespondents in c

$$N_c^{NR} = \left(en_c r_c N_l^{NR} \right) / \left(\sum_{c \in l} en_c r_c \right)$$

en_c is the expected sample size in c (i.e., $\sum_{k \in c} p_k$)

r_c is the relative rate of nonresponse in c (i.e., 2 for central city schools, 1 otherwise)

N_l^{NR} is the expected number of nonrespondents required for the analysis in stratum l

$$N_l^{NR} = \left(\sum_{c \in l} en_c \right) \times (\mathrm{X}/100)$$

It is easy to verify: 1) the expected nonresponse rate for s will be X%; 2) the relative rate of nonresponse in central city cells is twice that of other cells; and 3) larger schools, as measured by p_k, will have a higher probability of being selected as nonrespondents.

7.2 Simulation Sample Design

The Schools and Staffing Survey (SASS) is a stratified probability proportionate to size sample of elementary, secondary and combined schools. The selection is done systematically using the square root of the number of teachers per school as the measure of size. State by school level cells define the stratification. Before selection, schools are sorted to provide a good geographic distribution. Estimates are designed to provide state estimates.

In order to assure unbiased variance estimation using half-sample replication, the simulation design has been slightly altered. Each state/school level stratum has been split into a number of strata so that exactly two schools are selected within each stratum, while maintaining the original sample size. Another modification is that the selection within stratum is done with replacement.

7.3 Estimates

Three estimates per state are computed:

$$\hat{y}_e = \sum_{k \in s} S_k p_{ke}; \ \hat{y}_m = \sum_{k \in s} S_k p_{km}; \text{ and}$$

$$\hat{y}_h = \sum_{k \in s} S_k p_{kh}.$$

S_k is the known number of students in school k

p_{ke} is the proportion of the students in school k in grades pre-kindergarten to 6

p_{km} is the proportion of the students in school k in grades 7 to 9

p_{kh} is the proportion of the students in school k in grades 10 to 12

It is assumed that S_k is known for all k and that only the p's require collection. Therefore, when k requires imputation, a nearest neighbor donor's p will be applied to S_k. This is a common SASS imputation that can be duplicated from the frame.

7.4 Simulated Variance Estimates

All variances below, except for $V_T(\hat{y}_\bullet)$ and $V_2^{ppz}(\hat{d}^R)$, are estimated using a fully balanced half-sample replication variance estimator (Wolter,1985).

The four variance estimates computed within each sample and averaged across samples are:

$$\hat{V}_R = \hat{V}(\hat{Y})$$

$$\hat{V}_{EC}(\hat{Y}) = \hat{V}(\hat{Y}) - 2\,\hat{\text{cov}}(\bar{y}_\bullet - \hat{y}, \hat{y}) - V_2^{ppz}(\hat{d}^R)$$

$\hat{V}(\hat{y})$ is the variance estimate when all sample cases respond

$\hat{V}(\hat{y}_\bullet)$ is the variance estimate of $\hat{y}_\bullet$

The following estimates can not be computed from a single sample. However, they can be computed in the simulation setting and are used for comparison purposes. The first is the variance estimate using a true estimate of the covariance and the second is an estimate of the true variance.

$$\hat{V}_{TC}(\hat{Y}) = \hat{V}(\hat{Y}) - 2\,\text{cov}(\bar{y}_\bullet - \hat{y}, \hat{y}) - V_2^{ppz}(\hat{d}^R)$$

$$V_T(\hat{y}_\bullet) = 1/n \sum_{s=1}^{n} (\hat{y}_{\bullet s} - \bar{y}_{\bullet s})^2$$

$\hat{y}_{\bullet s}$ and $\bar{y}_{\bullet s}$ are the value of $\hat{y}_\bullet$ for the s^{th} simulation and the average of the $\hat{y}_{\bullet s}$, respectively.

8. **Analysis Statistics**

To evaluate the imputation methodology relative bias of the estimated standard error (RB) and coverage rates (C_t and C_m) are computed.

$$RB = \left(\sqrt{V_i(\hat{Y})} - \sqrt{V_T(\hat{y}_\bullet)}\right) / \sqrt{V_T(\hat{y}_\bullet)}$$

$V_i(\hat{Y})$ is one of the variance estimates defined above

C_t is the coverage rate testing whether the true estimate with complete response is in the 95% confidence interval

C_m is the coverage rate testing whether the average of the simulated estimates ($\bar{y}_{\bullet s}$) is in the 95% confidence interval

Since the estimates with imputation are biased, the levels for C_t are unknown. One however, expects them to be smaller than 95%. Since C_m should be closer to 95%, it is used for comparison.

9. **Results**

Tables 1-3 provide results for the populations of nonrespondents, 30%, 15% and 5%, respectively. By comparing C_t and C_m for $\hat{V}(\hat{y}_\bullet)$, it can be seen that the bias in $\hat{y}_\bullet$, relative to the coverage rates, is small. The only exception is state 24's $\hat{y}_h$ in table 1, where the coverage rates differ by 9 points (90-81).

The half-sample coverage rates (C_t's for $\hat{V}(\hat{y})$) are usually less than 95%. Since C_t for $\hat{V}(\hat{y})$ provides the coverage rate with complete response, it

will be assumed that the best imputation coverage rate, instead of being the one closest to 95%, will be the coverage rate coming from the procedure that gets closest to the C_t's for $\hat{V}(\hat{y})$.

From table 1, it can be seen that when the nonresponse rate is high $\hat{V}_{EC}(\hat{Y})$ is superior to $\hat{V}_R(\hat{Y})$. This is true for both relative bias and coverage rates. $\hat{V}_{EC}(\hat{Y})$ has better coverage rates 11 times, while $\hat{V}_R(\hat{Y})$ is better only once; $\hat{V}_{EC}(\hat{Y})$ has smaller relative biases 9 times, while $\hat{V}_R(\hat{Y})$ is better only 3 times. Therefore, $\hat{V}_{EC}(\hat{Y})$ should be used when the item nonresponse rate is high.

When the nonresponse rate is moderate in size, tables 2 show that $\hat{V}_R(\hat{Y})$ and $\hat{V}_{EC}(\hat{Y})$ are comparable with respect to coverage rates. $\hat{V}_{EC}(\hat{Y})$ is better or equal to $\hat{V}_R(\hat{Y})$ 7 times, while $\hat{V}_R(\hat{Y})$ is also better or equal to $\hat{V}_{EC}(\hat{Y})$ 7 times. With respect to relative bias, $\hat{V}_R(\hat{Y})$ is slightly better than $\hat{V}_{EC}(\hat{Y})$. $\hat{V}_R(\hat{Y})$ has smaller or equal relative biases 8 times, compared to $\hat{V}_{EC}(\hat{Y})$ being better or equal 5 times. Hence, when computation simplicity is important, $\hat{V}_R(\hat{Y})$ can be used. However, it may still be safer to use $\hat{V}_{EC}(\hat{Y})$.

When the nonresponse rate is low, table 3 shows that $\hat{V}_{EC}(\hat{Y})$ has better coverage rates than $\hat{V}_R(\hat{Y})$. $\hat{V}_{EC}(\hat{Y})$ is better or equal to $\hat{V}_R(\hat{Y})$ 10 times, while $\hat{V}_R(\hat{Y})$ is better or equal to $\hat{V}_{EC}(\hat{Y})$ 7 times. However, $\hat{V}_R(\hat{Y})$ has smaller relative biases. $\hat{V}_R(\hat{Y})$ is better or equal 9 times, while $\hat{V}_{EC}(\hat{Y})$ is better or equal 4 times. This probably means that it is safe to assume the terms subtracted in $\hat{V}_{EC}(\hat{Y})$ are close to zero and $\hat{V}_R(\hat{Y})$ can safely be used to estimate the total variance. However, since $\hat{V}_{EC}(\hat{Y})$ has better coverage rates, it should still be considered.

Tables 1-3 shows that $\hat{V}_{EC}(\hat{Y})$ and $\hat{V}_{TC}(\hat{Y})$ work equally well for relative biases. This means that for estimating the total variance, $\hat{\text{cov}}(\hat{d}, \hat{y}')$ is a good approximation for $\text{cov}(\bar{y}_{\bullet} - \hat{y}, \hat{y})$.

10. Conclusions

For the design and imputations simulated, $\hat{V}_{EC}(\hat{Y})$ should be used when the item nonresponse rate is high, although $\hat{V}_R(\hat{Y})$ is still an improvement over $\hat{V}(\hat{y}_{\bullet})$. When the item response rate is moderate or low $\hat{V}_R(\hat{Y})$ provides good results and given its simplicity, can be used to estimate the total variance.

Table - 1 Relative Bias (RB) and Coverage Rates (C_T and C_m) for Population with 30% Nonresponse

30% Nonresponse		Percent RB (C_T)		Percent RB (C_m)			
State	Estimate	$\hat{V}(\hat{y})$	$\hat{V}(\hat{y}_{\bullet})$	$\hat{V}(\hat{y}_{\bullet})$	$\hat{V}_{TC}(\hat{Y})$	$\hat{V}_{EC}(\hat{Y})$	$\hat{V}_R(\hat{Y})$
2	$\hat{y}_e$	0 (94)	-13 (90)	-13 (90)	0 (91)	9 (94)	2 (93)
	$\hat{y}_m$	1 (93)	-25 (81)	-25 (83)	-3 (90)	9 (93)	-3 (90)
	$\hat{y}_h$	1 (91)	-28 (80)	-28 (81)	-11 (85)	-5 (89)	-14 (86)
9	$\hat{y}_e$	1 (94)	-28 (83)	-28 (83)	2 (94)	0 (94)	-13 (89)
	$\hat{y}_m$	0 (94)	-29 (82)	-29 (82)	1 (93)	0 (93)	-14 (88)
	$\hat{y}_h$	1 (94)	-15 (88)	-15 (89)	3 (94)	-1 (93)	-7 (91)
10	$\hat{y}_e$	1 (91)	-22 (85)	-22 (85)	-7 (90)	-5 (91)	-13 (88)
	$\hat{y}_m$	0 (92)	-26 (83)	-26 (83)	-6 (91)	-3 (91)	-13 (88)
	$\hat{y}_h$	1 (87)	-31 (72)	-31 (73)	-1 (88)	-10 (83)	-21 (77)
24	$\hat{y}_e$	-1 (94)	-30 (82)	-30 (81)	-3 (93)	-1 (93)	-15 (88)
	$\hat{y}_m$	0 (93)	-29 (81)	-29 (81)	-2 (93)	2 (94)	-13 (88)
	$\hat{y}_h$	0 (93)	-13 (81)	-13 (90)	2 (93)	15 (96)	2 (93)

Table - 2 Relative Bias (RB) and Coverage Rates (C_T and C_m) for Population with 15% Nonresponse

15% Nonresponse		Percent RB (C_T)		Percent RB (C_m)			
State	Estimate	$\hat{V}(\hat{y})$	$\hat{V}(\hat{y}_\bullet)$	$\hat{V}(\hat{y}_\bullet)$	$\hat{V}_{TC}(\hat{Y})$	$\hat{V}_{EC}(\hat{Y})$	$\hat{V}_R(\hat{Y})$
2	$\hat{y}_e$	0 (94)	-6 (92)	-6 (92)	2 (93)	5 (95)	1 (94)
	$\hat{y}_m$	1 (93)	-9 (89)	-9 (91)	2 (92)	8 (93)	0 (92)
	$\hat{y}_h$	1 (91)	-1 (91)	-1 (92)	2 (91)	9 (93)	3 (92)
9	$\hat{y}_e$	1 (94)	-10 (90)	-10 (91)	4 (94)	2 (94)	-4 (92)
	$\hat{y}_m$	0 (94)	-11 (89)	-11 (90)	6 (95)	4 (94)	-4 (92)
	$\hat{y}_h$	1 (94)	-7 (92)	-7 (92)	7 (95)	3 (94)	-2 (93)
10	$\hat{y}_e$	1 (91)	-9 (89)	-9 (90)	2 (92)	3 (93)	-4 (91)
	$\hat{y}_m$	0 (92)	-10 (89)	-10 (91)	1 (93)	1 (94)	-5 (92)
	$\hat{y}_h$	-1 (87)	-10 (82)	-10 (86)	7 (90)	-3 (86)	-9 (86)
24	$\hat{y}_e$	-1 (94)	-14 (90)	-14 (90)	7 (94)	8 (94)	-4 (92)
	$\hat{y}_m$	0 (93)	-13 (90)	-13 (90)	8 (95)	11 (95)	-1 (93)
	$\hat{y}_h$	0 (93)	-5 (92)	-5 (91)	5 (93)	5 (93)	0 (93)

Table - 3 Relative Bias (RB) and Coverage Rates (C_T and C_m) for Population with 5% Nonresponse

5% Nonresponse		Percent RB (C_T)		Percent RB (C_m)			
State	Estimate	$\hat{V}(\hat{y})$	$\hat{V}(\hat{y}_\bullet)$	$\hat{V}(\hat{y}_\bullet)$	$\hat{V}_{TC}(\hat{Y})$	$\hat{V}_{EC}(\hat{Y})$	$\hat{V}_R(\hat{Y})$
2	$\hat{y}_e$	0 (94)	-4 (92)	-4 (93)	-1 (94)	-1 (94)	-2 (93)
	$\hat{y}_m$	1 (93)	-7 (88)	-7 (91)	11 (95)	8 (93)	0 (92)
	$\hat{y}_h$	1 (91)	-6 (88)	-6 (90)	-6 (87)	3 (91)	-2 (91)
9	$\hat{y}_e$	1 (94)	-3 (93)	-3 (93)	6 (95)	5 (95)	-1 (94)
	$\hat{y}_m$	0 (94)	-5 (92)	-5 (92)	2 (94)	3 (94)	-2 (93)
	$\hat{y}_h$	1 (94)	-1 (94)	-1 (94)	0 (93)	0 (93)	-1 (93)
10	$\hat{y}_e$	1 (91)	-2 (91)	-2 (91)	5 (92)	2 (91)	-1 (91)
	$\hat{y}_m$	0 (92)	-1 (92)	-1 (92)	2 (92)	2 (93)	0 (92)
	$\hat{y}_h$	-1 (87)	-2 (87)	-2 (87)	-2 (88)	-2 (87)	-2 (87)
24	$\hat{y}_e$	-1 (94)	-7 (91)	-7 (92)	1 (95)	1 (94)	-4 (93)
	$\hat{y}_m$	0 (93)	-5 (90)	-5 (92)	4 (94)	4 (93)	-2 (92)
	$\hat{y}_h$	0 (93)	1 (93)	1 (93)	1 (93)	2 (93)	1 (93)

References

Cochran, W.G. (1977), *Sampling Techniques*, New York: John Wiley & Sons, p253.

Rao, J.N.K., Shao, J (1992), "Jacknife Variance Estimation with Survey Data Under Hot Deck Imputation", *Biometrika*,79, pp. 811-822.

Rubin, D.B. (1978), *Multiple Imputation for Nonresponse in Surveys*, New York: John Wiley & Sons.

Särndal, C.E (1990), "Methods for Estimating the Precision of Survey Estimates When Imputation Has Been Used", *Proceedings of Statistics Canada's Symposium '90: Measurement and Improvement of Data Quality*, pp.369-380.

Särndal, C. E. (1994), "Estimation of the Variance in the Presents of Nearest Neighbor Imputation", *Proceedings of the Section on Survey Research Methods*, ASA Annual Meetings, pp. 888-893.

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

WHERE WILL IT ALL END?
SOME ALTERNATIVE SASS ESTIMATION RESEARCH OPPORTUNITIES

Steven Kaufman and Fritz Scheuren
Fritz Scheuren, The George Washington University, Washington, D.C. 20052-0001

KEY WORDS: GENERALIZED REWEIGHTING, MASS IMPUTATION

1. INTRODUCTION

For 1993-94, both NCES's Schools and Staffing Survey (SASS) and its Private School Survey (PSS) were conducted. SASS has a large private school component as part of its overall sample. PSS is essentially a census of private schools -- but with considerably less item content than SASS. Over the last two years at these Joint Statistical Meetings, a modified Generalized Least Squares (GLS) estimator has been explored in SASS to see if we could achieve simultaneous intersurvey consistency in comparable totals for schools, teachers, and students between SASS and PSS.

We were successful (Scheuren and LI 1996); but, in the end, the results were ultimately disappointing in that so little use had been made of the exceptionally rich PSS data. If our goal had been more general -- say, improving SASS estimates -- not just trying to achieve a limited consistency with PSS, then alternatives to GLS become attractive. In this year's paper we look generally at what can be done when conducting a survey in a data rich setting. These general observations are then applied to private school surveys. Among the methods advocated in this setting, "mass imputation" is given special attention, since it has not been used as often as its considerable strengths might warrant.

Organizationally, the present paper is divided into four parts: This introduction begins the discussion (Section 1). A summary of our results with Generalized Least Squares (GLS) estimators makes up Section 2. Section 3 presents an alternative to reweighting SASS which has been called "mass imputation." This technique, now roughly 20 years old (Colledge et al. 1978), imputes records from a survey back to its sampling frame; and, in a sense, operates in making estimates as if there had been a census. The final section discusses some "What Nexts."

2.GLS ESTIMATION IN SASS

As already noted, a Generalized Least Squares (GLS) technique was used to achieve simultaneous consistency or near consistency in totals for schools, teachers, and students between the private school component of the 1993-4 SASS and the 1993-4 PSS (Scheuren and Li, 1996). This technique is described briefly below (subsections 2.1 and 2.2), then our 1993-4 results summarized (subsection 2.3). A discussion of our (possibly misplaced) expectations concludes the section (in subsection 2.4).

2.1 Generalized Least Squares. -- Advocated by Deville and Särndal (1992), GLS can be used (as in Imbens and Hellerstein 1993) to achieve consistency between SASS and PSS. To see how GLS works in this setting 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,...,n.
- t_i is the SASS total of teachers for ith SASS observation, i=1,...,n.
- s_i is the SASS total of the 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 schools, 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 which can be written as the sum of squares, $\sum (u_i - w_i)^2$. Motivating this loss function here is outside our present scope, except to say that the sensitivity

of the final results to the loss function chosen (e.g.,Deville and Särndal 1992; Deville et al. 1993) seems not to be too great. Now 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's 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:

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

and $\underline{\lambda}$ is the vector of unknown GLS adjustment factors obtained from $\underline{\lambda} = M^{-1}\underline{d}$. 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; and, indeed, had we used another loss function, a weighted version of the **M** matrix could have been employed.

2.2 Olkin Modified GLS. -- Based on concerns about negative weights raised in our pilot application of GLS (Scheuren and Li 1995), it seemed worthwhile to see if a ratio adjustment step could be introduced before the GLS algorithm was employed. An old idea of Olkin (1958) formed our starting point. Assume we have a total τ, say, of student enrollment in the current application. Suppose further, as is the case here, that this is to be estimated from a sample. Following Olkin we tested a multivariate ratio estimator for τ of the form $Y = a_1 R_1 w_. + a_2 R_2 t_. + a_3 R_3 s_.$ where the a_i are positive and add to 1, where the $w_.$, $t_.$, and $s_.$ are sample totals, as before; and the R_i are the ratios $R_1 = S/N$, $R_2 = S/T$, and $R_3 = S/S$. In our application, the a_i were 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 is first carried out for small (appropriate) subdomains, then there would be a direct benefit from this step in those subdomains; and, among other things, the number of negative weights reduced..

2.3 Results of 1993-94 GLS Reweighting. -- For the nine typologies that make up the private school population separate GLS and Olkin GLS reweighting attempts were made. Sometimes this was straightforward; sometimes extremely difficult. In each case the complex nature of the PSS and SASS sample designs was considered, operational problems were documented, and independent comparisons were made to PSS school size and community type information. Measures of benefit and harm could be developed because of the comparisons possible. Extensive tabular, graphical, and analytic material have been looked at in making the assessments required (Scheuren and Li 1996). Table A at the end of this paper summarizes the results.

Our operational assessment of the Olkin GLS adjustment to SASS was judged to be good to excellent. In only one case, that for the Other Unaffiliated typology was the evidence unclear. We consider this typology "unclear" because the Olkin GLS did not work without a considerable amount of *ad hoc* tinkering.

Based on the independent assessment by community type and school size, the Olkin GLS seemed to do no apparent harm and may have even been of benefit -- beyond the basic consistency achieved with PSS. However, in the independent assessment we never judged the Olkin GLS as "excellent" because, especially by community type, the Olkin GLS was never best overall. Regularly, it did a "good" job, usually by school size, but even here the performance was less than hoped. In three cases, the Olkin GLS was judged only "fair." These were instances where very mixed results were achieved: Some estimates much improved, others quite adversely impacted.

2.4 Discussion. -- While a detailed comparison was also conducted for the Basic GLS, we have omitted any detailed comment on it here, because generally the Basic GLS was inferior to the Olkin GLS.

Frankly, as already noted, our expectations were really not satisfied, even in the Olkin GLS case. Upon reflection, we feel that part of our problem is that the expectations were misplaced. After all, why should introducing just three totals from PSS make a big improvement in SASS. Conversely, why should such a seemingly small change sometimes be so hard?

3 MASS IMPUTATION ALTERNATIVE

Because the positive benefits of the Olkin GLS were often disappointing, we began to question the entire reweighting approach. In order to avoid negative and small weights we had already set aside a few of the largest schools (about a half dozen to a dozen per typology) to be imputed rather than reweighted. We asked ourselves why not do more than just a few? In fact, why not impute the entire SASS file to the PSS in order to take full advantage of the opportunity that having PSS and SASS fielded for the same year offered? In other words, why not do "mass imputation"?

In particular, let us suppose that mass imputation were to be conducted as part of an overall change in SASS estimation. How would it be done? Suppose, for the sake of discussion, that we had PSS and SASS done in the same year. What would the steps be?

Take a specific typology, "Other Religious Unaffiliated" Schools. For the 1993-94 round of SASS, there were 329 schools in the SASS sample with this designation. In the corresponding PSS for the same period, there were 3,193 such schools. The original SASS estimate of students in other religious unaffiliated schools was 462,934. From PSS, the estimate was 37,578 smaller-- at 425,356 students.

An Olkin GLS reweighting approach was taken to this problem to "solve it." However, as noted earlier, we were not satisfied that we had done enough to use the PSS data to improve SASS. If number of students was the major predictive variable, a sensible mass imputation method that could be applied would be to simply impute the SASS records to nearby PSS cases where nearness is defined simply by student enrollment. For parts of the distribution where the SASS sample is sparse, the SASS observation could be used over and over as a donor perhaps up to, say, 1.5 times its original SASS weight.[1] Conversely, in parts of the distribution where there were lots of SASS cases relative to those in the PSS, the SASS cases would be used as donors less often than their original SASS weights would suggest. The SASS observation would always be used at least once, of course, to represent itself.

It may be useful to think of choosing a mass imputation approach after successively imputing SASS to each of the PSS variables separately and looking at how often each SASS observation was used as a donor. If this range of donor use is not too large, then a single, perhaps nearest neighbor, imputation model could work well. Widely discrepant values in terms of donor use would suggest that the imputation is sensitive to one's beliefs as to the predictive power of the variables being used in the imputation. In such settings a case can be made for doing several different imputations that might be made available to the final users for possibly different purposes. It may even make sense simply to use some convex combination of the separate imputations

In Kovar and Whitridge (1995), there is an excellent discussion of mass imputation. Among other things, they comment on the parallels that can exist between weighting and imputation. They call attention to the work of Folsom (1981) in this connection. Evidence that imputation model sensitivity can be a serious problem exists, as they point out -- citing Cox and Cohen (1985), among others.

Difficulties exist in calculating variances and covariances when using mass imputation. A multiple imputation approach to their estimation has been advocated (Rubin 1996) and could be workable since by design the missing data is missing at random.

In another application, by Wong and Ho (1991), bootstrapping was employed successfully. We think a form of bootstrapping might be the best approach for SASS. The presentation by Kaufman (1996), also given at these meetings, presents related work.

4. POSSIBLE NEXT STEPS

In this paper we have brainstormed at length about possible improvements in SASS that could be undertaken -- only one small aspect of which has been summarized in these Proceedings (See Scheuren 1996 for further details). We are not of one mind as to next steps. For one of us, some effort to try out mass imputation seems warranted. Maybe a single typology should be taken and alternative approaches tried out. We both agree that mass imputation could be a lot of fun.

On the other hand, efforts at reweighting SASS have not all been explored and an additional look at some form of poststratification, using GLS or some other calibration estimator makes a lot of sense too, especially since it is unlikely that PSS and SASS will ever be fielded again in the same year.

---

[1] This factor, about 1.5, is clearly arbitrary and depends on how much of a potential variance price one is willing to pay to get the "nearness" desired. In many weighting settings (e.g., Oh and Scheuren 1987), truncating factors under the square root of 2 work well.

REFERENCES

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

Colledge, M., Johnson, J., Pare, R. and Sande, I. (1978). "Large Scale Imputation of Survey Data," *Proceedings of the Survey Research Methods Section, American Statistical Association.*

Cox, B. and Cohen, S. (1985). *Methodological Issues for Health Care Surveys*, Marcel Decker: New York.

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.

Folsom, R. (1981). "The Equivalence of Generalized Double Sampling Regression Estimators, Weight Adjustments and Randomized Hot Deck Imputations, " *Proceedings of the Survey Research Methods Section, American Statistical Association.*

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

Kaufman, S. (1996). *Properties of the Schools and Staffing Survey's Bootstrap Estimator in Nearest Neighbor Matching.* Paper given at the Chicago meetings of the American Statistical Association.

Kovar, J. and Whitridge, P. (1995). "Imputation of Business Survey Data," in Cox, Binder, Chinnappa, Christianson, Colledge, and Kott, eds. *Business Survey Methods*, Wiley: New York.

Oh, H.L. and Scheuren, F. (1987). "Modified Raking Ratio Estimation in the Corporate Statistics of Income," *Survey Methodology*.

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

Scheuren, F., and Li, B. (1995). *Intersurvey Consistency in NCES Private School Surveys.* Prepared for U.S. Department of Education, National Center for Education Statistics.

Scheuren, F. and Li, B. (1996). *Intersurvey Consistency in NCES Private School Surveys* for 1993-94. Prepared for U.S. Department of Education, National Center for Education Statistics.

Whitridge, P. Bureau, M. and Kovar, J. (1990). *Mass Imputation at Statistics Canada.* U.S. Bureau of the Census, Sixth Annual Research Conference.

Wong, W. and Ho, C. 1991. "Bootstrapping Post-Stratification and Regression Estimates from a Highly Skewed Distribution," *1991 Proceedings of the American Statistical Association,* Section on Survey Research Methods.

Table A.– Olkin GLS Comparisons to Original Weighted SASS Data, By Typology

SASS Typology	Operational Assessment	Independent Assessment
Catholic Parochial	excellent	good
Catholic Diocesan	excellent	fair
Catholic Private	excellent	good
Conservative Christian	good	fair
Other Affiliated	excellent	good
Other Unaffiliated	unclear	good
Non-sectarian Regular	good	good
Non-sectarian Special Emphasis	good	good
Non-sectarian Special Education	good	fair

Notes: The admittedly subjective conventions employed in table A were devised to separate typologies by level of perceived difficulty or benefit. This was done as follows:

(1) Operationally typologies where a simple visual inspection was all that was needed to remove outliers are labelled "excellent" in the operational assessment column.

(2) Typologies labelled "good" operationally were ones where an analytic (potentially iterative) process was required to identify SASS cases that might best be treated by imputation to similar PSS cases rather than being reweighted.

(3) Operationally, only in one case, that of the "Other Unaffiliated" typology was the label "unclear" used. This was done because constructing the Olkin GLS weights proved enormously difficult and required great patience and persistence. (Parenthetically, this typology may, also, have been the most instructive in terms of learning more about how to employ the GLS.)

(4) Based on the independent assessment by community type and school size, the Olkin GLS seemed to do no apparent harm and may have even been of benefit -- beyond the basic consistency achieved with PSS. The comparisons made are to the original SASS weighted data.

(5) The independent assessment column was never coded "excellent" because, especially by community type, the Olkin GLS was never best overall. Regularly, it did a "good" job, usually by school size, but even here the performance was less than hoped. In three cases, the Olkin GLS was judged only "fair." These were instances where very mixed results were achieved: some estimates much improved, others quite negatively impacted.

This page intentionally left blank.

Estimating State Totals from the Private School Universe Survey

Easley Hoy, Beverley Causey, Leroy Bailey, Bureau of the Census, Steven Kaufman, NCES
Easley Hoy, Bureau of the Census, Room 3000-4, Washington, DC 20233

Key Words: Coverage Adjustment, Indirect Estimation, Raking

I. Background and History

The Private School Survey (PSS) is designed and conducted by the Census Bureau for the National Center for Education Statistic to collect data for all private secondary schools in 50 states and D.C. Every two years the survey collects data in an attempt to obtain a complete count for all private schools along with counts of students, teachers, and graduates.

The survey collects data from an administrative list of private schools. To improve the coverage of this list frame, additional lists were obtained from private school associations, state records, and other sources. These lists were matched and unduplicated with the list frame. These operations added about 4900 schools for 1991 and about 2300 schools for 1993. Despite these efforts, the private school list frame remains incomplete, with around 8% of private schools missing from the list. The list enumeration estimates are therefore supplemented by a followup area sample that aims to find and represent unlisted private schools. While direct estimation from the followup area sample produces estimates for unlisted schools of adequate precision for the four geographical regions, it fails to do so for individual states. This paper reports the empirical results of an alternative method for providing estimates of such state totals.

II. Current Methodology from the PSS

For this follow up survey, a stratified sample of primary sampling units (PSUs) is drawn with probability proportionate to size (PSS). The PSUs are comprised of counties or groups of counties and the strata cross state boundaries. Eight of the largest counties are included with certainty and about 115 PSUs are selected with noncertainty. In each sample PSU, seven different sources (e.g., yellow pages, local government offices, etc.) were used to identify missing schools which were not on the list frame. For 1991, the search identified a total of 355 missing schools, and for 1993 there were 421 such schools identified.

For direct estimation, each "added school" is first multiplied by its sampling weight (the reciprocal of the PSU's probability of selection). Then the weighted schools are added up to the PSU level, summed over PSUs within a state to obtain state totals, and summed over states to obtain the four region totals of the number of schools, missed by the list frame. Similar estimates can be obtained for students, teachers, and graduates. Such regional totals have adequate precision, but the state totals are dependent upon which of a states' PSUs are selected for the follow up stratified area sample. For the largest states, there should be no problem, as they will surely have at least one sample PSU with which to estimate its private school undercoverage rate. However, for the smaller states, they may not have a sample PSU in the follow up area survey and thus there would be no estimates of uncoverage rate for such states. The direct estimation approach does provide a unbiased estimator of total added schools to an area's total, and the sampling variance is relatively easily estimated.

III. Proposed Methodology

For the indirect estimator, each sample PSU receives the actual number of unweighted added schools from the followup survey. Each nonsample PSU school gets an upward adjustment based on the sample PSUs' estimation of the add rates.

1. Assume for the moment an across the board adjustment. Also suppose that in the area frame sample, the sample PSUs have a total of n schools

For school i let w_i be the weight associated with the school

$y_i = 1$ if school is an add
$= 0$ if school is an original (originally on the list frame)

Let p be the weighted proportion of school adds.

$$p = \left(\sum_{i=1}^{n} w_i y_i \right) / w. \text{ with } w. = \sum_{i=1}^{n} w_i$$

Then q = 1-p is the weighted proportion of original schools. Consequently if 1/q is the ratio of weighted total schools to the number of schools in the original listing. Let r = 1/q be the adjustment factor by which each school outside the sample PSUs is revised upward to reflect undercount. As it turns out, equivalently, we will be estimating for each list school outside the sample PSUs, p/q schools on average to be added. Since

$$r = 1 + \frac{p}{q} = 1 + \frac{p}{1-p} = \frac{1}{q}.$$

2. Knowledge of Postrata

Values of p and r can be quite different for different kinds of schools. NCES has identified 9 groups of such schools:

1. Catholic Parochial
2. Catholic Diocesan
3. Catholic Private
4. Conservative Christian
5. Other Religious, Affiliated
6. Other Religious, Unaffiliated
7. Nonsectarian Regular
8. Nonsectarian Special Emphasis
9. Nonsectarian Special Education

Such groupings helps to distinguish between different "coverage patterns for different schools. Therefore we will develop a value of p within each poststratum j. But also we can use that fact that big schools, (number of students) are easier to find than small schools. So rather than just fit a value p_j for poststratum j we fit a relationship $p_j(x)$ so that a school outside the sample belonging to poststratum j and size x receives an adjustment factor $r_j(x)$

Within the poststratum j we will drop the subscript j and fit a logistical regression.

$p = 1/[1+\exp(A + Bx)]$ subject to the constraints

$$y_i = 0 \qquad \sum w_i R_i = \sum w_i \qquad y=1 \qquad (1)$$

$$y_i = 0 \qquad \sum w_i x_i R_i = \sum w_i x_i \qquad y=1 \qquad (2)$$

Where $R_i = \frac{p_i}{q_i} = \frac{1}{\exp(A + B x_i)}$ and the summation over i on the left hand side of the equation is where $y_i=0$ (original schools) and the summation on the right hand side of the equation is for $y_i=1$ (add schools). We can now solve for coefficients A&B to be used for deriving the adjustment factors. The adjustment for the ith school is $r_i = 1 + R_i$.

Table 1
1993 Coefficients for Linear Adjustment

Poststratum	A	B
1	4.2144	.6137
2	4.5463	.0742
3	1.2842	1.6214
4	2.9289	.2379
5	.9181	1.0255
6	1.6024	.5807
7	1.6024	.5807
8	.9071	.6212
9	1.8757	2.3367

For the estimate of teacher adds, we refit A values with the x variable representing teachers in equation (2). For the estimate of graduate, data were sparse, so we collapsed across poststrata j to one strata.

IV. Results

A. Application Rules of Proposed Methodology for State Totals

The following summarizes the procedure for state estimation in the PSS:

1. A school, its students, teachers and graduates from a sample PSU are included in the adjusted state counts (including adds) but not in original count if it is an add.
2. A shool its students, teachers, and graduates from a non sample PSU are in the original count and with multiplied factor of r in the adjusted count.

For each state, we derived original and adjusted counts of schools, students, teachers, and graduates; also the ratio of students to teachers, and the ratio of adjusted to original counts were computed for each of these categories. In the full paper and the appendix, to be published in a separate document later, the adjustments for all 50 states and D.C. for both 1991 and 1993 will be provided. The range of adjusted values for most states was 8-12% for 1991 and 5-11% for 1993. The 1993 adjustment results for 6 states (California, Indiana, New York, Texas, Vermont, and Wyoming) are provided in Tables A, B, and C at the end of this paper as examples for discussion.

We begin the discussion with Table A. The numbers of Private Schools, Students, Teachers, Graduates, and Student-Teacher Ratio, by States 1993. Note for each state the ratio adjustment for schools is greater than that for students and teachers. For the states of California, New York, and Indiana, the ratio factors are near average. For the states of Texas, Vermont, and Wyoming, the ratio factors are above average.

B. Variance Estimates for Proposed Methodology

For the proposed methodology, there are two components of variance. The first component is the sampling error due to use of sample to estimate parameters A & B of the model. The second component reflects the variability due to the model. It turns out that the second component is many times larger than the first component.

Table B provides the adjusted estimates with standard deviations reflecting both components of variance.

Note that the standard deviations are small relative to the estimates. Also the first component is a smaller proportion to total error.

C. Raking Adjustment to Regional Totals

For precision and consistency sake, we want the

regional totals for the proposed method to equal those for the current method.

Table 2
NCES Totals for Regions

	Northeast	Midwest	South	West
Schools	6183	7146	7558	05207
Students	1275924	1309211	1386268	865039
Teachers	94622	81862	105509	56128
Graduates	77513	60547	67842	36965

For an area sample PSU's actual count, it includes the original list plus the unweighted adds. Subtract from the regional totals the area sample PSU actual count above to obtain a set of reduced regional totals. For each region we rake the nonsample PSU estimates across the board so that their sum equals the reduced total.

Table 2 above gives the 1993 regional totals based on the complete enumeration plus the follow up area sample estimates of the adds. The raking factors associated with the proposed procedure are given in Table 3.

Table 3
Raking Factors for Region

	Northeast	Midwest	South	West
Schools	.9789	.9725	1.0518	.9557
Students	.9933	.9813	1.0156	.9838
Teachers	.9880	.9796	1.0199	.9862
Graduates	.9962	.9973	1.0044	.9862

Note that the factors are all within 5% of 1.0000. It seems in the South our indirect estimator slightly over estimated adds.

Finally, for the raked counts by state for 1993 for the proposed method we have Table C. The table entries which are referred to as scaled counts are the adjusted state counts of Table B after the application of the raking procedure. The indicated errors are root mean square errors. The bias component of the estimates were derived from estimates of bias at the regional level, based on differences between totals using the current estimation procedure (unbiased) and the proposed indirect metnod. These regional bias estimates were proportionately allocated across the respective states.

V. Summary and Recommendations

On the basis of our test results, we recommend the indirect estimation approach to adjust the list frame counts for the PSS along with the raking to regional totals to obtain the adjusted state totals. We would use the poststratification via NCES and size of school vs an across the board adjustment or poststratification based on sample design. We have made preliminary evaluations of the results of the techniques on 1991 and 1993 data and plan to apply them on 1995 data.

VI. References

1. Broughman, S., Gerald, E., Bynum, L. And Stoner, K. (1994), Private School Universe Survey, 1991-92, National Center of Education Statistics, *Statistical Analysis Report NCES 94-350,* Washington, DC; U.S. Department of Education.

2. Hartley, H.O., J.N.K. Rao, and J. Kiefer (1962), "Sampling with Unequal Probabilities and Without Replacement," Annals of Mathematical Statistics, 33, 350-74.

3. Woodruff, R.S. (1971), "A Simple Method for Approximating the Variance of a Complicated Estimate," Journal of the American Statistical Association, 66, 411-14.

Table A
Numbers of Private Schools, Students, Teachers, Graduates, and Student-Teacher Ratio, by States, for 1993

CALIFORNIA	Schools	Students	Teachers	S-T Ratio	Graduates
Unadjusted	3009	562847	34501	16	23746
Adjusted	3224	576047	35718	16	24039
Ratio A/U	1.071609	1.023451	1.035274	0.988580	1.012336
INDIANA	**Schools**	**Students**	**Teachers**	**S-T Ratio**	**Graduates**
Unadjusted	619	91985	6138	14	4012
Adjusted	686	95447	6450	14	4097
Ratio A/U	1.108312	1.037631	1.050840	0.987429	1.021151
NEW YORK	**Schools**	**Students**	**Teachers**	**S-T Ratio**	**Graduates**
Unadjusted	1865	464172	33812	13	25682
Adjusted	1974	472562	34735	13	26048
Ratio A/U	1.058298	1.018076	1.027305	0.991016	1.014259
TEXAS	**Schools**	**Students**	**Teachers**	**S-T Ratio**	**Graduates**
Unadjusted	1024	186975	14529	12	7424
Adjusted	1177	199967	15728	12	7787
Ratio A/U	1.149020	1.069484	1.034274	0.987960	1.048821
VERMONT	**Schools**	**Students**	**Teachers**	**S-T Ratio**	**Graduates**
Unadjusted	84	9107	945	9	1081
Adjusted	98	9648	1021	9	1116
Ratio A/U	1.165207	1.059432	1.080959	0.980085	1.032137
WYOMING	**Schools**	**Students**	**Teachers**	**S-T Ratio**	**Graduates**
Unadjusted	34	1918	167	11	29
Adjusted	40	2112	192	11	35
Ratio A/U	1.185055	1.101031	1.148996	0.958255	1.210001

Table B
Adjusted Figures and Standard Deviations, with Percent of Variance Attributed to 1st Component, for 1993

CALIFORNIA	**Schools**	**Students**	**Teachers**	**S-T Ratio**	**Graduates**
Adjusted	3224	576047	35718	16	24039
Stand. Dev.	33	2967	240	0.0373	104
% 1st Comp.	28.982	15.746	20.568		6.478
INDIANA	**Schools**	**Students**	**Teachers**	**S-T Ratio**	**Graduates**
Adjusted	686	95447	6450	14	4097
Stand. Dev.	16	1213	96	0.0643	58
% 1st . Comp.	39.289	13.057	16.186		2.955
NEW YORK	**Schools**	**Students**	**Teachers**	**S-T Ratio**	**Graduates**
Adjusted	1974	472562	34735	13	26048
Stand. Dev.	19	2512	249	0.0337	131
% 1st Comp.	22.620	8.874	9.750		4.776
TEXAS	**Schools**	**Students**	**Teachers**	**S-T Ratio**	**Graduates**
Adjusted	1177	199967	15728	12	7787
Stand. Dev.	15	1829	160	0.0359	66
% 1st Comp.	23.996	11.081	14.921		4.893
VERMONT	**Schools**	**Students**	**Teachers**	**S-T Ratio**	**Graduates**
Adjusted	98	9648	1021	9	1116
Stand. Dev.	4	272	35	0.0975	38
% 1st Comp.	13.271	6.618	7.326		2.258
WYOMING	**Schools**	**Students**	**Teachers**	**S-T Ratio**	**Graduates**
Adjusted	40	2112	192	11	35
Stand. Dev.	2	93	11	0.2242	5
% 1st Comp.	9.514	3.925	5.630		1.258

Table C
Scaled Counts, and Student-Teacher Ratio, with Associated Errors, by State, for 1993: Proposed Method

CALIFORNIA	**Schools**	**Students**	**Teachers**	**S-T Ratio**	**Graduates**
Scaled Value	3081	566722	35116	16	23707
Error	140.045	9627.003	636.866	0.563	342.895
INDIANA	**Schools**	**Students**	**Teachers**	**S-T Ratio**	**Graduates**
Scaled Value	667	93663	6319	14	4086
Error	24.163	2117.049	159.529	0.853	58.888
NEW YORK	**Schools**	**Students**	**Teachers**	**S-T Ratio**	**Graduates**
Scaled Value	1932	469394	34317	13	25950
Error	44.925	4016.141	481.261	0.281	163.220
TEXAS	**Schools**	**Students**	**Teachers**	**S-T Ratio**	**Graduates**
Scaled Value	1238	203089	16041	12	7821
Error	66.289	3675.202	359.663	0.367	75.029
VERMONT	**Schools**	**Students**	**Teachers**	**S-T Ratio**	**Graduates**
Scaled Value	96	9583	1009	9	1112
Error	4.818	278.540	36.876	0.737	38.791
WYOMING	**Schools**	**Students**	**Teachers**	**S-T Ratio**	**Graduates**
Scaled Value	39	2109	191	11	36
Error	2.814	92.107	11.296	1.828	5.023

EFFECT OF HIGH SCHOOL PROGRAMS ON OUT-MIGRATION OF RURAL GRADUATES

Gary Huang, Synectics for Management Decisions
Michael P. Cohen, National Center for Education Statistics
Stanley Weng, and Fan Zhang, Synectics for Management Decisions
Stanley Weng, Synectics for Management Decisions, Inc., Arlington, VA 22201

Key Words: High School and Beyond, Scale transformation, Multilevel logit modeling, Random intercept, Weighting

This article presents a study which addressed a policy research issue regarding outmigration among rural high school graduates. The study analyzed data from a national longitudinal survey, the High School and Beyond (HS&B) by the National Center for Education Statistics, using two-level logit models for generating reliable estimates of organization effects on individual behavior. This study examined the post-school outmigration pattern in connection to students' coursework and schools' curriculum, focusing on the effect of academic program versus vocational program adjusting for the effects of local labor market condition and student sociodemographic background. This study also systematically examined the implementation of two-level logit models through the software MLn. Related statistical issues were addressed.

1. Introduction

Rural America has been experiencing chronic depression economically and psychologically, as a result of more than a century's industrialization and urbanization (Theobald, 1990). Rural-to-urban migration of educated rural youth is a factor leading to economic marginality and community decline in rural America (McGranahan, D.A. & Ghelfi, L.M., 1991; O'Hare, 1988; Reid, 1990). Some analysts say that the existing public school curriculum has contributed to the problems (Berry, 1990; Theobald, 1992; Snauwaert, 1990; Thompson & Kutach, 1990). Schools, they argue, have failed in educating the young to strengthen their community identity and to preserve the environment and community. Typical rural school programs provide predominatly occupational skills, which essentially prepare students for the urban labor market. Thus, rural schools are contributing to the decline of rural areas (Berry, 1990).

Some rural advocates suggest that it is necessary to strengthen liberal arts and humanities education, rather than occupational training, in rural schools (Berry, 1990, 1989,). The idea has drawn attention from some rural educators who support strengthening liberal arts and humanities curricula (Theobald, 1992; Snauwaert, 1990; Thompson & Kutach, 1990). They believe that liberal education can nurture the understanding and appreciation of the links between man and the land, local heritage and the global civilization, and help students explore meanings of rural life and maintain their attachment to the community. Only with a strong root in the community, is it possible for young people to contribute to the community development while achieving personal wellness.

On the other hand, many aruge that vocational and technical education oriented toward providing marketable skills to students are essential for local community development as well as for individual earnings (e.g., Klerman & Karoly,1995; Vandegrift & Danzig, 1993; Muraskin, 1993). In this perspective, vocational/technical training, often developed to fulfill the local market needs, not only helps students get jobs but also foster entrenprieniorship that contributes to local economic development (Finch, 1993; Sher, 1977).

Further, vocational and technical education may be particularly helpful for youth who are disadvantaged in socioeconomic background, disabilities, and rural locale by training them to learn cognitive skills in job-specific contexts (Teixeira & Swaim, 1991). The programs require a shorter period for student to complete than academic programs do, and allow them to promptly apply learned knowledge and skills to their workplace (Muraskin, 1993; Finch, 1993). Such features in vocational/technical education may contribute to graduates' success in labor market of local areas or elsewhere.

If we are to address problems in rural development, it is critical for research to understand the mechanism in which educated youth are moving out from rural areas (Teixeira, 1993). The question about the functions of the content of rural education regarding outmigration calls for systematic research. This study is a response to such research needs. We recognize, however, the differential benifits of specific school programs to different student subgroups, the unique concerns of local communities, and the diverse contexts wherein schooling takes place. The study is not meant to offer a uniform answer to the question as what kind of program is suited to a particular community.

Prior Studies

The research literature on rural outmigration has been largely motivated by economic and demographic concerns. In dealing with the factors responsible for rural outmigration, research often focused on job market; life quality; housing conditions; and demographic background of migrants such as race, age, income, and residential history (e.g., Cromartie, 1993, Sandefur & Jeon,1991; Voss & Futuitt, 1991; Johnson, 1989; Adamchak, 1987). Educational attainment has been studied as a predictor variable for migration, but the content of schooling in connection to subsequent mobility has been rarely studied. Little is known about the differential effects of secondary school curriculum and student coursework on later geographic mobility.

There is a study available (Pollard, O'Hare & Berg, 1990) that provides a ground for further research. This study examined a series of correlates to post-high school migration with HS&B data, including students' enrollment in different curricular programs. It describes a pattern in which a large portion of outmigrants-- relative to those who stayed in the community-- were enrolled in high school academic programs. The analysis however, is not adequate to answer the questions we pose. While dealing with a large number of variables in predicting migration, it has not focused on the impact of content of schooling, nor differentiated the effects of school program provision vis-a-vis individual student coursework.

Building upon this line of research, we examine the the issue by focusing on the joint effect of student coursework and school instructional programs in predicting post-school moving, adjusting for the effects of student background and community context. With the two perspectives about school curriculum as a conceptual framework, we analyze data simultaneously at student- and school-levels to sytematically investigate the differential effects on outmigration by academic and vocational\tecnical education in terms of both student couretaking and school program provisions.

The analysis links post-school migration to liberal arts education embodied in academic curriculum in contrast to vocational/technical education geared to acquiring occupational skills. Theoretically, academic learning differs from that of job-skill acquisition in influencing students' willingness to stay in the local community. We consider the school "compositional" effect and the effect of individual students' coursework as two distinctive variables contributing to post-school mobility. The compositional effect at the school level can be represented by high enrollment in a particular program. High enrollment in a program reflects the school emphasis on the given program. A large proportion of students in academic program often indicates the school's value of academic learning and its commitment of resource and staff to the program. Moreover, high enrollment in academic program probably suggests an overall school atomosphere pressing student toward academic learning. Likewise, high enrollment in vocational/technical programs represents the school emphaisis on job skill training and likely a strong climate that forsters student interest and effort to acquire such skills.

Normally, students who are in academic programs tend to be involved in more liberal arts education (literature, history, math, and science); whereas students in vocational and technical programs are more likely to learn practical, specific occupational skills. Specific coursetaking also should be a factor influencing the amount of students' exposure to liberal arts and vocational learning.

2. Study Design

2.1 Data

We used HS&B (1980-1992)'s 1980 sophomore cohort data combined with the senior cohort in order to have adequate within-school student samples. For the 1980 sophomore cohort data, we used: (1) the first followup (1982, the senior year) data on coursetaking (from transcripts file), curriculum program (academic vs. vocational), test scores (composite and 1982 IRT scores on reading and math), and other background measures; and (2) the third followup (1986) data on migration. For the senior cohort, similar data from the base year (1980, the senior year) and the second followup were used. Local economic indicators from the Census component were, as school-level variables, attached to each student record of both cohorts.

A group of students whose schools contained less than 3 students in the sample were removed from the analysis. The final file contains 875 schools with a total of 16,492 students. The within-school sample size averaged 18.8, with a standard deviation of 7.78, ranging from 3 to 48. The student sample represents the 1980 and 1982 public high school seniors in the U.S. who participated in the HS&B followup surveys as specified above. The school sample represents the public high schools to which these students went.

To improve the data quality, we edited data, including recoding, rescaling, and imputation for missing values. Data were largely compatible between the two cohorts with one exception, the curriculum credits. For sophomores, an official transcripts survey provided valid records of their coursework and composites for academic and vocational credits were in file. For seniors, only self-reported coursework is available. To retain the good quality of the sophomore transcript data while making the senior self-reported

credit data comparable, we took the following approach: first, scales for senior self-reported academic and vocational credits were developed; second, both senior and sophomore scales were standardized so that the measures were statistically compatible.

2.2 Variables

Following variables were selected for the analysis:

Outcome variable: Outmigration rate (MIGRATE). A binary indicator for outmigrant. It is defined by the student-reported residential location of 50 or more miles away from the original community four years after high school completion.

Student level variables:

Academic coursework measure (ACADCRC). A standardized score of two source of information: seniors' self-reported course-taking and sophomores' academic curriculum credits from the transcripts survey file.

Vocational coursework measure (VOCTCRC). Scaled with the same approach as for academic coursework.

Senior-year test score (FUTESTC). A standardized measure of academic achievement.

Parents' education (PCOLLEGE). As an indicator for student socioeconomic status because of substantial missing of student economical status composite data.

Minority (MINORITY). A dummy variable (Black, Hispanic, Native American and Others versus otherwise).

School level variables:

School academic program enrollment rate (ACADEMM).

School vocational program enrollment rate (VOCATM).

Locale indicator of rural versus others (RURAL).

County-level employment growth rate between 1980 and 1982 (CEMPG02). As an indicator of the local labor market conditions, used as a school-level variable. It was deemed more relevant to post-school outmigration compared with the other three indicators available from the local economic indicators file: metropolitan area employment growth rate, county per capita income relative to the national average, and metropolitan area per capita income relative to the national average.

Technical handling of the data included:

Imputation Missing values for most variables were a small portion and imputed with the sample means. MIGRATE, with approximately 3 percent nonresponse, was examined in relation with race, parents' education, program enrollment, coursework, and local economic indicators. A pattern identified with these predictors was used to generate imputing values. The rate of outmigration among the imputed cases is similar to that of the total sample.

Centering All student-level variables were centered around their school mean. Centered metrics was used since level-2 intra-unit correlation was found substantial (for the debate on the issue, see Plewis, et al., 1989).

2.3 Weighting After exploring a number of weighting procedures, we decided not to use any of them due to the restrictions imposed by the software and logistic regression procedures. The weighting issue will be discussed later in section 5.1 of this article.

3. Multilevel Logit Modeling

Multilevel logit models (Goldstein, 1995) were used to analyze the data. The software MLn (Rasbash, et al., 1995; Woodhouse, et al., 1995) was used to implement the models.

For binary responses y_{ij} for student i in school j, with outcome probability π_{ij}, the two-level logit model is described as follows (Goldstein, 1995). At level-1,

$$y_{ij} = \pi_{ij} + z_{ij}e_{ij}$$

where $z_{ij}e_{ij}$ is the level-1 binomial error, $z_{ij} = [\pi_{ij}(1-\pi_{ij})]^{1/2}$, $\sigma_e^2 = 1$, and

$$\log it(\pi_{ij}) = \log \frac{\pi_{ij}}{1-\pi_{ij}}$$

is modeled by a linear predictor that contains student explanatory variables with coefficients containing random components representing the variation of school characteristics. In a baseline model addressing the question as how much is the variance of outmigration across schools, the linear predictor consists of only a random intercept β_{0j} with

$$\beta_{0j} = \gamma_{00} + u_{0j} \qquad (1)$$

where γ_{00} is the fixed intercept and u_{0j} is the level-2 error associated with the intercept, $u_{0j} \sim N(0, \sigma_{u0}^2)$. More elaborate random-intercept-only model contains school-level explanatory variables:

$$\begin{aligned}\beta_{0j} &= \gamma_{00} + \gamma_{01}(\text{RURAL}) + \gamma_{02}(\text{CEMPG02}) \\ &+ \gamma_{03}(\text{ACADEMM}) + \gamma_{04}(\text{VOCATMM}) + \\ &\gamma_{05}(\text{RU\_ACADM}) + \gamma_{06}(\text{RU\_VOCTM}) + u_{0j}\end{aligned} \qquad (2)$$

addressing such question as how school-level variables contribute to the average outmigration rate, and how they explain the variance of the outmigration rate.

With all the five student variables included in the level-1 model, the linear predictor is formed as

$$\beta_{0j} + \beta_{1j}(\text{ACADCRC}) + \beta_{2j}(\text{VOCTCRC}) + \beta_{3j}(\text{FUTESTC}) + \beta_{4j}(\text{PCOLLEGE}) + \beta_{5j}(\text{MINORITY}).$$

For our data, we found the slopes in the level-1 model don't vary significantly at level-2. Thus we focused on the random-intercept-only model with fixed level-1 covariates.

Table 1 presents the estimates from two random-intercept models: without and with level-1 covariates.

Table 1. Estimates of Two-level Random-Intercept Logit Models (standard error in parenthesis): The sample of the 1980 and 1982 seniors of US public high schools

Estimate	Without student variable	With student variable
Fixed effects		
School mean outmigrt.	1.160(.134)	-1.245(.140)
RURAL	.499(.053)	.809(.214)
CEMPG02	-.009(.004)	-.009(.004)
ACADEMM	.806(.206)	.794(.215)
VOCATMM	-.911(.228)	-.893(.239)
RU_ACADM	-.578(.353)	-.655(.370)
RU_VOCTM	-.135(.385)	-.182(.403)
ACADCRC		.217(.023)
VOCATCRC		-.128(.021)
FUTESTC		.041(.003)
PCOLLEGE		.469(.042)
MINORITY		.042(.050)
Random effects		
School mean outmigrt.	.202(.023)	.232 (.025)
Log likelihood	19747.4	18304.2

Analysis of data on non-college goers .....

Findings from the two analyses are largely consistent. Similar to the pattern found in the total sample analysis, randome-intercept models with non-college goer data confirm that outmigration is positively related to school academic curriculum and student academic coursework (in the model with level-1 covariates, the two estimates are respectively .794 and .217 in logit), and negatively related to school vocational program enrollment and student vocational coursework (respectively, -.449 and -.131). The two analyses also generate similar estimates of the effects of other variables at both levels.

4. Discussion

This study systematically examined the effects of public school curriculum on rural youth outmigration with HS&B logitudinal data. Two-level random-intercept logistic regression models were tested, taking student-level logit of outmigration and school average outmigration as the outcome variables in simultaneous analyses. We decomposed the effects into school- and individual-levels components, controlled for the effects of local employment conditions and student test scores and sociodemographic background, and introduced interaction terms to determine the specific effects of curriculum on rural school average outmigration. .

The analyses revealed that (1) school average outmigration is positively related to the school emphasis on academic programs and negatively related to the school emphasis on vocational programs; (2) also, the probability of student outmigration is positively related to students' academic coursework and negatively related to students' vocational coursework; (3) these relationships hold in the subsample of the youth who did not go to college in the period of the four years after high school completion, as well as in the total sample; and (4) overall, curriculum effects do not differ in rural schools vis-a-vis schools elsewhere, except that among the subsample of non-college goers, school academic emphasis seem to reduce rural school average migration in a marginal fashion. In brief, vocational education seems to work better than academic programs in retaining youth in the rural areas.

As discussed earlier, the findings from this study are intended to further our general understanding about the the functions of public school curriculum in connection to postsecondary student mobility in rural areas. They cannot be applied to assessing the value of specific programs in local schools. An implication is that vocational and technical education programs tailored to local economic needs are capable of serving rural communities in retaining educated youth and consequently contribute to social and economic development in rural areas. An emerging consensus seems to favor curriculum integration that encompasses cooperative learning, experiential education, outdoor education, in addition to both conventional academic and vocational education.

Future research may be directed to look at more specific elements of school curriculum contents in explaining student postsecondary mobility. Within large categories of programs, sub-groups of courses

that are more relevant to the needs of local development may be identified and examined. For example, rural areas with tourist resources may benifit from educational programs that concentrate in tourist business; whereas areas with correctional facilities may need criminal justice and other related training programs.

Further, innovative approaches to developing community-oriented curricula should be studied. Community-based curriculum appears promising. Teaching and learning about local communities, including the historical and cultural studies of the place, and practical research into local problems and solutions, is an effective way to orient students to the communty needs. Such programs help students understand their communities and acquire skills useful for improving local conditions. More data on such innovative programs are needed for research.

The rapid development of telecommunication technology, for example, is making it possible for professionals to work in places that are remote from urban centers. Rural areas may anticipate opportunities for growth thanks to such development. In responding to such changes, schools are providing increasingly more courses on new technology. This sort of programs enable students to enter labor market with competitive skills or to pursue postsecondary education with good exposure to updated technology. The changing curriculum is likely to contribute to retaining educated youth in rural areas. How and to what extent this actually happens is an issue needs to study.

Finally, it is also interesting to focus on those who had extensive exposure to academic education yet took vocational programs as the basis for post high school career choice. This group, compared to those who merely took vocational studies and missed a large amount of liberal arts courses, perhaps would attain marketable skills while managing to stay in the community after high school graduation. We need to conceptualize such issues and examine them with empirical data.

5. Statistical issues

In this study, with the implementation of multilevel logit models, related statistical issues in weighting and model assessment were addressed methodologically and empirically.

5.1 Weighting

Weighting for survey data in multilevel logit modeling was an issue (Skinner ,1996). Literature review showed there were few studies reported in multilevel logit modeling of binary responses, and none of them involving use of weighted models (McArdle and Hamagami, 1994; Rodriguez and Goldman, 1995).

Both MLn and HLM (Bryk, et al., 1996), another widely used software for multilevel modeling, can perform weighted multilevel linear modeling, though they have different design of weighting procedure. However, neither MLn nor HLM is ready to perform weighted multilevel logit modeling. In fact, HLM's newly available procedure for hierarchical generalized linear modeling (HGLM) doesn't contain a weighting device. And for MLn, our examination showed its weighting procedure is applicable also only to linear modeling. We thus decided to use unweighted models for this study.

In an atempt to perform weighted multilevel ligit modeling, we explored the way of adapting MLn's weighting device to implement weighted logit models. A promising strategy is to use weights as frequencies for modeling binary data in the event/trial model syntax (SAS, 1992).

5.2 Intraclass correlation

It was desirable to have a simple measure assessing the effect of clustering in the context of multilevel logit model as does for multilevel linear model. We derived the intraclass correlation associated with a random-intercept-only logit model, as an analogue of the intraclass correlation for multilevel linear model (Goldstein, 1995; Bryk and Raudenbush, 1992), using the technique of linearization (Rodriguez and Goldman, 1995). It takes the following form:

$$\frac{\sigma^2_{u0}}{\sigma^2_{u0} + [\mu_0(1-\mu_0)]^{-1}}$$

where σ^2_{u0} is the level-2 variance associated with the intercept, and $\mu_0 = \log it^{-1}(\gamma_{00})$, where γ_{00} is the fixed intercept of the logit model. We used this measure to our data. The intraclass correlation is calculated as 0.054 for a simple random-intercept-only model (not containing school variables). Thus the level-2 variance accounts for about 5 percent of the total variance. This assessment is consistent with the models we obtained.

References

Adamchak, D.J. (1987). Further evidence on economic nd noneconomic reasons for turnaround migration. Ru*ral Sociology,* 52, 1, 108-118.

Aitkin, M., Anderson, D., Francis, B., and Hinde, J. (1989). *Statistical Modeling in GLIM*, Oxford: Oxford University Press.

Berry, W. (1990). *What are People for?* San Francisco: North Point Press.

Berry, W. (1989). *The Hidden wound.* San Francisco: North Point Press.

Bryk, A.S. and Raudenbuch, S.W. (1992). *Hierarchical Linear Models: Applications and Data Analysis Methods.* London: Sage.

Bryk, A.S., Raudenbush, S.W., and Congdon, R.T. (1996). *HLM: Hierarchical Linear and Nonlinear Modeling with the HLM/2L and HLM/3L Programs.* Chicago, IL: Scientific Software International.

Cromartie, J.B. (1993). "Leaving the Countryside: Young Adults Follow Complex Migration Patterns," *Rural Development Perspectives,* 8, 2, 22-27.

DeYoung, A. J. (1987). "The Status of American Rural Education Research: An Integrated Review and Commentary," *Review of Educational Research,* 57(2), 123-148.

Gober, P. (1993). "Americans on the Move," *Population Bulletin,* 48,32-40.

Goldstein, H. (1995). *Multilevel Statistical Models*, 2nd edition. London: Edward Arnold.

Johnson, K.M. (1989). "Recent Population Redistribution Trends in Nonmetropolitan America," *Rural Sociology,* 54, 3, 301-326

Lee, V.E. and Bryk A. S. (1989). "A Multilevel Model of the Social Distribution of High School Achievement," *Sociology of Education*, 62, 2, 172-192.

McArdle, J.J. and Hamagami, F. (1994). "Logit and Multilevel Logit Modeling of College Graduation for 1984-1985 Freshman Student-Athletes," *Journal of the American Statistical Association*, 89, 1107-1123.

O'Hare, W. (1988). *The Rise of Poverty in Rural America.* Washington, DC: Population Reference Bureau. ERIC Document Reproduction No. ED 302 350.

Paterson, L. (1992). "The Influence of Opportunity on Aspirations amongst Prospective University Entrants from Scottish Schools, 1970-1988," *Journal of the Royal Statistical Society, A*, 155, 37-60.

Pollard, K, O'Hare, W. & ??? (1991). "Selective Migration of Rural High School Seniors in the 1980s," *Staff Working Papers.* Washington, DC: Population Reference Bureau, Inc.

Rasbash, J., Woodhouse, G., Goldstein, H., Yang, M., Howarth, J. and Plewis, I. (1995). *MLn Command Reference*, London: Institute of Education, University of London.

Reid, J. (1990). "Education and Rural Development: A Review of Recent Evidence," presented at the annual meeting of American Educational Research Association, Boston, MA.

Rodriguez, G., and Goldman, N. (1995). "An Assessment of Estimation Procedures for Multilevel Methods with Binary Responses," *Journal of Royal Statistical Society, A*, 158, Part 1, 73-89.

Sandefur, G. d. & Jeon, J. (1991). "Migration, Race and Ethnicity, 1960-1980," *International Migration Review,* 25, 2, 392-407.

SAS (1992). *SAS/STAT User's Guide, Version 6, Edition 4.* Cary, NC: SAS System Institute.

Skinner, C. (1996). *Regression Models for Complex Survey Data*, presented at the Joint Program in Survey Methodology, University of Maryland at College Park.

Snauwaert, D. T. (1990). "Wendell Berry, Liberalism, and Democratic Theory: Implications for the Rural School," *Peabody Journal of Education*, 67, 4, 118-130.

Theobald, P. (1992). "Rural Philosophy for Education: Wendell Berry's Tradition," *ERIC Digest.* Charleston, WV: ERIC Clearinghouse on Rural Education and Small Schools. ERIC Document Reproduction No.

Theobald, P. (1990). "A Look at Rural Education in the United States," *Peabody Journal of Education*, 67, 4, 1-6.

Thompson, P. B. & Kutach, D. N. (1990). "Agricultural Ethics in Rural Education," *Peabody Journal of Education*, 67, 4, 131-153.

Voss, P.R. & Futuitt, G.V. (1991). "The Impact of Migration on Southern Rural Areas of Chronic Depression," *Rural Sociology,* 56, 4, 660-679.

Woodhouse, G., Rasbash, J., Goldstein, H., Yang, M., Howart, J., and Plewis, I. (1995). *A Guide to MLn for New Users.* London: Institute of Education, University of London.

Listing of NCES Working Papers to Date

Please contact Ruth R. Harris at (202) 219-1831
if you are interested in any of the following papers

Number	Title	Contact
94-01 (July)	Schools and Staffing Survey (SASS) Papers Presented at Meetings of the American Statistical Association	Dan Kasprzyk
94-02 (July)	Generalized Variance Estimate for Schools and Staffing Survey (SASS)	Dan Kasprzyk
94-03 (July)	1991 Schools and Staffing Survey (SASS) Reinterview Response Variance Report	Dan Kasprzyk
94-04 (July)	The Accuracy of Teachers' Self-reports on their Postsecondary Education: Teacher Transcript Study, Schools and Staffing Survey	Dan Kasprzyk
94-05 (July)	Cost-of-Education Differentials Across the States	William Fowler
94-06 (July)	Six Papers on Teachers from the 1990-91 Schools and Staffing Survey and Other Related Surveys	Dan Kasprzyk
94-07 (Nov.)	Data Comparability and Public Policy: New Interest in Public Library Data Papers Presented at Meetings of the American Statistical Association	Carrol Kindel
95-01 (Jan.)	Schools and Staffing Survey: 1994 Papers Presented at the 1994 Meeting of the American Statistical Association	Dan Kasprzyk
95-02 (Jan.)	QED Estimates of the 1990-91 Schools and Staffing Survey: Deriving and Comparing QED School Estimates with CCD Estimates	Dan Kasprzyk
95-03 (Jan.)	Schools and Staffing Survey: 1990-91 SASS Cross-Questionnaire Analysis	Dan Kasprzyk
95-04 (Jan.)	National Education Longitudinal Study of 1988: Second Follow-up Questionnaire Content Areas and Research Issues	Jeffrey Owings
95-05 (Jan.)	National Education Longitudinal Study of 1988: Conducting Trend Analyses of NLS-72, HS&B, and NELS:88 Seniors	Jeffrey Owings

Listing of NCES Working Papers to Date--Continued

Number	Title	Contact
95-06 (Jan.)	National Education Longitudinal Study of 1988: Conducting Cross-Cohort Comparisons Using HS&B, NAEP, and NELS:88 Academic Transcript Data	Jeffrey Owings
95-07 (Jan.)	National Education Longitudinal Study of 1988: Conducting Trend Analyses HS&B and NELS:88 Sophomore Cohort Dropouts	Jeffrey Owings
95-08 (Feb.)	CCD Adjustment to the 1990-91 SASS: A Comparison of Estimates	Dan Kasprzyk
95-09 (Feb.)	The Results of the 1993 Teacher List Validation Study (TLVS)	Dan Kasprzyk
95-10 (Feb.)	The Results of the 1991-92 Teacher Follow-up Survey (TFS) Reinterview and Extensive Reconciliation	Dan Kasprzyk
95-11 (Mar.)	Measuring Instruction, Curriculum Content, and Instructional Resources: The Status of Recent Work	Sharon Bobbitt & John Ralph
95-12 (Mar.)	Rural Education Data User's Guide	Samuel Peng
95-13 (Mar.)	Assessing Students with Disabilities and Limited English Proficiency	James Houser
95-14 (Mar.)	Empirical Evaluation of Social, Psychological, & Educational Construct Variables Used in NCES Surveys	Samuel Peng
95-15 (Apr.)	Classroom Instructional Processes: A Review of Existing Measurement Approaches and Their Applicability for the Teacher Follow-up Survey	Sharon Bobbitt
95-16 (Apr.)	Intersurvey Consistency in NCES Private School Surveys	Steven Kaufman
95-17 (May)	Estimates of Expenditures for Private K-12 Schools	Stephen Broughman
95-18 (Nov.)	An Agenda for Research on Teachers and Schools: Revisiting NCES' Schools and Staffing Survey	Dan Kasprzyk
96-01 (Jan.)	Methodological Issues in the Study of Teachers' Careers: Critical Features of a Truly Longitudinal Study	Dan Kasprzyk

Listing of NCES Working Papers to Date--Continued

Number	Title	Contact
96-02 (Feb.)	Schools and Staffing Survey (SASS): 1995 Selected papers presented at the 1995 Meeting of the American Statistical Association	Dan Kasprzyk
96-03 (Feb.)	National Education Longitudinal Study of 1988 (NELS:88) Research Framework and Issues	Jeffrey Owings
96-04 (Feb.)	Census Mapping Project/School District Data Book	Tai Phan
96-05 (Feb.)	Cognitive Research on the Teacher Listing Form for the Schools and Staffing Survey	Dan Kasprzyk
96-06 (Mar.)	The Schools and Staffing Survey (SASS) for 1998-99: Design Recommendations to Inform Broad Education Policy	Dan Kasprzyk
96-07 (Mar.)	Should SASS Measure Instructional Processes and Teacher Effectiveness?	Dan Kasprzyk
96-08 (Apr.)	How Accurate are Teacher Judgments of Students' Academic Performance?	Jerry West
96-09 (Apr.)	Making Data Relevant for Policy Discussions: Redesigning the School Administrator Questionnaire for the 1998-99 SASS	Dan Kasprzyk
96-10 (Apr.)	1998-99 Schools and Staffing Survey: Issues Related to Survey Depth	Dan Kasprzyk
96-11 (June)	Towards an Organizational Database on America's Schools: A Proposal for the Future of SASS, with comments on School Reform, Governance, and Finance	Dan Kasprzyk
96-12 (June)	Predictors of Retention, Transfer, and Attrition of Special and General Education Teachers: Data from the 1989 Teacher Followup Survey	Dan Kasprzyk
96-13 (June)	Estimation of Response Bias in the NHES:95 Adult Education Survey	Steven Kaufman
96-14 (June)	The 1995 National Household Education Survey: Reinterview Results for the Adult Education Component	Steven Kaufman

Listing of NCES Working Papers to Date--Continued

Number	Title	Contact
96-15 (June)	Nested Structures: District-Level Data in the Schools and Staffing Survey	Dan Kasprzyk
96-16 (June)	Strategies for Collecting Finance Data from Private Schools	Stephen Broughman
96-17 (July)	National Postsecondary Student Aid Study: 1996 Field Test Methodology Report	Andrew G. Malizio
96-18 (Aug.)	Assessment of Social Competence, Adaptive Behaviors, and Approaches to Learning with Young Children	Jerry West
96-19 (Oct.)	Assessment and Analysis of School-Level Expenditures	William Fowler
96-20 (Oct.)	1991 National Household Education Survey (NHES:91) Questionnaires: Screener, Early Childhood Education, and Adult Education	Kathryn Chandler
96-21 (Oct.)	1993 National Household Education Survey (NHES:93) Questionnaires: Screener, School Readiness, and School Safety and Discipline	Kathryn Chandler
96-22 (Oct.)	1995 National Household Education Survey (NHES:95) Questionnaires: Screener, Early Childhood Program Participation, and Adult Education	Kathryn Chandler
96-23 (Oct.)	Linking Student Data to SASS: Why, When, How	Dan Kasprzyk
96-24 (Oct.)	National Assessments of Teacher Quality	Dan Kasprzyk
96-25 (Oct.)	Measures of Inservice Professional Development: Suggested Items for the 1998-1999 Schools and Staffing Survey	Dan Kasprzyk
96-26 (Nov.)	Improving the Coverage of Private Elementary-Secondary Schools	Steven Kaufman
96-27 (Nov.)	Intersurvey Consistency in NCES Private School Surveys for 1993-94	Steven Kaufman

Listing of NCES Working Papers to Date--Continued

Number	Title	Contact
96-28 (Nov.)	Student Learning, Teaching Quality, and Professional Development: Theoretical Linkages, Current Measurement, and Recommendations for Future Data Collection	Mary Rollefson
96-29 (Nov.)	Undercoverage Bias in Estimates of Characteristics of Adults and 0- to 2-Year-Olds in the 1995 National Household Education Survey (NHES:95)	Kathryn Chandler
96-30 (Dec.)	Comparison of Estimates from the 1995 National Household Education Survey (NHES:95)	Kathryn Chandler
97-01 (Feb.)	Selected Papers on Education Surveys: Papers Presented at the 1996 Meeting of the American Statistical Association	Dan Kasprzyk