NATIONAL CENTER FOR EDUCATION STATISTICS

Working Paper Series

The Working Paper Series was initiated to promote the sharing of the valuable work experience and knowledge reflected in these preliminary reports. These reports are viewed as works in progress, and have not undergone a rigorous review for consistency with NCES Statistical Standards prior to inclusion in the Working Paper 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 1998 and 1999 ASA and 1999 AAPOR Meetings

Working Paper No. 2000-04 March 2000

Contact: Dan Kasprzyk
Elementary/Secondary and Libraries Studies Division
E-mail: sassdata@ed.gov
Web page: http://nces.ed.gov/surveys/sass.html

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
C. Kent McGuire
Assistant Secretary

National Center for Education Statistics
Gary W. Phillips
Acting 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
1990 K Street, NW
Washington, DC 20006-5651

The NCES World Wide Web Home Page is
http://nces.ed.gov

Suggested Citation

U.S. Department of Education. National Center for Education Statistics. *Selected Papers on Education Surveys: Papers Presented at the 1998 and 1999 ASA and 1999 AAPOR Meetings*, Working Paper No. 2000-04. Project Officer, Dan Kasprzyk. Washington, DC: 2000.

March 2000

Foreword

In addition to official NCES publications, NCES staff and individuals commissioned by NCES produce preliminary research reports that include analyses of survey results, and presentations of technical, methodological, and statistical evaluation issues.

The ***Working Paper Series*** was initiated to promote the sharing of the valuable work experience and knowledge reflected in these preliminary reports. These reports are viewed as works in progress, and have not undergone a rigorous review for consistency with NCES Statistical Standards prior to inclusion in the Working Paper Series.

Copies of Working Papers can be downloaded as pdf files from the NCES Electronic Catalog (http://nces.ed.gov/pubsearch/), or contact Sheilah Jupiter at (202) 502-7444, e-mail: sheilah_jupiter@ed.gov, or mail: U.S. Department of Education, Office of Educational Research and Improvement, National Center for Education Statistics, 1990 K Street, NW, Room 9048, Washington, DC 20006.

Marilyn M. McMillen
Chief Mathematical Statistician
Statistical Standards Program

Ralph Lee
Mathematical Statistician
Statistical Standards Program

This page intentionally left blank.

Table of Contents

Preface

The 14 papers contained in this volume were presented at either the 1998 or 1999 American Statistical Association (ASA) meeting, or the 1999 American Association for Public Opinion Research (AAPOR) Conference. This is the sixth collection of conference papers of particular interest to users of NCES survey data published in the Working Paper Series. The earlier collections are listed below:

- Working Paper 94-01, which included papers presented at the 1992 and 1993 ASA meetings and the 1993 ASA Conference on Establishment Surveys;
- Working Paper 95-01, which included papers from the 1994 ASA meeting;
- Working Paper 96-02, which included papers from the 1995 ASA meeting;
- Working Paper 97-01, which included papers from the 1996 ASA meeting; and
- Working Paper 97-41, which included papers from the 1997 ASA meeting.

This page intentionally left blank.

EVALUATION OF SOME POPULAR IMPUTATION ALGORITHMS

Mingxiu Hu, Sameena M. Salvucci, Synectics for Management Decisions, Michael P. Cohen, NCES
Sameena M. Salvucci, Synectics, 1901 North Moore St., Suite 900, Arlington, VA 22209

Key Words: Imputation Methods, Missing Data, Monte Carlo Simulation

I. INTRODUCTION. *Imputation* has become one of the most popular tools used to solve missing value problems in survey data analyses. A popular misunderstanding is that the goal of imputation is to predict individual missing values. This is popular because of hot deck imputation methods which attempt to find the best match (donor) for each missing case. A better estimate for each missing value not necessarily leads to a better overall estimate for the parameters of interest. As Rubin (1996) pointed out, imputation has two achievable objectives. The basic objective is to allow ultimate data users to apply their existing analysis tools to any data set with missing values using the same command structure and output standards as if there were no missing data. Most imputation methods satisfy this basic objective and so have a certain appeal. But it is certainly not enough to just achieve this basic goal. The additional desirable objective is to obtain statistically valid inference. This goal can be achieved through some imputation methods, but not through others.

Many imputation techniques and imputation software packages have been developed over the years. Section II gives a brief review on some thirty imputation methods. Different methods may work well under different circumstances. The major part of this paper evaluated eleven popular imputation methods according to six evaluation criteria for four types of distributions, five types of missing mechanisms, and four types of missing rates, through a simulation study.

II. IMPUTATION METHODS. Imputation methods are conventionally classified into two categories: *random* (or *stochastic*) imputation and *deterministic* imputation. A deterministic imputation method determines one and only one possible value for imputing each missing case. On the other hand, a random imputation method draws imputation values randomly either from observed data or from a predicted distribution. In this paper, we divide imputation methods into five categories: *simple deterministic*, *simple random*, *model-based deterministic*, *model-based random*, and *Bayesian-theory-based* imputation methods. This is not a mutually exclusive partition, but it provides a clearer picture about the property of each imputation method.

1. Simple deterministic imputation. This type of method generally distorts the distribution of the data and leads to underestimation of the variance (except the deductive imputation method). However, it is still widely used in practice because of its simplicity. The most popular ones follow.

Deductive imputation. Missing values may be deduced from available information such as similar items in previous surveys, related items of current surveys, etc. The cold deck may be counted as this type.

Mean imputation. This is the simplest but least attractive method. The concentration of all imputed values at the mean creates spikes in the distribution, and the variances will be materially underestimated.

Deterministic hot-deck imputation. This method is used very often in early imputation practice because it intuitively makes sense to many practitioners. It does not employ any explicit statistical model. This method has many versions. Among the most popular ones are: *(1) Sequential nearest neighbor hot deck imputation* (or *traditional hot deck imputation*). A major attraction of this method is that all imputations are made from a single pass through the data file. A disadvantage of this method is that it may easily give rise to multiple use of donors, a feature which leads to a loss of precision for survey estimators; (2) *Multivariate matching*. This method is not convenient to implement using computer programs. An approximately equivalent algorithm may be used to replace it: First sort the data file with the same auxiliary variables, and then impute the nearest response value for each missing case; (3) *Distance function matching*. This method imputes the nearest response value according to some univariate distance function of auxiliary variables, such as the Mahalanobis distance, the difference between the predicted values from a regression model, etc.

2. Simple random imputation. This type of method adds some uncertainty about imputed values and much less likely to distort the distribution of the data comparing to simple deterministic methods. It may still underestimate the variance if no within-imputation variation is considered. The following methods belong to this category.

Mean imputation with random disturbance. A random disturbance is added to the mean imputation.

Overall random imputation. The overall random imputation generally refers to draw imputation randomly from observed data using different sampling schemes. This is one of the easiest method to implement. But it does not use any auxiliary variables and will not be able to reduce non-response biases.

Within-class random imputation. This widely used method involves two steps: to form imputation classes and to draw imputations within each class. Imputation classes may be formed using: (i) regression predicted

values from a multivariate regression model. This method was used by imputation software PROC IMPUTE; (ii) a propensity score. Rosenbaum and Rubin (1983) show that the best score function for constructing imputation classes is the propensity score, defined as the conditional probability of observing the target variables Y given covariates X. With a propensity score e(X), the property that the missing mechanism is independent of Y given X, carries over to independence given the propensity score e(X). We may use a logistic regression model to estimate the propensity scores. The r*andom hot-deck method,* which randomly draws imputations from observed data according to the weighted or unweighted frequency, is a specific within-class random imputation method.

3. Model-based deterministic imputation. Model-based approaches will produce more accurate imputations than randomization-based methods if the model assumptions hold. But those assumptions are usually unverifiable in practice. A good model-based approach is required to work well for a wide range of underlying data distributions and missing mechanisms. Again, the deterministic nature of this type of method will lead to distortion of the distribution and under-estimation of the variance.

Ratio imputation. This widely applied method may be able to provide very accurate imputations if the missingness of the target variable mainly depends on only one highly correlated auxiliary variable. If missing values depend on several auxiliary variables, the ratio imputation may not be fully efficient.

Predicted regression imputation. This method is also widely used in early imputation practice. It uses predicted values from a regression model as imputations for all missing cases. The disadvantage of this method is "the shrinkage to the mean" phenomenon.

EM algorithm. Although the EM algorithm can be used to create imputation for individual missing values, it is more often used to obtain parameter estimates. Convergence may be slow and not guaranteed with the EM algorithm especially with sparse data. This method also suffers "the shrinkage to the mean" phenomenon. An advantage of EM algorithm is its stable convergence; that is, iterations always increase the likelihood.

Dear's principal component method, *General iterative principal component method,* and *Singular value decomposition (SVD) method* also belong to this category and enjoy similar properties as EM algorithm. See Bello (1993) for details on these methods.

4. Model-based random imputation. This type of method shares disadvantages and advantages as model-based approaches stated in the preceding section, but it enjoys the advantages of random imputation over deterministic imputation. The following methods belong to this category.

Draw imputations from predicted distributions. If there is some information available about the type of the distribution, we may draw imputations from a predicted distribution. With this method, we assume a distribution for the data and use the observed data to estimate the unknown parameters in the assumed distribution. If the distribution assumption is approximately true, this method will give much better imputations than any method which draws imputations from observed data.

Random regression imputation. A small random disturbance may be added to the regression imputation. The disturbance may be drawn from: (1) a distribution with mean 0 and variance estimated from observed data; (2) respondents' residuals of the regression model; (3) residuals of those respondents which have similar values on matching variables to protect against non-linearity and non-additivity in regression models.

Ratio with random disturbance imputation. Similar to the random regression imputation, we could add a small random disturbance to the ratio imputation.

Modeling non-ignorable missing mechanism. Most imputation methods only model the target variable with missing values but not the missing indicator variable. These methods explicitly or implicitly assume that the missing values occur at random given the conditioning variables. Greenless, Reece and Zieschagn (1982) used two models: a logistic regression model for the missing indicator variable and an ordinary regression model for the target variable. The method is more sensitive to the model specification. It is rarely used in practice because of the unverifiability of missing mechanisms and the complexity of the model specifications.

5. Bayesian-theory-based imputation. This type of method not only adds variation to the imputed data but also to the parameters of the model by drawing parameter estimates from their posterior distribution. The following methods belong to this category.

Approximate Bayesian Bootstrap (ABB). The ABB method first draws a resample from the observed data and then draws imputation from the resample. The extra step of taking a resample first introduces additional variation to the imputation, which makes the ABB method approximately "proper" for multiple imputation according to Rubin's theory (1987).

Bayesian Bootstrap (BB). The ABB and the BB are approximately equivalent. The only difference between them is that the underlying parameter of the data, which gives the probabilities of each possible value in the observed data, is being drawn from a scaled multi-nomial with the ABB rather than a Dirichlet distribution with the BB. These distributions have the same means and correlations, but the variances for ABB are $(1+1/r)$ times the variances for the BB, where r is the number of observed data.

Data augmentation. This Bayesian iterative method is proposed by Tanner and Wong (1987). Their method of constructing the complete data sets is closely related to the Gibbs sampler. It efficiently uses relationship among variables for constructing imputations. It generally gives both good point estimates and variance estimates if the distribution assumptions on the data are approximately satisfied. The disadvantage of the data augmentation method is that it requires iterations, and similar to the EM algorithm, convergence can be slow.

Adjusted data augmentation. If the distribution assumption in the data augmentation method is in question, it is desirable to let the observed data Y_{obs} influence the shape of the distribution of values imputed for Y_{mis}. Rubin and Scheker (1986) adjusted the normal model as follows. First, the parameters μ^* and σ^{*2} are drawn from their posterior distributions as in the data augmentation method. Second, $X_1,\ldots,X_m$ are drawn with replacement from Y_{obs}, and standardized through $Z_i = (X_i - \bar{y}_r)/\sqrt{(r-1)s_r^2/r}$. Finally, the m missing values are imputed using $\mu^* + \sigma^* Z_i$, i=1, 2, …, m.

Sequential imputation method. Kong, Liu and Wong (1994) proposed this method. According to the authors, the sequential imputation has three advantages over the data augmentation: (1) it does not require iterations; (2) it can directly estimate the model likelihood; (3) it can cheaply perform sensitivity analysis and influence analysis. But, so far, this method only has its theoretical value.

III. SIMULATION STUDY. We compared 11 popular imputation methods according to 6 evaluation criteria for 4 types of distributions, five types of missing mechanisms, and four types of missing rates. Detailed description of the design factors follow.

Distribution Four sets of variables were generated:
(1) **Normal**: Norm1, Norm2, Norm3, Norm4, Norm5;
(2) **Double Exponential**: Dexp1, Dexp2, Dexp3, Dexp4, and Dexp5;
(3) **Contaminated Normal**: MixNorm1, Mix-Norm2, MixNorm3, MixNorm4, and MixNorm5 from a mixed normal distribution of 95% N(μ, 1) and 5% N(μ, 3^2);
(4) **Mixer of Normal and Chi-square**: MixNChi1, MixNChi2, MixNChi3, MixNChi4, and MixNChi5 from mixed normal distributions of 95% N(μ, 1) and 5% $\chi^2(4) - 4 + \mu$.

The first three sets of variables are symmetric about their means, while the fourth set of variables is right skewed. Each set of five variables are correlated with the following correlation matrix:

$$\begin{pmatrix} 1 & 0.9 & 0.7 & 0.5 & 0.3 \\ 0.9 & 1 & 0.8 & 0.6 & 0.4 \\ 0.7 & 0.8 & 1 & 0.7 & 0.5 \\ 0.5 & 0.6 & 0.7 & 1 & 0.6 \\ 0.3 & 0.4 & 0.5 & 0.6 & 1 \end{pmatrix}$$

Missing Mechanism. The five types of missing mechanisms are: (1) *MCAR*; (2) *Tail values more likely missing*: missing values were created with probability of $exp(-\lambda |X-\mu|)$, where λ was determined so that the desired missing rates were created; (3) *Large values more likely missing*: missing values were created with probability of $exp[-\lambda (X-\mu)]$; (4) *Center values more likely missing:* missing values were created with probability of $1-exp[-\lambda |X-\mu|]$; (5) *Tail values more likely missing*: missing values in Y were created with probability of $1-exp[-\lambda |Y-\mu|]$. Only mechanism (5) is confounded; that is, missingness of Y depends on itself.

Missing Rate. For missing mechanisms (1), (2), (4), and (5), the four missing rates are 10%, 20%, 30%, and 40%, while for missing mechanism (3), the four missing rates are 5%, 10%, 15%, and 20%.

For each setting formed by the above simulation design factors, 200 datasets were generated and the imputation methods were assessed based on their average performance over the 200 replications. The sample size for each replicate data set is 100.

Imputation Methods. The 11 imputation methods included in this simulation study are:
(1) *Mean Imputation* (deterministic);
(2) *Ratio Imputation* (deterministic): Norm1, Norm2, Norm3, and Norm4 served as auxiliary variables for Norm2, Norm3, Norm4, and Norm5, respectively. No imputations were created for Norm1. Other types of variables are handled similarly;
(3) *Sequential nearest neighbor hot deck method* (deterministic);
(4) *Overall random imputation* (simple random);
(5) *Mean imputation with disturbance* (random);
(6) *Ratio imputation with disturbance* (random);
(7) *The ABB method* (random);
(8) *The BB method* (random);
(9) *PROC IMPUTE* (random);
(10) *Data Augmentation* (random): *Schafer's software* was used to implement this method in our simulation;
(11) *Adjusted data augmentation method* (random).

IV. SIMULATION RESULTS. The evaluation criteria are: bias of parameter estimates (mean, median, first and third quartiles), bias of variance estimates, coverage probability, confidence interval width, and average imputation error. Biases of quartile estimates, and average imputation error are not given in this paper because of space limitation. Results based on the other

criteria are given in Tables 1~6 and a brief summary is described below.

1. Bias of parameter estimates. Table 1 shows that ratio imputation with or without disturbance, Schafer's software, PROC IMPUTE, and hot deck are all very effective in improving the biases of mean estimates caused by missing mechanism (3) where large values were more likely to be missing. For all other missing mechanisms, biases are very small with the incomplete data. The ratio imputation method does so well because we used the same auxiliary variables to create and to impute the missing values in this method, and because the correlation coefficients between the target variables and the auxiliary variables are high (at least 0.6).

In terms of bias of quartile estimates (Table 2), the mean imputation method is obviously the worst across all five missing mechanisms. For missing mechanism (2) and (3), Schafer's software, ratio with and without disturbance imputation, PROC IMPUTE, and hot deck, have evident advantages over the other methods. For missing mechanism (4), the hot deck method has the best overall performance, followed by PROC IMPUTE and Schafer's software. For the confounded missing mechanism (5), the ratio with disturbance imputation method obviously has the best performance.

2. Bias of variance estimates. Table 3 reports the relative biases of variance estimates based on the data imputed via single imputation. For the MCAR, all methods provide acceptable variance estimates except the mean imputation whose estimates need to be adjusted with a factor of *(n-1)/(r-1)*. For unconfounded missing mechanisms, Schafer's software has the best performance, and ratio imputation, PROC IMPUTE, and the hot deck method are all able to improve the biases of variance estimates dramatically, but the ratio with disturbance imputation tends to overestimate the variance. For the confounded missing mechanism, only the ratio imputation with or without disturbance have substantial improvement on the biases of variance estimates. The random, ABB, BB, and mean imputation with disturbance have almost no improvement over the variance estimates based on the incomplete data, while the adjusted data augmentation method always helps a little, but never much.

Table 4 presents the relative biases of variance estimates of the mean based on five sets of imputations. The ratio with disturbance imputation method always overestimate the variances for all types of missing mechanisms. For this method, the idea of multiple imputation is obviously inappropriate. PROC IMPUTE seems to have the least between-imputation variation and provides approximately unbiased variance estimates for the MCAR and all unconfounded missing mechanisms. The ABB and BB methods introduce the most between-imputation variation for the MCAR and missing mechanism (4) when the incomplete data are more diversified than the true distribution.

3. Coverage rates. Schafer's software has almost perfect coverage rates across all five missing mechanisms. The adjusted data augmentation method also has almost perfect coverage rates for all except mechanism (3). These seem to suggest that imputation methods based on Bayesian theory give better coverage rates. Ratio and ratio with disturbance imputation methods have great coverage rates for missing mechanisms (2), (3), and (5). PROC IMPUTE has very good coverage rates except for missing mechanism (5). The sequential hot deck method is significantly worse than PROC IMPUTE in terms of coverage rates, but it is better than the other methods which do not use any auxiliary information, especially for missing mechanism (3). Not much difference has been found among the other methods. Some rates of these methods are too low, especially for missing mechanisms (3) and (5).

4. Confidence interval width. From Table 6, overall, Schafer's software and the adjusted data augmentation method have the shortest confidence intervals across the five missing mechanisms. We also found in the preceding section that the two methods also gave the best coverage rates except for missing mechanism (3) with the adjusted data augmentation method. Therefore, the two methods are least likely to provide bad estimates. The other methods seem not to have substantial advantage over each other in terms of confidence interval width.

V. REFERENCES.

Bello, A. L. (1993). Choosing among imputation techniques for incomplete multivariate data: a simulation study. *Comm. in Stat.*, 22(3), 853-877.

Greenless, J. S., Reece, W. S. and Zieschang, K. D. (1982). Imputation of missing values when the probability of response depends on the variable being imputed. *JASA*, 77, 251-261.

Kong, A., Liu, J. S. and Wong, W. H. (1994). Sequential imputations and Bayesian missing data problems. *JASA*, Vol. 89, No. 425, 278-288.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70, 41-55.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley and Sons.

Rubin, D. B. (1996). Multiple imputation after 18+ years. *JASA*, Vol. 91, No. 343, 473-489.

Rubin, D. B. and Schenker, N. (1986). Multiple imputation for interval estimation from simple random sam-ples with ignorable nonresponse. *JASA*, 81, 366-374.

Tanner, M. A. and Wong, W. H. (1987). The calculation of posterior distributions by data augmentation (with discussion). *JASA*, 82, 528-550.

Table 1 Bias of population mean estimates (overall #)

Missing Mechanism	Distribution	Mean Imp.	Ratio Imp.	Hot Deck	Random	Mean +error	Ratio +error	ABB	BB	Proc Impute	Schafer	Adj. DA
1. MCAR	Normal	-0.005		0.012	-0.007	-0.009		-0.003	-0.008	-0.003	-0.006	-0.004
	Dexp	-0.004		0.014	0.001	0.000		-0.009	-0.015	-0.003	-0.004	0.003
	MixNorm	0.003		0.025	0.003	0.009		0.009	0.002	-0.004	-0.005	0.001
	MixNChi	0.009		0.079	0.011	0.011		0.011	0.033	0.014	0.008	0.022
2. Unconfounded	Normal	0.005	-0.002	0.000	0.008	0.006	-0.007	0.006	0.007	-0.002	-0.001	0.006
(tail values	Dexp	-0.003	-0.011	-0.008	-0.007	-0.004	-0.009	0.004	-0.007	0.001	-0.003	-0.007
more likely	MixNorm	0.003	-0.009	0.001	0.000	0.001	-0.003	0.004	0.006	0.002	-0.004	0.001
missing)	MixNChi	-0.014	-0.034	0.016	-0.011	-0.012	-0.033	-0.011	-0.011	-0.023	0.000	-0.010
3. Unconfounded	Normal	-0.094	0.002	-0.021	-0.095	-0.094	0.004	-0.093	-0.094	0.010	0.001	-0.085
(large values	Dexp	-0.118	0.003	-0.034	-0.116	-0.119	0.002	-0.119	-0.112	0.020	0.003	-0.103
more likely	MixNorm	-0.109	0.001	-0.024	-0.109	-0.110	0.004	-0.112	-0.104	0.011	0.001	-0.098
missing)	MixNChi	-0.159	-0.001	-0.061	-0.160	-0.157	-0.001	-0.151	-0.154	-0.045	-0.007	-0.143
4. Unconfounded	Normal	0.013	0.032	0.009	0.016	0.010	0.032	0.012	0.012	-0.002	0.004	0.013
(Center values	Dexp	-0.006	0.022	-0.014	-0.007	0.000	0.027	0.000	-0.007	-0.016	-0.005	-0.010
more likely	MixNorm	0.010	0.031	0.008	0.007	0.010	0.030	0.018	0.016	-0.004	-0.002	0.007
missing)	MixNChi	0.016	0.048	0.022	0.025	0.024	0.054	0.020	0.018	-0.012	-0.004	0.022
5. Confounded	Normal	-0.003	0.002	0.001	0.000	-0.005	0.004	-0.008	-0.001	-0.008	-0.006	-0.004
(tail values	Dexp	0.005	0.016	0.012	0.010	0.004	0.012	0.006	0.003	0.006	0.006	0.006
more likely	MixNorm	-0.010	-0.005	-0.004	-0.009	-0.013	-0.007	-0.011	-0.006	-0.014	-0.019	-0.006
missing)	MixNChi	-0.076	-0.022	-0.045	-0.071	-0.070	-0.015	-0.072	-0.078	-0.065	-0.032	-0.062

Table 2 Biases of the first quartile estimates (overall #)

Missing Mechanism	Distribution	Mean Imp.	Ratio Imp.	Hot Deck	Random	Mean +error	Ratio +error	ABB	BB	Proc Impute	Schafer	Adj. DA
1. MCAR	Normal	0.251		0.038	-0.001	-0.006		0.007	-0.004	-0.016	-0.013	0.001
	Dexp	0.289		0.028	-0.004	-0.062		-0.004	-0.010	0.004	-0.045	-0.007
	MixNorm	0.271		0.033	-0.003	-0.012		0.007	0.004	0.004	-0.015	-0.011
	MixNChi	0.290		0.044	-0.003	-0.084		0.002	0.008	0.049	-0.058	-0.027
2. Unconfounded	Normal	0.221	-0.027	-0.014	0.066	0.066	-0.019	0.068	0.056	-0.034	-0.003	0.054
(tail values	Dexp	0.272	-0.017	0.003	0.094	0.074	-0.015	0.092	0.092	-0.001	-0.002	0.076
more likely	MixNorm	0.247	-0.015	-0.003	0.071	0.061	-0.004	0.072	0.072	-0.004	-0.001	0.059
missing)	MixNChi	0.245	-0.018	0.022	0.082	0.033	-0.016	0.076	0.086	0.021	0.003	0.047
3. Unconfounded	Normal	0.005	0.005	-0.008	-0.066	-0.073	-0.021	-0.068	-0.074	0.001	0.000	-0.060
(large values	Dexp	0.015	0.015	-0.013	-0.080	-0.097	-0.018	-0.084	-0.077	0.006	0.003	-0.073
more likely	MixNorm	0.008	0.008	-0.009	-0.088	-0.085	-0.022	-0.083	-0.079	0.002	0.001	-0.082
missing)	MixNChi	0.009	0.009	-0.020	-0.087	-0.123	-0.051	-0.085	-0.086	-0.011	-0.022	-0.084
4. Unconfounded	Normal	0.209	0.123	0.008	-0.039	-0.038	-0.031	-0.044	-0.046	0.036	0.001	-0.033
(Center values	Dexp	0.173	0.118	-0.024	-0.091	-0.099	-0.082	-0.083	-0.092	0.017	-0.032	-0.085
more likely	MixNorm	0.193	0.111	0.006	-0.064	-0.065	-0.062	-0.061	-0.056	0.023	-0.024	-0.056
missing)	MixNChi	0.238	0.138	-0.014	-0.118	-0.207	-0.197	-0.121	-0.121	0.049	-0.137	-0.112
5. Confounded	Normal	0.331	0.096	0.120	0.131	0.116	0.045	0.115	0.123	0.142	0.111	0.121
(tail values	Dexp	0.463	0.143	0.173	0.201	0.177	0.061	0.190	0.189	0.203	0.153	0.191
more likely	MixNorm	0.388	0.096	0.137	0.140	0.127	0.033	0.146	0.150	0.157	0.103	0.135
missing)	MixNChi	0.467	0.124	0.189	0.206	0.162	0.021	0.192	0.197	0.135	0.143	0.172

Table 3 Relative bias of variance estimates with single imputation (overall #)

Missing Mechanism	Distribution	Mean Imp.	Ratio Imp.	Hot Deck	Random	Mean +error	Ratio +error	ABB	BB	Proc Impute	Schafer	Adj. DA
1. MCAR	Normal	-0.250		-0.039	-0.019	-0.010		-0.008	-0.009	-0.027	0.012	-0.010
	Dexp	-0.234		-0.020	0.019	0.024		0.006	0.010	0.001	0.014	0.024
	MixNorm	-0.247		-0.039	-0.004	-0.004		-0.006	-0.028	-0.027	0.004	0.006
	MixNChi	-0.242		0.195	-0.011	0.007		-0.008	0.064	-0.044	0.026	0.018
2. Unconfounded	Normal	-0.279	0.033	-0.001	-0.123	-0.132	0.172	-0.130	-0.121	0.080	0.004	-0.097
(tail values	Dexp	-0.372	0.057	-0.065	-0.244	-0.237	0.174	-0.244	-0.240	-0.012	-0.009	-0.199
more likely	MixNorm	-0.341	0.064	-0.025	-0.205	-0.193	0.206	-0.205	-0.196	-0.006	-0.002	-0.162
missing)	MixNChi	-0.519	0.008	-0.204	-0.421	-0.429	0.097	-0.415	-0.426	-0.110	-0.005	-0.357
3. Unconfounded	Normal	-0.137	-0.018	-0.029	-0.050	-0.048	0.080	-0.046	-0.046	0.029	0.004	-0.041
(large values	Dexp	-0.131	-0.022	-0.024	-0.040	-0.040	0.058	-0.041	-0.045	0.042	0.003	-0.032
more likely	MixNorm	-0.138	-0.020	-0.024	-0.051	-0.051	0.068	-0.049	-0.041	0.041	0.004	-0.044
missing)	MixNChi	-0.190	-0.023	-0.052	-0.117	-0.107	0.057	-0.108	-0.098	-0.072	-0.009	-0.108
4. Unconfounded	Normal	-0.136	-0.082	0.014	0.114	0.118	0.171	0.119	0.119	-0.036	0.004	0.092
(Center values	Dexp	-0.113	-0.084	0.017	0.109	0.110	0.133	0.110	0.111	-0.041	-0.006	0.088
more likely	MixNorm	-0.123	-0.083	-0.002	0.121	0.115	0.162	0.122	0.123	-0.036	-0.002	0.095
missing)	MixNChi	-0.144	-0.126	0.011	0.165	0.137	0.148	0.186	0.123	-0.099	-0.021	0.117
5. Confounded	Normal	-0.444	-0.146	-0.255	-0.282	-0.278	0.106	-0.269	-0.278	-0.309	-0.247	-0.267
(tail values	Dexp	-0.510	-0.162	-0.321	-0.358	-0.360	0.055	-0.354	-0.353	-0.373	-0.317	-0.344
more likely	MixNorm	-0.514	-0.178	-0.330	-0.353	-0.351	0.054	-0.361	-0.353	-0.375	-0.323	-0.338
missing)	MixNChi	-0.750	-0.228	-0.629	-0.678	-0.676	-0.075	-0.676	-0.680	-0.488	-0.550	-0.644

The "Overall" combined missing rate is about 10% for missing mechanism 4 and 25% for the others. Results for each missing rate category are available from the authors.

Table 4 Relative bias of variance estimates with five sets of imputations (overall #)

Missing Mechanism	Distribution	Random	Mean +error	Ratio +error	ABB	BB	Proc Impute	Schafer	Adj DA
1. MCAR	Normal	0.254	0.272		0.459	0.365	0.018	0.065	0.280
	Dexp	0.327	0.323		0.458	0.449	0.021	0.087	0.327
	MixNorm	0.283	0.303		0.400	0.348	-0.003	0.059	0.289
	MixNChi	0.304	0.320		0.393	0.557	-0.010	0.069	0.324
2. Unconfounded	Normal	0.060	0.046	0.364	0.102	0.065	0.094	0.030	0.122
(tail values	Dexp	-0.088	-0.086	0.343	-0.014	-0.059	0.000	0.016	0.010
more likely	MixNorm	-0.026	-0.017	0.359	0.024	-0.021	0.010	0.033	0.062
missing)	MixNChi	-0.291	-0.307	0.205	-0.290	-0.296	-0.082	0.022	-0.147
3. Unconfounded	Normal	0.069	0.064	0.164	0.083	0.047	0.038	0.035	0.086
(large values	Dexp	0.065	0.059	0.160	0.084	0.049	0.059	0.036	0.079
more likely	MixNorm	0.062	0.059	0.177	0.067	0.057	0.053	0.040	0.079
missing)	MixNChi	0.000	-0.016	0.173	0.004	-0.003	-0.050	0.022	0.018
4. Unconfounded	Normal	0.409	0.415	0.484	0.558	0.494	0.011	0.130	0.358
(Center values	Dexp	0.350	0.354	0.379	0.452	0.410	-0.006	0.113	0.306
more likely	MixNorm	0.433	0.396	0.438	0.475	0.463	0.012	0.120	0.373
missing)	MixNChi	0.569	0.477	0.482	0.752	0.571	-0.079	0.096	0.446
5. Confounded	Normal	-0.055	-0.064	0.342	0.046	-0.009	-0.248	-0.093	-0.029
(tail values	Dexp	-0.170	-0.172	0.326	-0.102	-0.093	-0.322	-0.187	-0.148
more likely	MixNorm	-0.156	-0.171	0.314	-0.021	-0.127	-0.328	-0.181	-0.126
missing)	MixNChi	-0.586	-0.584	0.105	-0.548	-0.561	-0.450	-0.491	-0.504

Table 5 Coverage rates with single imputation (overall #)

Missing Mechanism	Distribution	Mean Imp.	Ratio Imp.	Hot Deck	Random	Mean + error	Ratio +error	ABB	BB	Proc Impute	Schafer	Adj. DA
1. MCAR	Normal	84.5%		93.5%	87.5%	86.5%		85.5%	85.5%	92.0%	96.0%	93.5%
	Dexp	85.0%		87.5%	88.5%	88.0%		84.5%	86.0%	93.0%	94.5%	94.5%
	MixNorm	85.0%		91.5%	89.5%	85.0%		84.0%	87.0%	93.0%	95.0%	95.5%
	MixNChi	84.0%		87.0%	88.5%	87.5%		86.5%	86.0%	92.5%	94.5%	95.5%
2. Unconfounded	Normal	89.5%	96.5%	92.0%	92.0%	89.0%	95.0%	96.0%	93.0%	93.5%	96.5%	96.5%
(tail values	Dexp	94.0%	96.5%	88.5%	92.0%	93.5%	96.5%	91.5%	94.5%	96.0%	97.0%	97.0%
more likely	MixNorm	84.5%	94.5%	85.5%	87.5%	88.0%	96.0%	84.5%	87.0%	94.0%	95.0%	92.5%
missing)	MixNChi	87.5%	94.0%	88.5%	90.0%	89.0%	93.5%	88.5%	91.0%	90.5%	93.5%	97.0%
3. Unconfounded	Normal	80.5%	94.5%	93.5%	81.5%	79.5%	94.5%	81.0%	81.5%	95.0%	96.5%	87.0%
(large values	Dexp	82.0%	94.5%	92.0%	80.5%	81.0%	93.0%	80.5%	82.5%	92.0%	94.0%	85.5%
more likely	MixNorm	76.0%	92.0%	91.0%	80.0%	77.0%	93.5%	80.0%	76.5%	93.0%	94.0%	82.5%
missing)	MixNChi	82.0%	93.0%	91.5%	83.0%	84.0%	94.0%	83.5%	81.5%	93.5%	96.5%	89.0%
4. Unconfounded	Normal	88.0%	91.5%	88.5%	90.5%	89.0%	91.5%	89.0%	90.5%	94.0%	97.0%	96.5%
(Center values	Dexp	88.5%	91.0%	86.0%	90.0%	90.0%	93.5%	89.5%	90.5%	90.0%	93.5%	95.0%
more likely	MixNorm	88.5%	92.0%	85.5%	88.5%	87.0%	93.5%	88.0%	89.0%	90.0%	96.5%	96.5%
missing)	MixNChi	86.0%	89.5%	88.0%	89.0%	86.5%	87.0%	87.5%	91.0%	92.0%	94.0%	95.5%
5. Confounded	Normal	87.0%	95.0%	89.0%	91.5%	87.5%	92.5%	90.0%	86.0%	91.0%	95.5%	96.0%
(tail values	Dexp	84.0%	96.0%	91.0%	89.0%	84.5%	94.0%	87.5%	88.0%	88.5%	95.5%	98.0%
more likely	MixNorm	84.5%	95.5%	85.0%	88.5%	88.0%	95.5%	85.0%	86.0%	84.0%	94.5%	96.0%
missing)	MixNChi	74.5%	96.0%	81.0%	75.0%	81.0%	95.0%	74.0%	77.0%	85.0%	95.0%	90.5%

Table 6 Confidence interval width with single imputation (overall #)

Missing Mechanism	Distribution	Mean Imp.	Ratio Imp.	Hot Deck	Random	Mean +error	Ratio +error	ABB	BB	Proc Impute	Schafer	Adj. DA
1. MCAR	Normal	0.453		0.417	0.496	0.518		0.491	0.488	0.466	0.390	0.393
	Dexp	0.629		0.689	0.610	0.713		0.681	0.685	0.598	0.557	0.497
	MixNorm	0.494		0.532	0.598	0.618		0.634	0.585	0.478	0.428	0.481
	MixNChi	1.015		1.504	1.179	1.094		1.134	1.289	0.847	0.841	0.959
2. Unconfounded	Normal	0.383	0.374	0.441	0.425	0.419	0.415	0.355	0.402	0.437	0.364	0.358
(tail values	Dexp	0.463	0.545	0.635	0.494	0.490	0.550	0.495	0.472	0.530	0.496	0.444
more likely	MixNorm	0.481	0.459	0.618	0.515	0.507	0.515	0.544	0.538	0.465	0.447	0.444
missing)	MixNChi	0.658	0.878	1.122	0.801	0.729	0.953	0.720	0.766	0.834	0.878	0.722
3. Unconfounded	Normal	0.434	0.394	0.423	0.477	0.447	0.448	0.446	0.431	0.395	0.377	0.422
(large values	Dexp	0.567	0.572	0.550	0.589	0.545	0.571	0.663	0.588	0.643	0.562	0.546
more likely	MixNorm	0.527	0.532	0.493	0.525	0.550	0.543	0.567	0.510	0.519	0.488	0.465
missing)	MixNChi	0.781	0.866	0.805	0.877	0.848	0.870	0.832	0.846	0.895	0.770	0.825
4. Unconfounded	Normal	0.443	0.402	0.519	0.507	0.499	0.438	0.549	0.517	0.408	0.360	0.377
(Center values	Dexp	0.707	0.632	0.762	0.727	0.720	0.616	0.688	0.783	0.584	0.582	0.562
more likely	MixNorm	0.554	0.496	0.617	0.612	0.600	0.523	0.601	0.622	0.564	0.436	0.474
missing)	MixNChi	1.118	0.997	1.130	1.114	1.310	1.123	1.324	1.026	0.974	0.919	0.936
5. Confounded	Normal	0.379	0.361	0.395	0.377	0.407	0.441	0.418	0.424	0.355	0.312	0.283
(tail values	Dexp	0.460	0.469	0.552	0.501	0.495	0.565	0.483	0.547	0.547	0.446	0.381
more likely	MixNorm	0.432	0.388	0.512	0.436	0.473	0.450	0.529	0.492	0.429	0.376	0.353
missing)	MixNChi	0.627	0.677	0.678	0.685	0.658	0.769	0.698	0.637	0.773	0.622	0.578

The "Overall" combined missing rate is about 10% for missing mechanism 4 and 25% for the others. Results for each missing rate category are available from the authors.

IMPUTATION VARIANCE ESTIMATION IN SCHOOLS AND STAFFING SURVEY

Fan Zhang, Synectics for Management Decisions, Inc. Mike Brick, Westat, Inc., Steve Kaufman, National Center for Education Statistics, Elizabeth Walter, Synectics
Fan Zhang, Synectics, 1901 North Moore St., Suite 900, Arlington, VA 22209

Key Words: Imputation, Variance Estimation, Schools and Staffing Survey

1. Introduction

Missing data is a common problem in virtually all surveys. In cross-sectional surveys, missing data may mean no responses are obtained for a whole unit being surveyed (unit nonresponse), or that responses are obtained for some of the items for a unit but not for other items (item nonresponse). Unit and item nonresponse cause a variety of problems for survey analysts. Missing data can contribute to bias in the estimates and make the analyses harder to conduct and results harder to present.

The most frequently used method to compensate for item nonresponse in National Center for Education Statistics (NCES) surveys is imputation.

In practice, imputed values are often used as true values to estimate the population parameters. However, it is no longer appropriate to use the standard formulae to estimate the variance when there is imputed data. Treating imputed values as observed values can lead to underestimating variances if standard formulae are used. This underestimation may become more appreciable as the proportion of imputed items increases.

Analysts have developed a number of procedures to handle variance estimation of imputed survey data. In particular, Rubin (1987) proposed a multiple imputation procedure to estimate the variance due to imputation by replicating the process a number of times and estimating the between replicate variation. Särndal (1992) outlined a number of model-assisted estimators of variance, while Rao and Shao (1992) proposed a technique that adjusts the imputed values to correct the usual or naive jackknife variance estimator for hot deck imputation. Kaufman (1996) proposed a variance estimation method similar to Särndal's method that can be used with a nearest neighbor imputation approach. Shao and Sitter (1996) proposed to perform an imputation procedure on each bootstrap sub-sample to incorporate the imputation variability. This proposed bootstrap procedure is consistent irrespective of the sampling design, the imputation method, or the type of statistic used in inference. Shao and Sitter's method does not require any model or explicit variance formulae. Once the imputation procedure is programmed appropriately, Shao and Sitter's method is easy to implement. However, since B imputations should be performed for each item, extensive computation is required for large scale surveys. Maintaining the large amount of imputed data can be operationally difficult.

In this study, we applied Shao and Sitter's bootstrap method to the Schools and Staffing Survey (SASS) 1993-94 Public School Teacher Survey component to assess the magnitude of imputation variance.

2. 1993-94 Schools and Staffing Survey (SASS)

SASS 1993-94 Public School Teacher Survey has a two stage stratified sampling design. First, public schools are stratified. Within each stratum, schools are sorted and systematically selected using a probability proportionate to size algorithm. Then within each selected school, teachers are stratified. Within each school and teacher stratum, teachers are selected systematically with equal probability. The SASS 1993-94 Public School Teacher Survey data contains information on the 47,105 public school teachers who responded to the survey. The range of item response rates is 71-100%.

3. SASS 93/94 Imputation Procedure

Four types of imputation methods are used in SASS 1993-94. They are (paraphrasing from Abramson et al., 1996, page 80):

(1) Using data from other items of the same unit on the questionnaire;
(2) Extracting data from a related component of SASS;
(3) Extracting data from the frame file (the information about the sample case from the sampling frame);
(4) Extracting data from the record for a sample case with similar characteristics ("hot deck ").

Imputation methods (1) – (3) are deductive or logical imputation. Whenever it was possible, a item nonresponse was imputed by methods (1) – (3). If a missing item can not be imputed by methods (1) – (3), then imputation method (4) was used. Method (4) is a (sequential) hot deck method. The procedure started with the specification of imputation classes defined by certain relevant variables (matching variables). Then the records were sorted by STGROUP (Groups of states with similar schools) / STATE / TEALEVEL (Instructional level for teacher) / GRADELEV (Grade

levels taught this year) / URB (Type of community where school located) / TEAFIELD (Teaching assignment field) / ENROLMNT (Number of students enrolled in the school). The records were then treated sequentially. A nonmissing *y*-variable was used as a starting point for the process. If a record had a response for the *y*-variable, that value replaced the value previously stored for its imputation class. If the record had a missing response, it was assigned the value currently stored for its imputation class. If there was no donor in the class, the class was collapsed with another class.

For imputation method (1), the imputed values are from other observed items of the same unit and in method (3) the imputed values are from the sampling frame file (PSS or CCD). For imputation method (2), the LEA's (Local Education Agency – another component of SASS) missing item is imputed through information from the sampled school which belongs to that LEA. According to Abramson et al. (1996), this type of imputation was performed only to the one-school LEAs. Therefore, the imputed values by methods (1), (2), or (3) are independent of the sample and the sample design. Assume the simplest response mechanism: respondents always respond and nonrespondents never respond. Then if the population is $\{y_1, y_2, ..., y_N\}$, the imputed values can be assumed to be $\{z_1, z_2, ..., z_N\}$. Here if y_k is actually observed, then $z_k = y_k$, otherwise z_k equals the value imputed by any method of (1), (2), or (3). Let $t_y = \sum_{k=1}^{N} y_k$ be the population total of *y*, $t_z = \sum_{k=1}^{N} z_k$ be the population total of *z*, and $\hat{t}_z = \sum_s z_k / \pi_k$ be the Horvitz-Thompson estimator of t_z (here π_k is the inclusion probability of unit k). We have the following decomposition

$$MSE(\hat{t}_z) = V(\hat{t}_z) + (t_z - t_y)^2 .$$

The first part, $V(\hat{t}_z)$, can be estimated by treating the imputed values as observed values while the second part is the bias of the imputation and assessing this bias is out of the scope of this study. If the imputation bias is small, then treating the values imputed by any method of (1), (2), or (3) as observed values and using a standard variance estimation formula will not underestimate the variance.

For method (4)—the hot deck imputation, however, the imputed data can not be treated as observed data. Actually every imputed value is a function of the sample, therefore the imputed values cannot be represented as a set of fixed values as $\{z_1, z_2, ..., z_N\}$. Therefore in this study, we investigated the imputation variance of method (4) – the hot deck method.

4. Imputation Variance Estimation Procedure

SASS surveys are designed to produce reliable state estimates, and samples are selected systematically without replacement with large sampling rates within strata. To reflect the increase in precision due to large sampling rates, a without replacement bootstrap variance estimator procedure has been implemented for the 1993-94 SASS. Instead of drawing a simple random sample with replacement from the original sample, the bootstrap is done systematically without replacement with probability proportional to size as the original sampling was performed (Abramson et al., 1996).

In SASS 1993-94 components, 48 replicate weights were created to estimate variance using the bootstrap method. These replicate weights were subjected to various adjustments, including a sampling adjustment, a noninterview adjustment, and a ratio adjustment. In order to reflect these adjustments, these replicate weights should be used in the variance estimation. To this end, we used the Shao and Sitter's method in the following manner:

(1) For each set of replicate weights $\{w_{ik}\}_{k=1,2,...,n}$ (i = 1, 2, …, 48), cases with $w_{ik} = 0$ are dropped. Denote the remaining cases, which make up a bootstrap sub-sample, as $\boldsymbol{Y}_{Ii} = \{y_k : k \in A_{Ri}, \eta_k : k \in A_{Mi}\}$ (i = 1, 2, …, 48). Here A_{Ri} is the set of observed values and A_{Mi} is the set of missing values.

(2) Apply the same imputation method as was used to create the full sample imputation values and use $\{y_k : k \in A_{Ri}\}$ to impute $\{\eta_{ik}^* : k \in A_{Mi}\}$ (i = 1, 2, …, 48). This re-imputed bootstrap sub-sample is denoted as s_i. That is

$$s_i = \{y_k : k \in A_{Ri}\} \cup \{\eta_{ik}^{\bullet} : k \in A_{Mi}\},$$

here $\eta_{ik}^{\bullet}$ is imputed value. The missing values in the full sample are also imputed using the nonmissing values in the full sample. This set of imputed values is denoted as

$$s_0 = \{y_k : k \in A_R\} \cup \{\eta_k^* : k \in A_M\}.$$

Thus, 48 sets of imputed bootstrap sub-samples and 1 set of imputed full sample are obtained.

(3) Calculate the $\hat{\theta}_i$ of interest from s_i, weighted by replicate weights $\{w_{ik}\}$ $(i = 1,...48)$, and the $\hat{\theta}$ from full sample s_0, weighted by the full sample weight $\{w_k\}$. The variance of $\hat{\theta}$ is estimated by

$$v(\hat{\theta}) = \frac{1}{48} \sum_{i=1}^{48} (\hat{\theta}_i - \hat{\theta})^2 .$$

Another difference between the variance estimator we used above and Shao-Sitter's estimator is that in our formula the deviation is around the full sample estimate $\hat{\theta}$ whereas in Shao-Sitter's formula the deviation is around the average of the bootstrap estimates $\bar{\theta}^*$. The balanced repeated replication method (BRR) is implemented in WesVar PC, but the bootstrap method is not. Abramson et al. (1996) suggests that with any BRR software package, the BRR option should be specified for 1993-94 SASS data analysis. The formulae used in WesVar PC for the BRR option is the formula we used above. In general,

$$\sum_{i=1}^{B}\left(\hat{\theta}_i-\bar{\theta}^*\right)^2 \le \sum_{i=1}^{B}\left(\hat{\theta}_i-\hat{\theta}\right)^2 = \frac{1}{B}\sum_{i=1}^{B}\left(\hat{\theta}_i-\bar{\theta}^*\right)^2+\left(\bar{\theta}^*-\hat{\theta}\right)^2$$

here $\bar{\theta}^* = B^{-1}\sum_{i=1}^{B}\hat{\theta}_i$. Notice

$$E\left(\bar{\theta}^*-\hat{\theta}\right)^2 = E_P E_B\left(\bar{\theta}^*-\hat{\theta}\right)^2.$$

Here E_P is with respect to sample design, E_B is with respect to bootstrap subsampling, and typically $E_B\left(\bar{\theta}^*\right)=\hat{\theta}$. Therefore $E_B\left(\bar{\theta}^*-\hat{\theta}\right)^2 = Var_B\left(\bar{\theta}^*\right)$. An unbiased estimator of $Var_B\left(\bar{\theta}^*\right)$ is

$$\hat{V}_B\left(\bar{\theta}^*\right)=\frac{1}{B}\frac{1}{B-1}\sum_{i=1}^{B}\left(\hat{\theta}_i-\bar{\theta}^*\right)^2.$$

Therefore

$$\frac{1}{B}\sum_{i=1}^{B}\left(\hat{\theta}_i-\hat{\theta}\right)^2 \approx \left(1+\frac{1}{B-1}\right)\frac{1}{B}\sum_{i=1}^{B}\left(\hat{\theta}_i-\bar{\theta}^*\right)^2.$$

When B is large the bias in variance estimation is small and can be easily corrected by factor $(B-1)/B$. In our study, we compare standard error estimates instead of variance estimates and $B=48$, so the adjustment factor is $\sqrt{47/48}\approx 0.99$. We do not apply this adjustment because it is close to 1. In addition, we use the same formula to calculate both the standard error estimates cooperating imputation variance and the standard error estimates without cooperating imputation variance. And the ratio of these two types of standard error estimates is used as the measurement of the difference. Therefore, the adjustment factor has no effect on this ratio.

The variables used for this study include 6 categorical variables and 7 continuous variables. Their stage 2 imputation—method (4), rates range from 2 percent to 25 percent (see table 1).

Most of the variables used for sorting or matching the records are not included in the data file; they had to be reconstructed by using other variables in the data file. This caused a discrepancy between the data imputed for this study and the original imputed data in the file. To prevent confounding the imputation difference with imputation variance, we imputed the full sample with our sorting and matching variables and denote this imputed full sample as s_0. This is the sample used in the variance estimation (see imputation procedure step 3 above).

5. Imputation Variance Estimates

From Table 2 to Table 4, we compare standard errors which do not take the imputation variance into account ($ste(\hat{\theta})$) with the standard errors incorporated with imputation variance ($ste_I(\hat{\theta})$). It is important to emphasize that both $ste_I(\hat{\theta})$ and $ste(\hat{\theta})$ are estimates of standard errors instead of true standard errors and therefore both of them are also subjected to sampling errors.

Table 2 compares standard errors for the total estimator and the average estimator of continuous variables. The output shows the imputation does not inflate the variance for the total very much. For variable T0985, the standard error increases only 7 percent even though the imputation rate is as high as 27 percent. For the average per person estimators of continuous variables, the underlying estimator is actually a nonlinear estimator. When the imputation rate is high, inflation to the variance can be very high, too. For example, variable T0985 now shows $ste_I(\hat{\theta})$ is 41 percent higher than $ste(\hat{\theta})$. So if the imputed data are treated as true values, the underestimation can be severe.

Table 3 compares standard errors for the ratio estimators of continuous variables. Variable BASIC is the ratio of teacher's basic salary to teacher's total income. Variable INSCH is the ratio of teacher's total income at school to teacher's total income. OUTSCH is the ratio of teacher's total income from outside of school to teacher's total income. ADITION is teacher's other income from school (total income inside school minus base salary) to teacher's total income. IN_OUT is teacher's total income inside school to teacher's total income outside school. Although some variables used for the ratios have high imputation rates (T1440, for example, has a 21.3% imputation rate) the increase in standard errors are very small. Again, for continuous variables, we observed smaller inflation in standard error.

Table 4 compares standard errors for the total estimator and percentage estimator of categorical variables. Here the total estimates are estimated total counts in each category and the percentage is the estimated percent of units in each category. Notice the inflation in variance is larger than the continuous variables. This might be due

to the fact that the sample sizes of the categorical variables are smaller (there is more legitimate skipping for these items). It also shows that when imputation rates get higher, the increase in standard errors also gets larger. Now variable T0040 shows the biggest inflation: 2.04.

6. Summary

The techniques developed so far for the variance estimation of imputed data are not yet easy to implement or operationally convenient. Shao and Sitter's method is appealing but requires repeated imputations, so for large scale surveys the data files become too large.

For the deductive imputation methods (1) – (3), the imputed value can be treated as observed value and the use of standard formula should not cause variance underestimation.

Our empirical study shows that using the hot deck imputation method in the 1993-94 SASS can seriously affect the standard error especially for the discrete variables with small sample size.

But notice that the majority of items have very low hot deck imputation rates. For the SASS 1993-94 Public School Teacher component, only 11 out of 249 items had hot deck imputation rates above 10 percent (see Gruber, Rohr, and Fondelier, 1996, figure VIII-24, pp. 231-235). We used six of those items for this study. And, when the imputation rate is low, the inflation in variance is not severe, especially for continuous type variables with large sample size, no matter it is a linear or ratio estimator.

References

Abramson, R., Cole, C., Fondelier, S., Jackson, B., Parmer, R., and Kaufman, S. 1996. *1993-94 Schools and Staffing: Survey Sample Design and Estimation*. (NCES 96-089). Washington, D.C.: U.S. Department of Education, Office of Educational Research and Improvement. National Center for Education Statistics.

Gruber, K., Rohr, C., and Fondelier, S. 1996. *1993-94 Schools and Staffing Survey: Data File User's Manual, Volume I: Survey Documentation.* (NCES 96-142). Washington, D.C.: U.S. Department of Education, Office of Educational Research and Improvement. National Center for Education Statistics.

Kaufman, S. 1996. Estimating the variance in the presence of imputation using a residual. In *1996 Proceedings of the Section on Survey Research Methods* (pp. 423-428). Alexandria, VA: American Statistical Association.

Rao, J. N. K. and Shao, J. 1992. Jackknife variance estimation with survey data under hot deck imputation. *Biometrika* 79(4): 811-822.

Rubin, D. B. 1987. *Multiple Imputation for Nonresponse in Surveys*. New York: J. Wiley & Sons, Inc.

Särndal, C. E. 1992. Methods for estimating the precision of survey estimates when imputation has been used, *Survey Methodology*, 18(2): 241-252.

Shao, J. and Sitter, R. R. 1996. Bootstrap for imputed survey data, *Journal of the American Statistical Association*, 91: 1278-1288.

Table 1: Variables used in this study

Name	Label	Hot Deck imputation rate (%)	Type
T0030	2 Full/Part-time teacher at this school	11.8	5 Categories
T0035	3A Have other assignment at this sch	9.8	Dichotomous
T0040	3B What is other assignment at this sch	24.0	6 Categories
T0140	11D Consecutive yrs teaching since break	5.2	Continuous
T0435	28A Any mathematics courses taken	5.7	Dichotomous
T0645	32B Programs changed views on teaching	2.0	5 Categories
T0860	40B(4) Number of students in the class	13.6	Continuous
T0985	41C Number of separate classes taught	27.0	Continuous
T1420	53B(1) Academic yr base tchng salary	8.3	Continuous
T1430	53B(2) Additional compensation earned	4.0	Continuous
T1440	53B(3) Earning from job outside sch sys	21.3	Continuous
T1455	53B(5) Income earned from other source	5.9	Continuous
T1520	55 Total income of all HHD family member	25.0	12 Categories

Source: Abramson et al. (1996).

Table 2: Standard error comparison for total estimates and average estimates of continuous variables

Name	Hot Deck imputation rate (%)	Total Estimate	$ste_I(\hat{\theta})/ste(\hat{\theta})$	Average Estimate*	$ste_I(\hat{\theta})/ste(\hat{\theta})$
T0140	5.2	8985367	0.99	11.01	0.96
T0860	13.6	24958128	1.01	22.79	1.10
T0985	27.0	2107888	1.07	12.79	1.41
T1420	8.3	86349560396	1.00	33713.26	1.01
T1430	4.0	1865774738	1.03	2093.88	1.05
T1440	21.3	2179435663	1.03	4384.44	1.05
T1455	5.9	588847739	1.01	1676.05	1.03

- These estimates are average per teacher.

Table 3: Standard error comparison for ratio estimates of continuous variables

Name	Hot Deck Imputation rate (%)	Estimate	$ste_I(\hat{\theta})/ste(\hat{\theta})$
Basic	--	0.94907	1.01
Insch	--	0.96957	1.03
Outsch	--	0.02395	1.02
Addition	--	0.02051	1.05
In_out	--	31.87	1.03

Basic = T1420/(T1420 + T1430 + T1440 + T1455)
Insch = (T1420 + T1430)/(T1420 + T1430 + T1440 + T1455)
Outsch=T1440/(T1420 + T1430 + T1440 + T1455)
Addition=T1430/(T1420 + T1430 + T1440 + T1455)
In_out=(T1420 + T1430)/(T1440 + T1455)

Table 4: Standard error comparison for total estimates and percentage estimates of discrete variables

Name	Hot Deck imputation rate (%)	Categories	Total Estimate	$ste_I(\hat{\theta})/\ ste(\hat{\theta})$	Percentage Estimate (%)	$ste_I(\hat{\theta})/\ ste(\hat{\theta})$
T0030	11.8					
		1	12994	1.10	5.61	1.10
		2	31489	1.14	13.60	1.18
		3	97607	1.12	42.15	1.19
		4	52767	1.11	22.79	1.13
		5	36706	1.38	15.85	1.35
T0035	9.8					
		1	54006	1.08	24.45	1.09
		2	166845	1.00	75.55	1.09
T0040	24.0					
		1	9613	1.44	13.49	1.54
		2	11737	2.04	16.47	2.09
		3	5093	1.26	7.15	1.29
		4	12311	1.73	17.28	1.66
		5	26962	1.27	37.84	1.52
		6	5543	1.62	7.78	1.71
T0435	5.7					
		1	2001004	0.99	78.12	0.98
		2	560289	1.00	21.88	0.98
T0645	2.0					
		1	122310	0.99	5.42	0.98
		2	822249	1.01	36.41	1.01
		3	498908	1.00	22.09	1.03
		4	711355	1.01	31.50	1.01
		5	103472	0.98	4.58	0.97
T1520	25.0					
		1	173	1.45	0.01	1.60
		2	863	1.63	0.03	1.68
		3	8850	1.03	0.35	1.04
		4	72952	1.18	2.85	1.15
		5	123771	1.19	4.83	1.22
		6	154036	1.10	6.01	1.12
		7	174850	1.18	6.83	1.16
		8	404821	1.18	15.81	1.30
		9	434259	1.08	16.95	1.14
		10	523142	1.27	20.42	1.26
		11	438739	1.12	17.13	1.19
		12	224836	1.22	8.78	1.21

A BOOTSTRAP VARIANCE ESTIMATOR FOR SYSTEMATIC PPS SAMPLING

Steven Kaufman, National Center for Education Statistics
Room 422d, 555 New Jersey Ave. NW, Washington, DC 20208

Key Words: Simulation, Half-Sample Replication

1. Introduction

Systematic probability proportionate to size (PPS) sampling procedures (Wolter, 1985, section 7.6) are efficient in terms of ease of selection and lowering sampling error. For this reason they are used extensively in large-scale surveys. Since each stratum systematic sample is selected using a single random start, the sample can be viewed as a sample of size one, where each sample consists of a single sample cluster of n_h primary sampling units (PSUs). Therefore, it is impossible to produce an unbiased variance estimator since the sample size is one. However, a number of biased methodologies are used for variance estimation.

These methodologies generally take one of two forms: 1) assume the systematic sample can be approximated by a simpler sample design with a known variance estimator or 2) assume the response variable follows some super-population model and a variance estimator is produced appropriate for that model. Both these approaches allow for grouping of PSUs, so variances can be computed within groups. Wolter (1985, chapter 7) provides a good discussion of a number of systematic sample variance estimators that can be classified into one of these two forms. An example, using balanced half-sample replication (BHR) is provided below.

BHR is a widely used variance replication methodology for complex survey designs. It is designed for samples where two PSUs within each stratum are selected with replacement. With BHR, choosing one PSU within each stratum generates a half-sample. A number of half-samples are generated by alternating which PSU, within stratum, go into the half-samples. The BHR variance is the simple variance of the half-sample estimates. Through a balancing process of the half-samples, the BHR variance estimate, for linear estimates, equals the direct sample variance estimate.

BHR can be adapted to designs where more than two PSUs are selected in a stratum by consecutively pairing selected PSUs, after placing them in the original order of selection; and assuming each pair is a stratum for variance estimation (variance stratum). If without-replacement sampling is used then a finite population adjustment can be applied. See (Wolter, 1985 pp. 110-152) for a more complete description of BHR.

In order to use BHR with systematic PPS sampling, it must be assumed that a PPS selection can be approximated by the deep stratification induced by the pairing described above. This assumption is reasonable, considering that the first sort variable, ignoring the lack of independence between breaks in the variable, can be considered an implicit stratification. However, BHR also assumes that the variance estimate is proportional to the inverse of the sample size. (This follows from $V_{BHR}(X) = V((X_1 + X_2)/2) = 1/2V(X_1)$, where subscript 1 and 2 represents the estimate based on the first and second PSUs respectively selected in each stratum.) In section 2.0, it will be demonstrated, through a simulation study, that systematic sampling variances are not necessarily inversely proportional to the sample size.

2.0 Using the BHR Model with Systematic Sampling

For BHR model to work, the stratum variances must be proportional to $1/n_h$, as n_h increases or decreases, since BHR makes this assumption. If this assumption is not true then the BHR model is unlikely to produce accurate results. To investigate the $1/n_h$ assumption, a simulation study is done, using the sample design described in section 4.2. Four thousand systematic PPS samples are selected with sample sizes of n_h and $0.5n_h$. By computing the simple variance of the 4,000 simulation estimate, an estimate of the true variance is computed. This is done for estimates of total students, teachers and schools. If the variance is proportional to $1/n_h$, then the ratio, $R100/50 = \hat{V}_{n_h}(X_l)/\hat{V}_{0.5n_h}(X_l) \times 0.5 - 1$ should be close to 0; where l represents the estimate type (total students, teachers or schools). When the ratio is less (greater) than 0, the systematic sample variance decreases faster (slower) than the $1/n_h$ assumption would imply. A negative (positive) ratio means that BHR should overestimate (underestimates) the variance.

Table 1 demonstrates that sometimes the ratio is close to 0. Other times, it is a great deal different than 0. The systematic PPS sampling variance does not necessarily decrease faster than the $1/n_h$ assumption would imply; sometimes its decrease is slower. This is an indication that BHR will not necessarily produce an overestimate of the variance, which is a common assumption among sampling statisticians. When there is a large difference from 0, the magnitude is dependent on the variable. This seems to imply, since the sampling rates are not high, that the violation of the $1/n_h$ assumption is due to the initial sort ordering (i.e., within sample correlation).

It should be noted that the table 1 results exaggerate the true impact of the $1/n_h$ assumption. Using the $1/n_h$ assumption, the ratio, used in the table, adjusts the

variance with the smaller sample size to approximate the variance with the larger sample size. This approximation uses the smaller sample estimate's unknown finite population correction. Since the true finite population correction is likely larger than the one used in the approximation, the absolute value of the true impact of the $1/n_h$ assumption should be expected to be smaller than what table 1 indicates.

The important conclusion from this example is that variance estimates, based on designs using systematic sampling, will not necessarily be proportional to $1/n_h$, as n_h increases or decreases. When this occurs, an important BHR assumption is violated, and the BHR variance estimator should not be expected to perform well when the magnitude of the violation is large.

The statements concerning the proportionality of the variance estimate are qualified with 'as n_h increases or decreases'. The importance of this qualification can be seen with equal probability systematic sampling. Here, the variance can be expressed proportional to $1/n_h$ (e.g., $V(\bar{y}_h) = [(N_h - n_h)/N_h][S^2_{wst}/n_h][1+(n_h-1)\rho_{wst}]$, see (Cochran, 1977, pp. 209)). If S^2_{wst} and ρ_{wst} are constant for an arbitrary n_h then $V(\bar{y}_h)$ would be approximately proportional to $1/n_h$, as n_h increases or decreases. However, both S^2_{wst} and ρ_{wst} are within systematic sample population estimates. This implies that as n_h changes, the systematic samples change; hence S^2_{wst} and ρ_{wst} also change by some unknown function of n_h. Therefore, even though $V(\bar{y}_h)$ is proportional to $1/n_h$ for fixed n_h, as n_h increases or decreases, the variance may not be proportional or even closely proportional to $1/n_h$.

3.0 Bootstrap Variance Model

To address the situation when the systematic variance in not proportional to $1/n_h$, a bootstrap variance estimator is proposed in this paper, which is less dependent on the $1/n_h$ assumption than the BHR estimator. This section first describes the consistency theorem for the bootstrap estimator; by example, the super-population model, used in the proposed bootstrap procedure, is demonstrated; next, the mechanics of the bootstrap procedure is presented; and finally, the consistency of the bootstrap procedure is established. We begin by describing the super-population model.

3.1 The Consistency Result

Theorem

The required assumptions are:

1) a systematic PPS sample (s_{ih}) has a known partition (i.e., $s_{ih} = \bigcup_{c=1}^{C_{ih}} s_{ihc}$);
2) $\hat{X} = \sum_h \sum_{j\in h} w_{hj} x_j = 1/n \sum_h \sum_{j\in h} y_j$ is the estimate of interest, with w_{hj} being the sampling weight and x_j being the variable of interest;
3) as n increases the sample allocation between stratum remains constant;
4) for PSUs in s_{ihc}, the y_j's are conditionally i.i.d given s_{ihc} and are generated from an otherwise unspecified distribution function $F_{hc}(y)$ satisfies conditions for Mallows' distance;

and

5) between partitions, the y_j's are conditionally independent given the s_{ihc}'s, but not identically distributed.

Bullets 1-5 specify the super-population model.

It then follows that the bootstrap variance estimator of $\hat{X}$ given s_{ih} generated from the bootstrap estimates $\hat{X}^*_b = 1/n \sum_h \sum_{j\in h} y^*_j$, where the y^*_j's are generated from $\hat{F}_{hc}(y)$, is consistent, as $n \to \infty$, provided $\hat{F}_{hc}(y) \to F_{hc}(y)$ and $\mu^*_{yhc} \to \mu_{yhc}$, as $n \to \infty$. μ^*_{yhc} is the bootstrap expectation of y within a partition.

The proof follows from the super-population assumptions using the argument in example 3.1 from (Shao and Tu, 1995). The details are provided in (Kaufman, 1998).

3.2 Bootstrap Model Example

In practice, the statistician never knows the required partitioning ($s_{ih} = \bigcup_{c=1}^{C_h} s_{ihc}$). However, the statistician usually orders the frame before sample selection. With this ordering, the statistician is implicitly assuming that nearby PSUs are similar, at least in terms of the most important response variables. This implicit assumption can be used to develop a partitioning that approximately meets the required assumptions.

An example is provided below.

For a fixed even numbered sample size (n_h), the elements of the partition (s_{ich}) can be determined by pairing the sets of PSUs within consecutive sampling intervals, after the frame has been placed in its original sort ordering. All samples have the same partitioning (i.e., the partitioning is only a function of stratum, -- s_{ch}, $c = 1 \text{ to } C_h$) and each s_{ich} (s_{ch}) has exactly two PSUs. In terms of consistency, it is assumed that the

partitioning remains fixed as the sample size increases and more PSUs are selected within a partition. This "type" of partitioning is used in the bootstrap procedures proposed in this paper.

An additional observation about this partitioning is:

If the partitioning methodology described above correctly models the distribution of X; the n_h's are even and increase by multiples of C_h then the $\underset{2}{E}(\hat{X}_i | s_h's) = K$, a constant; where $\underset{2}{E}$ refers to the expectation with respect to the super-population model. Therefore,

$$V(\hat{X}_i) = \underset{1}{E}\underset{2}{V}(\hat{X}_i | s_h's) + \underset{1}{V}\underset{2}{E}(\hat{X}_i | s_h's) = \underset{1}{E}\underset{2}{V}(\hat{X}_i | s_h's),$$

where 1 refers to the selection of the s_h's. Since the bootstrap variance estimator is consistent for $\underset{2}{V}(\hat{X}_i | s_h's)$, the bootstrap variance is consistent to an unbiased estimator for the unconditional variance.

3.3 Bootstrap Sample Size (n_h^*)

Since it is assumed that the relationship between the variance and n_h is unknown, the actual bootstrap sample size (n_h^*) used in the bootstrap selections must be computed through a series of trial and error simulations. This is done by comparing and estimate of the true variance with the bootstrap variance for a specific bootstrap sample size.

Determining n_h^* through a simulation provides a robust variance estimate because $V^*(\hat{X}_h)$, by construction, will be almost unbiased, even if the model assumptions are false. The disadvantage of the simulation is that it can only be implemented with frame variables. However, if n_h^* is relatively flat for non-frame variables, the bootstrap replicate weights should be applicable for those variables, too.

3.4 Bootstrap Implementation

To perform the simulation study, frame variables are used, so estimates can be computed for any selected sample. The statistician always has three estimates available for this purpose. One is the measure of size or some function of the measure of size. The second is the estimate of the total number of PSUs (sum of the sample weights). The third is the average measure of size per PSU or the average per PSU of some function of the measure of size. If the measure of size is used in the simulation, it will be necessary to use a different year's data to produce estimates; otherwise, the variances will be zero.

To determine the appropriate n_h^*'s, the simulations must first be applied to individual stratum estimates Θ_h. The simulation process for estimating the bootstrap variance, $V^*(\Theta_h)$ for an estimator Θ_h, works as follows:

3.4.1 Bootstrap Procedures

1. Select a sample (s_i) from the original frame, using the PPS methodology of the original sample design.
2. For the initial bootstrap sample size values, n_h^*, use n_h. After the initial simulation, n_h^* will likely require adjustment for at least some of the strata.
3. Generate a bootstrap frame based on the selected sample. For each selected PSU j, w_j bootstrap PSUs (bj) are generated by replicating the j^{th} PSU w_j times. The bj^{th} bootstrap-PSU has the following measure of size (m_{bj}):

   $m_{bj} = I_{bj} \cdot 1/w_j$,

   $$I_{bj} = \begin{cases} 1, \text{ if } bj \text{ is an integer component of } w_j \\ C_i, \text{ if } bj \text{ is a noninteger component of } w_j \\ \quad C_j \text{ being the noninteger component} \end{cases}$$
4. Randomize the bootstrap frame according to super-population model specification. This is accomplished by placing the bj bootstrap-PSUs generated from PSU j within stratum h and sample s_i in their original order of selection. Next, bootstrap-PSUs generated from the first PSU are paired with the next set of bootstrap-PSUs generated from the second PSU. The third set of bootstrap-PSUs is paired with the fourth set. This process continues until all bootstrap-PSUs are paired. If there are an odd number of PSUs then the last set of groupings of bootstrap-PSUs contains the bootstrap-PSUs generated from the last three PSUs in stratum h. This is repeated for every stratum in s_i. Now, the bootstrap-PSUs are randomized within their respective pair.
5. The bootstrap frame, bootstrap frame ordering, measure of size (m_{bj}), and bootstrap sample size (n_h^*) have been specified. Select B bootstrap samples, after re-randomizing the bootstrap-PSUs after each selection, using the same procedures used to select the original systematic PPS sample. The one exception to this is that a bootstrap-PSU generated from noncertainy PSUs that become certainty in the bootstrap selection should not be eliminated from the selection process and taken in sample with probability 1. Their selection probability should remain unchanged and if the bootstrap-PSU is selected multiple times that should be reflected in the bootstrap weight (see 6 below).

6. For each bootstrap sample, b, compute a set of bootstrap weights, w_j^{*b}, and then compute Θ_{ibh}^* by using w_j^{*b} instead of w_j in the formula for Θ_{ih}.

   The bootstrap-PSU weight, w_j^{*b}, is:

   $w_j^{*b} = \sum_{bj \in S_j^B} w_{bj}^p$, S_j^B is the set of all bj generated from j that are selected in the b^{th} bootstrap sample.

   and $w_{bj}^p = I_{bj} \cdot M_{bj} / p_{bj}$

   M_{bj} : is the number of times the bj^{th} bootstrap- PSU is selected,

   p_{bj} : is the bootstrap selection probability for the bj^{th} bootstrap-PSU.

   $p_{bj} = m_{bj} / SI_h$, $SI_h = \sum_{bj \in s_h} m_{bj} / n_h^*$.

7. The bootstrap variance is:

   $V^*(\Theta_{ih}) = 1/(B-1)\sum_{b=1}^{B}(\Theta_{ibh}^* - \overline{\Theta}_{ih}^*)^2$,

8. Repeat steps 1-7, for a large number of samples (T).
9. Compute the simple variance of Θ_{ih} from $i = 1$ to T, $\hat{V}(\Theta_h)$, as a measure of the true variance; and compute the average bootstrap variance $\overline{V}^*(\Theta_h)$, averaged over the T, $V^*(\Theta_{ih})$ estimates.
10. Compare $\overline{V}^*(\Theta_h)$ with $\hat{V}(\Theta_h)$ and adjust n_h^* to reduce the bias between $\overline{V}^*(\Theta_h)$ and $\hat{V}(\Theta_h)$.
11. Repeat steps 1-10, until this bias has been reduced to a satisfactory level.
12. Using the n_h^* from step 11, repeat steps 3-6 for the actual collected sample, generating a set of bootstrap replicate weights, w_j^{*b} that can be used to compute variances of other, more complex statistics that are not necessarily computed within h.

3.5 Consistency of the Bootstrap Estimator

$\hat{F}_{hc}(y) \to F_{hc}(y)$ and $\mu_{yhc}^* \to \mu_{yhc}$, as $n \to \infty$ follows from $E_*(\sum_{bj \in D} w_{bj}^p X_{bj}) = \sum_{bj \in D} I_{bj} X_{bj} = \sum_{j \in D} w_j X_j$, where D is a domain and E_* is the bootstrap expectation. See (Kaufman, 98) for details.

4.0 Simulation

To demonstrate the advantages of the bootstrap variance estimator, a simulation study is presented comparing BHR and the bootstrap variance estimator. Two thousand simulations, denoted by *s*, are generated using frame variables. The frame is the National Center for Education Statistics' (NCES) Private School Survey (PSS). The PSS is NCES's school frame for private elementary and secondary schools. Three totals (number of schools, number of teachers, and number of students), two averages (average students and average teachers per school), and one ratio (ratio of number of students to number of teachers) are estimated in the simulation. In tables 3-5, estimates are computed by each stratification variable (affiliation, region and school level), as well as one of the sort variables (Urbanicity). The School and Staffing Survey (SASS) sample design is used to select the simulation samples. Relative error, relative mean square error, and coverage rates are used to measure performance.

4.1 Comparison Statistics

In this section, the statistics used to compare the bootstrap and BHR variances are described.

4.1.1 Relative Error

Rel. Error $= (\overline{V}_e(\Theta)^{1/2} / V_t(\Theta)^{1/2} - 1) \cdot 100$

Where: $\overline{V}_e(\Theta)$ is the average of the variance estimates ($V_e(\Theta_s)$) from either the bootstrap or BHR procedure.

$V_t(\Theta) = 1/1999 \sum_{s=1}^{2,000} (\Theta_s - \overline{\Theta})^2$.

4.1.2 Relative Mean Square Error

$\{[VV_e(\Theta) + (\overline{V}_e(\Theta) - V_t(\Theta))^2]^{1/2} / V_t(\Theta)\} \cdot 100$,

Where: $VV_e(\Theta) = 1/1999 \sum_{s=1}^{2,000} (V_e(\Theta_s) - \overline{V}_e(\Theta))^2$.

4.1.3 Coverage Rates

The coverage rate is the percent of the time the 95% confidence interval contains the true value.

4.2 SASS Sample Design

The sample frame, used in the simulation, is the list frame component of NCES's Private School Survey (PSS). The list frame is stratified by detailed School Association (19 groups), within Association by Census Region (4 levels), and within Region by school level (elementary, secondary and combined). The school sample is selected using the systematic probability proportionate to size sampling procedure, described in the introduction. The measure of size is square root of the number of teachers in the school. Before sample selection, the school frame is ordered by state, school highest grade, urbanicity, zip code, and school enrollment. To reduce the time to complete 2,000 simulation only one detailed school association is used.

4.3 Determining n_h^* for the Bootstrap Variance

As described in section 3.3, the determination of n_h^* requires a simulation study in itself. For each stratum, a series of simulations was done for various n_h^*. The optimum n_h^* is likely dependent on the estimate of interest. Since we want only one set of replicate weights, a compromise n_h^* is determined that works reasonably

well for all estimates. The results presented below use the compromise set of n_h^*. Table 2 presents the values for n_h and n_h^*. Each simulation used in the determination of n_h^* had at least 250 samples.

4.4 BHR Variances

The r^{th} school half-sample replicate is formed using the usual textbook methodology (Wolter, 1985) for establishment surveys with more than 2 units per stratum. This is described in the introduction. Two BHR variance estimates are presented. The first (BHR without FPC Adjustment) is the variance estimates described above. This estimate does not make any type of Finite Population Correction (FPC) adjustments. The second BHR variance estimate (BHR with FPC Adjustment) adjusts the first variance estimator by $1 - P_h$, where P_h is the average of the selection probabilities for the selected units within stratum h.

4.5 Number of Replicates

Thirty-two and thirty replicates have been used in the BHR and bootstrap variances, respectively.

4.6 Results

Because of space consideration, three tables have been excluded from this paper. These tables are included in (Kaufman, 1998).

According to tables 3-5, in terms of extremes, the bootstrap variance estimator is better than either BHR variance estimator with respect to relative error, relative MSE, or coverage rate. The bootstrap relative errors are large in absolute value (greater than 20% or less than –20%) once, while the BHR, with and without FPC adjustment, relative errors are large 8 and 4 times, respectively.

Only 5 of the bootstrap relative MSEs are larger than 50% and none are greater than 100%. The BHR without FPC adjustment has 18 relative MSEs larger than 50% and 3 greater than 100%. The FPC adjusted BHR has 14 relative MSEs larger than 50% and 2 larger than 100%.

The bootstrap procedure has no high coverage rates (coverage rate greater than 98%) and 1 low coverage rates (coverage rate less than 89%). The bootstrap has no coverage rate greater than 99%. The BHR without FPC adjustment has 7 high coverage rates, no low coverage rate and 5 larger than 99%. Even with a FPC adjustment, the BHR has 6 high coverage rates, 1 low coverage rate, and 5 coverage rates greater than 99%.

The difference between the bootstrap and BHR is largest for the Urbanicity estimates. For these estimates the BHR relative MSE can be almost 4 times larger than the bootstrap relative error (see tables 3 and 4 Urban). One possible explanation for this may be that the Urbanicity sample size is indirectly controlled by the third sort variable, while the other estimates are directly controlled by the stratification.

4.7 Conclusion

This paper discussed how BHR can be used to measure the variances from surveys utilizing systematic PPS selection procedures. Two assumptions are necessary: 1) the extra stratification introduced by the variance stratum is sufficient to reflect the systematic process and 2) the variance is inversely proportional to the sample size. In table 1, it has been observed that systematic PPS sampling variances may not be inversely proportional to the sample size.

To correct this problem, a bootstrap variance estimator has been introduced which does not make the inverse sample size assumption. Given an appropriate super-population model, the bootstrap procedure produces consistent variance estimates. Based on the simulation of the SASS survey design (Tables 3-5), the bootstrap variance estimator performs better than the BHR with respect to relative error, relative MSE and coverage rates. This is especially true with the Urbanicity estimates. One drawback of the proposed bootstrap procedure is that the determination of an appropriate bootstrap sample size can only be implemented using frame variables. However, with appropriate frame variables, the bootstrap variances are close to unbiased, even when the super-population model assumption fails.

5.0 References

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

Kaufman, S. (1998), *A Bootstrap Variance Estimator for Systematic PPS Sampling*, Working Paper No. 98-12, U.S. Dept. of Ed, NCES, Washington, D.C.

Shao S. and Tu, D. (1995). *The Jackknife and Bootstrap*. New York: Springer-Verlag.

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

Tables 1 – Measurement of degree the true systematic sampling variance is proportional to $1/n_h$ with respect to different sample sizes

Stratum (h)	n_h / N_h (%)	Teacher R100/50 (%)	Student R100/50 (%)	Schools R100/50 (%)
01911	2.0	-31.2	14.1	-28.0
01912	2.8	-27.0	-2.5	-14.3
01914	3.4	-23.8	-5.3	-19.3
01931	4.5	23.7	18.0	4.6
01932	4.9	2.4	-25.8	3.9
01934	4.3	-20.4	-26.3	-7.0

Table 2 – Original (n_h) and Bootstrap (n_h^*) Sample Size by Stratum

Stratum	n_h	n_h^*	Stratum	n_h	n_h^*	Stratum	n_h	n_h^*
01911	14	12	01921	10	5	01931	48	35
01912	16	11	01922	10	8	01932	46	33
01913	52	28	01923	10	10	01933	114	81
01914	34	24	01924	10	10	01934	52	40

Table 3 -- % Relative Error, % relative Mean Square Error and % coverage rates for the Bootstrap and BHR variance estimator for Schools estimates by Affiliation, Region, Level and Urbanicity

Estimate	Bootstrap			BHR without FPC Adjustment			BHR with FPC Adjustment		
	Rel. Error	Rel. MSE	Cov. Rate	Rel. Error	Rel. MSE	Cov. Rate	Rel. Error	Rel. MSE	Cov. Rate
Other Affil.	-4.5	27.8	93.1	12.7	40.7	97.0	9.7	35.4	97.0
Northeast	4.3	43.6	94.6	10.3	52.3	94.9	8.0	49.0	94.9
Midwest	4.2	42.8	92.9	12.5	51.4	98.3	9.8	46.9	95.7
South	-10.9	32.7	90.7	-6.6	26.5	89.6	-10.3	29.0	89.4
West	-2.4	35.1	92.9	7.8	43.7	92.2	5.1	40.0	92.2
Elementary	1.3	34.9	93.6	16.1	57.0	95.9	14.0	52.9	95.9
Secondary	-2.9	57.0	90.5	26.3	107.1	97.2	14.8	81.9	95.9
Combined	-6.2	29.5	91.2	-1.1	28.2	92.3	-4.2	27.7	92.3
Rural	7.5	36.8	95.7	24.2	71.2	98.7	20.9	63.8	98.7
Suburban	6.5	36.6	95.0	23.1	67.5	97.4	19.9	60.6	97.4
Urban	11.5	43.2	96.1	53.7	147.6	97.5	49.5	135	97.5

Table 4 -- % Relative Error, % relative Mean Square Error and % coverage rates for the Bootstrap and BHR variance estimator for Teachers per School estimates by Affiliation, Region, Level and Urbanicity

Estimate	Bootstrap			BHR without FPC Adjustment			BHR with FPC Adjustment		
	Rel. Error	Rel. MSE	Cov. Rate	Rel. Error	Rel. MSE	Cov. Rate	Rel. Error	Rel. MSE	Cov. Rate
Other Affil.	-5.8	28.4	92.4	4.0	27.3	95.9	1.0	24.6	95.8
Northeast	2.1	42.1	93.7	0.6	41.2	90.9	-1.7	39.4	90.8
Midwest	-0.7	37.7	92.5	18.2	60.7	99.7	15.0	54.0	99.7
South	-10.9	32.5	89.4	-9.6	28.0	89.4	-13.2	31.4	88.1
West	5.5	41.3	95.1	12.1	45.4	93.6	9.2	40.1	93.6
Elementary	4.6	38.7	94.0	17.7	57.2	97.1	15.3	52.4	97.1
Secondary	8.6	54.2	95.2	29.4	93.3	97.4	16.6	63.8	93.7
Combined	-6.9	29.9	91.6	-4.1	26.2	92.3	-7.2	27.1	91.0
Rural	1.1	37.4	93.2	27.9	83.6	99.6	24.2	74.6	99.6
Suburban	-10.7	34.4	89.7	-2.9	34.9	91.8	-5.5	34.3	91.8
Urban	10.6	44.9	95.5	61.6	177.4	99.8	56.7	161.4	99.8

Table 5 -- % Relative Error, % relative Mean Square Error and % coverage rates for the Bootstrap and BHR variance estimator for Students/Teacher Ratio estimates by Affiliation, Region, Level and Urbanicity

Estimate	Bootstrap			BHR without FPC Adjustment			BHR with FPC Adjustment		
	Rel. Error	Rel. MSE	Cov. Rate	Rel. Error	Rel. MSE	Cov. Rate	Rel. Error	Rel. MSE	Cov. Rate
Other Affil.	-0.3	31.9	94.2	12.3	46.0	96.1	9.1	40.6	94.7
Northeast	-4.3	56.3	91.6	9.3	70.4	94.7	6.7	66.4	94.7
Midwest	-5.5	67.1	91.2	3.1	43.9	95.4	0.5	41.6	94.1
South	6.7	45.2	95.4	3.2	31.5	95.9	-0.8	28.6	93.4
West	-1.0	38.3	93.9	9.8	46.9	97.3	7.1	42.8	97.3
Elementary	-2.5	43.9	93.2	11.8	52.2	99.5	9.5	48.5	99.5
Secondary	-25.3	49.1	81.1	1.2	33.2	94.1	-9.3	32.5	91.4
Combined	9.7	46.2	95.8	16.3	53.0	95.9	12.6	45.7	95.7
Rural	7.0	59.8	95.3	22.7	84.2	99.9	18.9	75.6	99.8
Suburban	1.1	37.6	93.5	18.0	61.6	97.1	14.8	55.3	97.1
Urban	5.4	45.6	94.7	15.0	58.3	93.8	11.6	52.1	93.7

SAMPLE DESIGN ISSUES FOR THE BASE YEAR OF A LONGITUDINAL SURVEY OF KINDERGARTEN CHILDREN

John Burke, Thanh Lê, John Michael Brick, Westat Inc.,
Thanh Lê, Westat Inc., 1650 Research Boulevard, Rockville, Maryland 20850

Key Words: Longitudinal Survey, Multi-Stage Sampling, Oversampling, Coverage, Precision of Estimates

1. Introduction

The Early Childhood Longitudinal Study: Kindergarten Class of 1998-99 (ECLS-K) is sponsored by the U.S. Department of Education, National Center for Education Statistics (NCES). It will provide national data on children's characteristics as they progress from kindergarten through the fifth grade. It will also provide information on key analytical issues such as school readiness; transition to kindergarten and subsequent grades; kindergarten and first grade student performance; and, cognitive growth and student progress.

The ECLS-K will collect data on a nationally representative sample of approximately 20,000 children enrolled in about 1,000 kindergarten programs beginning with the 1998-99 school year. During this school year there will be two data collections, one at the beginning (fall) and one near the end (spring). Thereafter, most follow-up studies will be conducted in the spring, beginning with spring 2000. In the fall of 1999, data will be collected on a 25 percent subsample of first graders that will provide information to assess children's growth over the summer vacations. Data collection will consist of direct assessments of the students themselves, interviews with their parents, as well as abstracts of school records. Teachers and school administrators will complete self-administered questionnaires. In the base year, the sample of children is selected using a multi-stage probability design. The first-stage or primary sampling units (PSUs) are geographic areas that are counties or groups of counties. In the second stage, samples of public and private schools with kindergarten programs are selected within the sampled PSUs. Both PSUs and schools are selected with probability proportional to measures of size that take into account the desired oversampling of Asians and Pacific Islanders (APIs). The third stage sampling units are children of kindergarten age, selected within each sampled school.

In this paper, we discuss the evaluation of alternative designs for sampling PSUs, the method of sampling within PSUs, and the sampling of students within schools. The school sampling frames are also described, as well as procedures adopted to improve the school coverage, separately for public, Catholic and non-Catholic private schools. Other features of the design, such as procedures used to include students in the follow-up collections, are not discussed here.

2. Sampling PSUs for the ECLS-K

2.1 Issues Under Consideration

In the base year, the design for the ECLS-K involves a clustered sample of PSUs that are counties or groups of counties; in the second stage, about 800 public and 200 private schools are selected from the sampled PSUs; the final stage is the selection of a fixed number of about 24 students from each sampled school. In subsequent years, students will be followed as they move to first grade and beyond, with subsampling of students that will largely be determined by probabilities that are a function of how many sampled students move into the same school so that the cost of data collection can be contained. The number of study schools in the subsequent years is expected to be substantially larger than the number in the base year because of this migration. The clustered design is necessary to limit the costs of data collection that are highly related to the dispersion of the children.

The primary focus of the analysis of the ECLS-K data will be at the student level, as indicated by the issues of interest such as school readiness and transition to kindergarten and subsequent grades. The optimal sample design for student level estimates is to sample students with probabilities that are approximately the same for each student. In most designs, this is achieved by sampling PSUs and schools with probabilities proportional to the number of students and selecting a fixed number of students per school. An equal probability student sample in the subsequent years would also be optimal if the data collection costs were roughly equal, but unequal probabilities may be necessary to account for the cost efficiency associated with sampling students clustered in the same first grade school.

On the other hand, school level estimates in the base year are more efficient if the schools have equal probabilities of selection, irrespective of the number of students in the school. A compromise scheme that is very useful when both school and student estimates are of equal importance is to select schools with probabilities proportional to the square root of size (this is between equal probability and probability proportional to the number of students). However, if this procedure is followed and an epsem sample of students is to be achieved, students have to be subsampled at rates that equalize their probabilities of selection. This results in a different number of students being sampled per school. In the ECLS-K, a fixed student sample size per school is highly desirable for burden and cost reasons, so that this option is not viable.

This brief review of the relationship between the sampling probabilities and the analysis issues for the ECLS-K is intended to provide some background for subsequent discussions of the alternatives for sampling PSUs, schools, and students. Two design options, A and B, are considered below.

2.2 Two Alternative Designs

Option A involves using the existing Private School Survey[1] (PSS) first stage sample of PSUs. In PSS, about 124 PSUs are sampled with probabilities proportional to the square root of total population. Let f be the overall student-sampling fraction in the ECLS-K. Since the PSU selection probability P_1 is known from the PSS (PSUs are already selected), and the conditional probability of sampling students within a sampled school P_3 is fixed (24 per school), the conditional probability of selection of a school in a PSU (P_2) can be calculated as $P_2 = f/(P_1 P_3)$. It is easy to show that the number of schools sampled per PSU will vary if the PSS PSUs are used in order to obtain an equal probability selection of students. Also, the overall probability of selecting a school under this option is the product of P_1 and P_2 and this is not an equal (or a square root of size) probability sample of schools.

Option B involves selecting a new sample of 100 PSUs, with probabilities proportional to the number of kindergarten students (or the count of five-year-old children in a PSU, which is a close surrogate for this). The second stage sampling of schools is proportional to the number of kindergarten children in the school and, in the third stage, 24 students are sampled per school. The number of schools sampled per PSU should be approximately constant (on average 8 public and 2 private schools would be sampled per PSU if 100 PSUs and 1,000 schools are sampled). The overall probability of selection for a school is close to proportional to the number of students in the school.

We evaluate the two options by looking at the student coverage and the precision of the estimates. These are two factors that are most likely to have different characteristics under the two options. There are also cost differences that are discussed at the end of this section.

2.2.1 Coverage

Coverage of all students in the base year sample is of great importance to the ECLS-K. In a longitudinal survey, biases in the base year are often carried throughout the multiple years of the study. The most important coverage concern in the ECLS-K is the coverage of students in private schools. Private schools have much greater rates of both openings and closings than do public schools so that using old sampling frames can lead to coverage bias.

Under both options, the most current Common Core of Data[2] (CCD) sampling frame is used for the public school sector and the most recent PSS list frame is used for the private school sector. These sampling frames can be partitioned into the sampled PSUs so that samples of schools within each PSU can be selected.

Data from previous PSS samples indicate that updating the area frame increases the estimated number of all private schools by about 8 percent and the estimated number of all private school students by about 3 percent. For kindergarten, the percentage increases are probably greater. If nothing is done about the coverage, the ECLS-K could exclude up to 5 to 7 percent of children in private kindergarten programs and about 1 percent of all children in kindergarten (5 to 7 percent of about 15 percent of children in private kindergarten). While the overall loss is small, the loss within the private school sector is too large to ignore.

Option A was thought to be better because the PSS is an on-going survey that includes an area sample to improve coverage of private schools and students, and because of its potential integration with the ECLS-K. The search for new private schools is conducted every two years, so the frame is more complete within these PSUs. However, the private school sampling frame in the PSS is outdated because no fieldwork has been done since 1995. The search for new schools in PSS for 1997-98 was not completed in time for ECLS-K school sampling. As a result, the same work would be required for both options, and there would be no coverage advantage if the PSS PSUs were used.

2.2.2 Precision of the Estimates

The precision of the estimates is affected by the sample design in various ways. For the ECLS-K, the two main factors that cause losses in precision relative to a simple random sample design are the clustering of schools and students within the sampled PSUs and the

[1] The PSS, conducted by the Bureau of the Census for NCES, is designed to build a universe of private schools in the United States. The main component of the PSS is the list frame. Data sources for building the list frame are commercial lists, state lists and private school association lists. An area frame is included to identify schools overlooked in the list frame. This area search for additional schools is conducted in randomly selected counties. For more details, see NCES (1998) and McMillen (1993).

[2] The CCD is the NCES database of elementary and secondary public schools in the United States and its territories. It collects data on schools and state and local school districts (or education agencies), mostly from administrative records. The database contains information on schools, school districts, students and staff, as well as fiscal data. For more details, see McMillen (1993).

variability in the sampling rates or weights of the units. These two issues are discussed below using estimates of the number of sampled schools and students.

In almost all clustered samples, the precision of the estimates is reduced relative to simple random sampling because units within the same cluster tend to be more homogeneous than units across the entire population. This will be true in the ECLS-K since the clusters are geographic areas, and schools and students within the same geographic area are almost always more homogeneous. In most multi-stage samples the effect of clustering (on the variance of the estimate) at the PSU level is approximated well by the expression (Kish, 1965)

$$D_1 = 1 + \rho(b-1)$$

where ρ is the intra-class correlation coefficient indicating the degree of homogeneity within the PSU and *b* is the average sample size in the cluster (in this case, the number of sampled students).

This formulation breaks down if the average sample size per cluster is not constant. This is exactly what occurs in Option A. In this case, a better approximation (Holt, 1980) is given by

$$D_2 = 1 + \rho(b'-1)$$

where $b' = \frac{\sum b_i^2}{\sum b_i}$. Note that if the sample size per cluster is a constant across PSUs, then the two expressions are equal.

Because of the variation in the cluster sample sizes, we used D_2 to compute the effect of varying the student cluster sample size for both options. The ratio $R = D_2 / D_1$ is the expected increase in the variance of public school student level estimates. The results are shown in Table 1 for the two options and different values of ρ. The variability in the Option A student sample sizes by PSU results in a substantial increase in the variance of estimates of public school students. This increase may be understated because the comparison assumes equal values of ρ under the two designs while under Option A the PSUs are smaller and likely to have larger values of ρ. Similarly results apply to private schools.

Table 1. D_2 and $R = D_2 / D_1$ for characteristics of students from public schools

ρ	Option A		Option B	
	D_2	$R = D_2/D_1$	D_2	$R = D_2/D_1$
.01	3.13	1.23	2.84	1.03
.02	5.25	1.28	4.68	1.03
.03	7.38	1.31	6.52	1.04
.05	11.63	1.33	10.20	1.04
.10	22.26	1.35	19.41	1.04

The discussion of the increase in the variance for student level estimates under both options does not recognize the effect of differential weights on the estimates. This is appropriate because both options have approximately self-weighting samples of students. The same approach is not appropriate for school level estimates because neither option results in a self-weighting sample of schools. The effect of weights on the school sample was computed for both options using the following formula (Kish, 1976)

$$D_W = 1 + L = \left(\sum U_i k_i\right)\left(\sum U_i / k_i\right)$$

where U_i is the size of unit *i* and k_i is the base weight of unit *i*.

Using this formula, the effect due to weighting for school level estimates D_W is 2.17 for Option A and 2.63 for Option B for characteristics of public schools. This factor must then be multiplied by the appropriate D_2^* factor (computed using the number of schools sampled per PSU rather then the number of students per PSU) for each of these options to get the combined effect of the design on the variance of public school estimates. The results of this multiplication, which include the effect associated with the different numbers of PSUs in the two designs, are shown in Table 2 for the two options and different values of ρ.

Table 2. D_2^* and $D_s = D_2^* D_W$ for characteristics of public schools

ρ	Option A		Option B	
	D_2^*	$D_s = D_2^* D_W$	D_2^*	$D_s = D_2^* D_W$
.01	1.08	2.34	1.07	2.81
.02	1.16	2.51	1.13	2.98
.03	1.24	2.68	1.20	3.16
.05	1.40	3.03	1.34	3.51
.10	1.79	3.88	1.67	4.40

The increase in the variance is less under Option A than Option B for school level estimates. As mentioned before, the value for ρ is probably smaller for Option B than for Option A. Because student

estimates are more important to the study objectives, Option B was used for ECLS-K.

2.3 The ECLS-K PSU Sample

Following the parameters of the Option B design, the ECLS-K sample consists of 100 PSUs which are counties or groups of counties. The distribution of five-year-olds based on 1994 population estimates by race/ethnicity was used to form PSUs with a minimum size of 320 five-year-olds and to construct a measure of size that took into account the oversampling of API children. The PSUs were stratified into self-representing and non-self-representing. There are 24 self-representing PSUs. For the non-self-representing PSUs, 38 strata of roughly equal measure of size were created, and two PSUs were selected in each stratum, yielding 76 non-self-representing PSUs. The variables used for stratifying the non-self-representing PSUs were MSA/non MSA status, and region. In the next level of stratification, size class, race/ethnicity (high concentration of API, Black or Hispanic) and per capita income were used for MSAs, and race/ethnicity and per capita income were used for non-MSAs.

The measure of size used for selecting PSUs takes into account the oversampling of APIs. The weighted measure of size is calculated as $2.5 \times n_{API} + n_{other}$, where 2.5 is the oversampling rate for APIs, n_{API} and n_{other} are the counts of five-year-old APIs, and all others, respectively.

3. Sampling Within PSUs

3.1 School Sampling Frames

In the second sampling stage, public and private schools offering kindergarten programs were selected. The target number of schools was set at 800 public and 200 private schools from within the ECLS-K sampled PSUs. The number of schools selected is the target number of schools adjusted upward by an expected school response and eligibility rate. In total, 934 public schools and 346 private schools were selected with probability proportional to the measure of size described below.

The school frame for the ECLS-K was built using several data sources: the 1995-96 CCD, the 1995-96 PSS and the 1996 lists of schools run by the Bureau of Indian Affairs and the Department of Defense. Data from the 1997-98 PSS list frame and the Quality of Education school and district files were used to update school location information. The constructed ECLS-K school frame included 18,891 public schools and 12,412 private schools with kindergarten programs within the sampled PSUs. The school frame was augmented in the spring of 1998 to include schools that are operational but were not included in the frame, as discussed in Section 4.

3.2 School Measure of Size

Schools were selected with probability proportional to size. The measure of size was constructed taking into account the oversampling of APIs, separately for public and private schools. The measure of size for school *j* in PSU *i* is calculated as

$$SCHMOS_{ij} = 2.5 \times n_{API,ij} + n_{other,ij}$$

where 2.5 is the oversampling rate for APIs, $n_{API,ij}$ and $n_{other,ij}$ are the counts of API kindergarten students, and all other kindergarten students, respectively, in school *j* of PSU *i*.

3.3 Clustering of Schools

Schools with fewer than 24 students (public) or 12 students (private) were clustered together within PSUs in order to obtain a sample that is closer to self-weighting. For example, if a school with 12 students was not clustered the students from that school would be sampled at about half the probability as students in larger schools. The goal was to group small numbers of schools to form heterogeneous clusters with an aggregate number of students as close to 24 as possible. This goal was set so that if a cluster was selected we would not need to recruit many small schools; furthermore, the heterogeneity of schools improves the reliability of the estimates. We defined heterogeneity for public schools by school size and for private schools by religious affiliation and school size.

3.4 Stratification of Schools

The schools were stratified implicitly within each PSU. For public schools, (clusters of) schools were sorted by the measure of size and separated into three size classes of roughly equal size (high, medium, and low). Within each size class, they were sorted by the proportion of APIs in a serpentine manner. In private schools, each cluster was identified as religious, mixed, or non-religious. The list of clusters was then sorted by these three categories. Within each category, clusters were sorted in a serpentine manner by the measure of size.

3.5 Sampling Students

In the third stage, 24 students will be selected for the base-year study in each school (or fewer when the school does not have 24 students), with oversampling of API students. For the ECLS-K, PSUs and schools were

sampled assuming that API students would be oversampled by a factor of 2.5. However, in about 40 percent of the school sample, it may not be possible to select a total of 24 students while oversampling API by a factor of 2.5. We determined that the oversampling factor would have to be as high as 5.5 in order to meet the target. Increasing the oversampling factor would have the unfavorable effect of increasing the variability of the weights and increasing variances. Therefore, we have chosen to oversample the API students by a factor of 3. In subsequent years, API students may be followed at a higher rate as they transfer to new schools to reduce the attrition in this domain.

4. Improving Coverage of Schools

The sampling frames used for the main sampling of schools offering kindergarten programs were augmented to include newly opened schools that were not included in the frame. Some schools that were in the CCD and PSS but not included in the ECLS-K frame for various reasons were also included in this process. Procedures for augmenting the frames were different for public schools, Catholic schools and non-Catholic private schools. Each is discussed below.

4.1 Public Schools

The sample of 934 public schools falls in 535 school districts in 41 states and the District of Columbia. The sampled districts were asked if any school expected to offer a kindergarten program in 1998 or any ungraded school was missing from a list sent to them (developed from the original frame). Districts that were in the sampled PSUs but were not reported in the CCD as having any schools with kindergarten programs were also contacted. If they had any schools that would offer a kindergarten or ungraded program in fall 1998, information on these schools was collected. The information obtained from the school districts was checked against the ECLS-K public school frame to confirm that they truly were new or newly eligible. Checking was not restricted to within the school district but was done within state in order to ensure that each "new" school was not already listed under a different district and that it was new in this district due to district reorganization. Through this process, 252 new public schools were identified. A sample of 19 schools was selected. Since a district identifies a new school, each school was selected with a probability conditioned on the within stratum probability of selecting that district.

4.2 Catholic Schools

There are 117 Catholic schools in the ECLS-K sample in 59 dioceses. The procedure for contacting the dioceses and for obtaining new school information was exactly the same as for public schools. Since a diocese can cover more than one city or county and can sometimes cut across state, checking included an additional step of placing each school on the list sent by the diocese in the correct county and hence the correct PSU, before checking for new schools. A total of 117 new Catholic schools were identified, and 6 were sampled. As for public schools, the new school selection probability is conditioned on the within stratum probability of selecting the diocese that identifies the new school.

4.3 Non-Catholic Private Schools

The main source used to search for non-Catholic private schools was the telephone book Yellow Pages. In addition, local education agencies (LEAs) and local government offices were also contacted for information on non-Catholic private schools in their areas, but this was only implemented in 22 PSUs with large PSU weight (greater than 20).

For all the counties in the ECLS-K sample, electronic Yellow Pages listings of elementary schools, private and parochial schools, special education schools, preschools, nurseries and kindergartens were created. The procedures involved matching these listings in the sampled PSUs to various ECLS-K data files in order to purge, to the extent possible, schools that were already in the ECLS-K frames. Schools that were on the PSS file but were out-of-scope or did not contain any kindergarten children according to the PSS were also included. The files were matched and school names of matches and near-matches were examined in order to decide on true matches. Non-matches were put through a screening of school names using keywords to exclude any that had 'high school', etc. in the name. The private school frame constructed using the Yellow Pages had 11,405 schools in the sampled PSUs. A sample of 279 schools was selected and then screened for eligibility. Of these, only 85 schools reported that they were private, would be open in fall 1998 and would have kindergarten or kindergarten-age students. These schools were added to the main sample.

The 22 PSUs with largest PSU weights cover 53 counties. In these counties, 135 LEAs and 218 cities/towns were identified. In each city/town, a list of local government offices was compiled using the electronic Blue Pages. The telephone interviewers contacted all LEAs. However, in cities/towns with multiple government offices, they contacted more than one only if the first call did not yield any information on private schools. Of the 135 LEAs, only 54 had information on private schools in their area. After the information collected was unduplicated against PSS and the Yellow Pages, 30 new private schools were identified. Of the 218 cities/towns, only 75 yielded information on private schools. After the information was unduplicated against PSS, LEAs and the Yellow Pages, 19 new private schools were identified. In

addition to these procedures, three new private schools were reported by the field staff working in the area. Of the 52 new schools identified, 24 were sampled and screened to ensure that they are eligible for the ECLS-K.

In total, 134 new schools (public, Catholic and non-Catholic private) were added to the original sample of 1280 schools. Since the grade span of some of these schools was not known at the time of sampling, particularly the non-Catholic private schools, the actual number of productive new schools added will be smaller.

5. Conclusion

In this paper, two alternatives for sampling PSUs for the ECLS-K were examined. A new sample of PSUs with a measure of size appropriate for the study was found to provide much greater reliability than using a sample with a less efficient measure of size. The sample of schools was selected from a school frame that is somewhat dated, and then augmented to improve coverage, particularly important in the case of private schools.

6. References

Holt, D. (1980). Discussion of the paper 'Sample designs and sampling errors for the World Fertility Surveys' by Verma, Scott and O'Muircheartaigh. *Journal of the Royal Statistical Society* (Series A), 143, Part 4, p. 468.

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

Kish, L. (1976). Optima and proxima in linear sample designs. *Journal of the Royal Statistical Society* (Series A), 139, Part 1, p. 80.

McMillen, M., Kasprzyk, D., and Planchon P. (1993). Sampling frames at the United States National Center for Education Statistics. In *Proceedings of the American Statistical Association Conference on Establishment Surveys*, pp. 237-243.

National Center for Education Statistics (1998). *Private School Universe Survey, 1995-96.* NCES 98-229. Washington, DC: National Center for Education Statistics.

DEVELOPMENT OF A PROTOTYPE WAREHOUSE SYSTEM FOR SURVEY DATA

Stephen R. Wenck, Sameena M. Salvucci, James K. Tyson, Synectics for Management Decisions,
Steven Kaufman, National Center for Education Statistics
Stephen R. Wenck, Synectics, 1901 North Moore St., Suite 900, Arlington, VA 22209

Key Words: Data Warehouse, Linked Data, Dimensional Model, Star Schema

Introduction

In the past 20 years, there has been a significant change in the way data are accessed and used. The procedural tools of the seventies have given way to graphical user interfaces (GUIs) which allow direct manipulation of data. This evolution has created an opportunity to place powerful tools for information retrieval and manipulation in the hands of users.

The National Center for Education (NCES) has been a catalyst in the development and innovative use of technology, including the addition of user tools with its survey data. It fueled advances such as the electronic code book (ECB), the data analysis system (DAS), and other data products including the Common Core of Data (CCD) and the Integrated Postsecondary Education System (IPEDS) CD-ROMs with graphical user interfaces. These were not just technical advances; they provided a paradigm shift from merely providing information to placing user-friendly tools along with the data directly in the hands of users in order to improve the accessibility and usefulness of the information.

Recently, NCES began providing its survey data and tools through the Internet, a step which has greatly expanded accessibility to these data. However, both NCES staff and its customers have a further need for the ability to directly access "linked" data from multiple surveys, multiple components within a survey, and multiple time points, for methodological and analytic purposes.

The purpose of the project was to advance NCES' capabilities to support the dissemination of linked data[1]. This enhanced capability acknowledges the critical role of providing access to linked historical data in improving analytical capabilities, improving sampling and estimation techniques, ensuring data quality, and improving customer service. An essential element of this study is the development of a software prototype system to facilitate NCES data customers' access to linked historical data (i.e., a data warehouse).

The basic concept of an NCES data warehouse is to facilitate data management such that specific information is easily accessible to all users of NCES data.

The major expected benefits to NCES of such a data warehouse system are:

- Elimination of the need to re-create links every time a research purpose requires it, thereby substantially reducing the effort involved in ad-hoc survey linkages;
- Wider dissemination and use of its survey data through more user friendly access to multiple-linked data products;
- Establishment of "standards" for documentation of data and metadata related to future releases of NCES data.

Research Towards Developing a Prototype

The research spanned three basic areas that relate to NCES' interest in data warehouses:

- Identifying and comparing commercial off-the-shelf products and technologies;
- Reviewing activities by other organizations with similar mission and scope; and
- Reviewing white papers, articles and textbooks.

The activities and findings in each of the areas of primary research undertaken are briefly summarized below.

Commercial-off-the-shelf (COTS) products and technologies

Investigation of COTS products focused on 1) characterizing the broad range of vendors and products

[1] "Data linking" and "linked data" refer to the concept of associating data from *across* various survey data sets through common or related elements. This may include linking heterogeneous data sets (such as via common data elements), as well as linking homogeneous data sets (such as to aggregate results across different survey years), or both.

that are positioned to support the general goals of this project, and 2) evaluating selected products in somewhat closer detail towards the development of a data linking prototype.

The current market for COTS software products that support data warehousing is at a stage of rapid growth. Many competing products from numerous vendors characterize the current market. One Internet-based list of data warehousing products (maintained by Larry Greenfield, LGI Systems at http://pwp.starnetinc.com/larryg/index.html) recently identified 822 products in 17 categories.

Two categories of products, Report and Query tools, and OLAP tools, were determined to be the most appropriate types of tools to use to demonstrate the capabilities of linked survey data.

Report and query tools produce tabular reports with simple summations and aggregations, typically (but not necessarily) based on the contents of a relational database. Examples of popular products with this capability include Cognos Impromptu, Microsoft Access, Brio BrioQuery, Seagate Software Crystal Reports, and the SAS System.

On line analytical processing (OLAP) is a popularly used to describe an interactive approach to decision support[2]. OLAP query tools produce reports with more complex processing requirements, and typically work against (star-schema[3]) relational databases and/or multidimensional databases (such as Red Brick or Essbase). Examples of popular products with this capability include Cognos PowerPlay, Brio BrioQuery, IQ Software IQ/Vision, Seagate Software Crystal Info., and the SAS System.

Similar activities by other government organizations

The activities of several Federal government agencies with missions similar to NCES were reviewed, including the National Agriculture Statistics Service, the Environmental Protection Agency, the Bureau of Labor Statistics, the Census Bureau, and the National Science Foundation.

General findings show that these agencies are taking or have taken an approach to linking statistical data similar to NCES' current approach, that is to plan and implement various data warehousing technologies to support the efficient organization and dissemination of data and metadata.

The National Agricultural Statistics Service (NASS) of the Department of Agriculture has developed a relational database system employing a star schema model. In particular, the Department of Agriculture has successfully developed a system that links various survey data sets (including time series data) using a multidimensional or star-schema model. The Department of Agriculture staff who developed this system emphasized the practical value of the star-schema data model as their starting point, particularly as the use of a common model allows them flexibility in the choice of COTS products that can support or work with their data. By design, the NASS warehouse is not available on the web. Their warehouse contains restricted data and is only available on the NASS LAN.

EPA has developed an on-line (web based) relational database called Envirofacts. This database integrates data extracted monthly from five facility (site) based EPA program systems. Envirofacts also contains a grant information database, three integrating databases, and mapping applications. Information in the Envirofacts database can be freely accessed through the use of predetermined or user-developed queries. While the complexity of user-developed queries is unlimited, queries that return a large volume of data may terminate prematurely due to system limitations. The Envirofacts Query allows users to retrieve the environmental profile of facilities that match the query specifications. The Envirofacts database contains only data available under the Freedom of Information Act and therefore full access is granted to all users.

The Bureau of Labor Statistics (BLS), Department of Labor has developed a web-based data warehouse. The data are not linked across different surveys. The system allows extraction of timeseries data for a specific database. The only output selection is raw ASCII data. Within each database one can select any one or all of the timeseries shown and choose desired date ranges and output options.

The Census Bureau has developed the "Data Extraction System" (DES) as a web-based data warehouse. The

[2] Decision support activities involve producing reports and views of aggregated data, such as cross tabulations and various statistical measures to support inferential decision making.

[3] Star Schema is a name that database designers have used to describe dimensional models because the diagram of this type of model looks like a star, with one large central table and a set of smaller attendant tables displayed in a radial pattern around the central table.

Data Extraction System is a tool for extracting records (rows) and fields (columns) from very large, public-information, data files (for example: survey and census records.) The system produces custom extracts in selectable data file formats which, can then be picked up at the Census Bureau's FTP site. This system does not produce tabulations, it produces only raw data. No confidential data are available via this service. As with the BLS system, data are not linked across databases, but rather allow for extraction of data from a particular data source.

The National Science Foundation has designed WebCASPAR, a database system to provide quick and convenient access to a wide range of statistical data focusing on U.S. universities and colleges and their science and engineering resources. WebCASPAR users can specify the statistical data of interest and either view the data through their web browsers or transfer the data to their own computers as spreadsheets or other data files.

Review of Industry Literature

A common theme of the literature was the importance and value of developing appropriate data models to represent the information that is to be linked or warehoused. While there was some debate over the relative benefits of alternative technologies for implementing these models, there was broad and strong consensus that the so-called dimensional or star-schema data model best represents data and metadata[4] in forms useful to analysts.

It should be noted that the multidimensional or star-schema model is a *conceptual* model, and theoretically can be implemented through any of a variety of *physical* designs. For example, there is an ongoing debate over the relative merits of general-purpose relational database management systems (RDBMS) versus special-purpose "multidimensional database systems" (MDDSs) towards supporting star-schema warehouses. Such discussions are important for developing an architecture to support production operations but, for purposes of this prototype, it was determined that a standard RDBMS already supported at NCES would be adequate to demonstrate the feasibility and utility of linking NCES data sets through a star-schema model.

[4] Metadata is a general term for data about data.

Prototype Warehouse: Metadata and the Data Model

The objective of this prototype warehouse is to provide a highly interactive *ad hoc* analysis system with the ability to access data from multiple NCES surveys simultaneously. Users expect to view this data from different perspectives—enrollment by race/ethnicity by year, number of full-time equivalent teachers by urbanicity of school by region, etc.—and expect to switch interactively among these perspectives. Users need to see information at different levels of detail, looking for insights with summary data (e.g., enrollment by race/ethnicity by year), then "drilling down" to increasing levels of detail (e.g., enrollment by race/ethnicity by school district by year), in order to understand root causes and anomalies.

Metadata

Metadata is data about data. There are two categories of metadata: technical and warehouse-dependent. Technical metadata is the description of the data needed by various tools to store, manipulate, or move data. These tools include relational databases, application development tools, and OLAP tools just to name a few. Warehouse-dependent metadata, on the other hand, is the description of the data needed by users to understand the context and meaning of the data. It is any information that makes the data more useable by the researchers.

Building data warehouses is resource-intensive, so it is important to make the results of these projects usable by researchers. Implementing full warehouse-dependent metadata produces the significant benefit of making the information visible, understandable, and available. In short, it can be the difference between success and failure of a warehouse effort.

Data Model

A data model provides an *abstract* view of the data, including a description of *what* data are actually to be stored in a database, and the relationships that exist among the data. In particular, the dimensional or star-schema data model is a logical design technique that seeks to present the data in a standard, intuitive framework that allows for high-performance access in a data warehouse.

A star-schema model was developed for the selected longitudinal CCD LEA data. The star schema represents to the end user a simple and query-centric view of the data by partitioning the data and the warehouse-

dependent metadata into two types of tables: fact tables and dimension tables.

In the NCES prototype, fact tables represent the actual data from the multiple CCD data sets. The most useful facts in the fact tables are numeric and additive. Additivity is crucial because data warehouse application users almost never retrieve a single fact table record; rather, they fetch back hundreds and thousands of these records at a time, and the only useful thing to do with so many records is to add them up and present the information in an aggregate form. Information in the dimensions is what describes the facts.

Dimension tables, by contrast, most often contain descriptive textual information including the warehouse-dependent metadata. Dimension attributes are used as the source of most of the interesting constraints in data warehouse queries, and they are virtually always the source of the row and column headers in the users answer set. The power of the warehouse database is proportional to the quality and depth of the dimension tables.

A fact table has columns of anonymous index keys, one for each dimension, that relate back to exactly one row in each of the dimension tables. The unique combination of keys in the fact table acts as a cross-reference specifying the intersection of each dimension at that data point.

The creation of unique dimension table keys is very important. These keys have to be unique to identify each record in the dimension table, but they also have to be generic enough so that as other surveys are added, a completely new key does not have to be added.

Because the single key in a dimension table is a row, metadata about that key can exist on that row and be available for querying. In fact, the dimensional intersection specified by the unique combination of foreign keys in the fact table defines the data point by presenting the metadata and the data item together.

Prototype Warehouse: Implementation

The approach on the implementation of the data-linking prototype was to use standard methods and tools that facilitate end user access to linked or linkable NCES data. Concentrating on standard methods and tools ensures that the successful elements of the data-linking prototype can readily be integrated or incorporated within broader NCES or Department of Education plans, standards and technologies.

The prototype development process included the following steps:

1. Development of a flexible data model for the target data using an industry standard approach (i.e., the "dimensional" or "star schema" model). Dimensional modeling is a technique for visualizing the data as a "cube" of three, four, or even five or more dimensions. This can be illustrated with a simple example using the CCD:

   "States report data every year on district enrollment and locale type"

   For the data warehouse design, special emphasis is added as follows:

   "*States* report data *every year* on enrollment and *locale type*"

   Most people find it easy to think of this as a **cube** of data, with labels on each of the edges of the cube. For the description above, the edges of the cube can be labeled as *State, Time, and Locale Type*. The points inside the cube are where the enrollment measurements for that combination of State, Time, and Locale Type are stored. This is the dimensional model. Star Schema is a name that database designers have used to describe dimensional models because the diagram of this type of model looks like a star, with one large central table and a set of smaller attendant tables displayed in a radial pattern around the central table. There is one large dominant table in the center of the schema. It is the only table in the schema with multiple joins (relationships) connecting it to other tables. The other tables all have only a single join attaching them to the central table.

2. Implementation of the data model using Oracle, a standard relational database management system (RDBMS). A database management system consists of a collection of interrelated data and a set of programs to access that data. The collection of data is usually referred to as the database. A standard language for accessing relational databases is the standard query language (SQL).

3. Population of the model with data and metadata from multiple years of the CCD School District Universe Survey.

4. Demonstration of the model's utility with a commercial "front end" query tool called Cognos.

The conceptual model for the prototype system consists of components or processes that, taken together, meet the specific goals of the project. For each component or process, several alternative commercial products or methods may be available. The benefit of the standards-based approach is that any other component or process that supports the same common standards may replace any or all of the components or processes.

The following provides a description of each of these components and processes. These descriptions are of necessity somewhat technical.

1. A work process extracts data from relevant sources (i.e., NCES survey data sets) and loads it into a repository that is based on a standard "star-schema" or "dimensional" (the terms are synonymous) data model. This process can be accomplished via a wide variety of tools, methods, and technologies.

2. The star-schema repository is implemented under a standard SQL accessible RDBMS. SQL is a formal standard, controlled by the American National Standards Institute and endorsed via the National Institute for Standards and Technology of the Department of Commerce as a Federal Information Processing Standard. Examples of SQL RDBMS include Oracle7, Microsoft SQL Server, Sybase, Informix, and DB2. A star-schema repository can be quickly and easily transferred from one SQL RDBMS to any other. There are no practical, standards-based alternatives to SQL RDBMS for database management. There are certain data management products that are optimized to support star-schema data "warehouses" (e.g., Red Brick, Essbase). These products purportedly offer better performance than standard SQL RDBMSs in support of analytical queries against a star- schema database. The use of such products was beyond the scope of this task. In any event, the star schema model is equally as portable to this class of product, so no flexibility was sacrificed.

3. Data that are implemented as a star-schema under a SQL RDBMS can be accessed via a wide variety of end-user analytical tools (e.g. Cognos, Brio, SAS, Excel, etc.). These tools generate SQL queries that are transmitted to the RDBMS over a network. The RDBMS executes the SQL, and returns the results (data) to the end-user tool. These tools use one (or both) of two methods for communicating with the RDBMS. One method is through a mature, reliable *de jure* standard technology known as Open Data Base Connectivity (ODBC). ODBC allows any client tool that supports ODBC to communicate with any RDBMS that supports ODBC. ODBC is currently supported by every leading RDBMS product (including all of those identified above), as well as by every leading vendor of data warehousing analytical tools. (The second method is to use "native" data access drivers, i.e., to communicate directly with the RDBMS (using SQL), to avoid the processing overhead of ODBC translations. "Native" data access generally offers better performance than ODBC data access, however these differences are largely inconsequential for data access for analytical purposes). The goal was to identify leading products that can best illustrate the feasibility and benefits of linking survey data through a multidimensional or star-schema data warehouse model.

In summary, the data linking prototype may be considered as consisting of a set of components, where each component can be replaced with any other of the many that support the same basic technical standards. This mainstream standards-based approach was chosen so that the work products of this project will have the broadest utility and applicability toward larger NCES and Department of Education plans and goals.

ORACLE Warehouse

The first step in converting the SAS datasets into the ORACLE data warehouse was to create a crosswalk of the variable names and format types across all years of the CCD datasets. This crosswalk allowed us to change variable names and types so that for every year of data the same information had the same variable name and same data type.

Related to this standardization process was the standardization of the record layout. The variables on the datasets were in different order, depending on year. An identical record layout was made for each year to reduce the burden necessary when reading the ASCII files into ORACLE.

The percent of students in each ethnic group (white, black, Asian, Hispanic, Native American) for each district had been calculated and added in the CCD data sets. Summing up percents across districts would not make much sense, so these percents had to be backed out into raw numbers. This process also had to be performed on the Percent of Children in Poverty. The front-end tool will allow for the calculation of all percents. For a warehouse, it is important that all values be numerical.

The final step in SAS, prior to converting the datasets to ASCII, was to standardize some values in particular variables. For numeric variables such as the ethnicity variables, values of "M," "N," or "**" have been converted to null. This is the NCES way of indicating that data are Missing or N/A.

At this point data were converted from SAS data sets into ASCII files. These ASCII files were read into Oracle as a rectangular database. The rectangular database served as the starting point for the fact table. We then split off the dimension tables from the fact table in Oracle.

Cognos Tools

After a search of a number of commercial off-the-shelf vendors of OLAP tools, the Cognos suite of tools was chosen. Cognos was selected for its quality product, as well as their willingness to provide unlimited technical support during the creation of a prototype warehouse.

The two main tools of the Cognos suite are PowerPlay and Impromptu. The main tool, PowerPlay, has two views - Explorer and Reporter. In PowerPlay, a three-dimensional view of the data, called a PowerCube, is generated from a subset of the data available on the warehouse.

In PowerPlay Explorer, the three-dimensional view of the data, or PowerCube, can be manipulated to display any of the included categorical data as rows, columns, layers (for 3 or more dimensional tables), or filters. Any one of the continuous variables included in the 3-dimensional view are used as the data in the main body of the tables (cells). This tool allows the user to drill up or down to any level of specificity pre-defined by the categorical data. PowerPlay is a very powerful exploratory data analysis tool. It allows for the quick calculation of the viewed data as percents, raw numbers, or even as any of a number of graphs or charts.

PowerPlay Reporter opens with a blank screen and from there the user specifies the exact variables wanted for rows, columns, layers, and filters. Here there are no restrictions on the types of variables used for rows or columns or layers as there are in PowerPlay Explorer. With this tool, tables that took many runs to create for table production are quickly created on the fly by the researcher with no programming at all.

Impromptu, the other major Cognos tool, is a report writer tool. This tool allows access to the entire warehouse of data. As it is linked directly to the warehouse, data access time is much longer than for either of the PowerPlay tools. The report tools available in Impromptu are much more limited than those available in PowerPlay. There is a simple one-way frequency, as well as a simple crosstab. The power of Impromptu lies in its ability to access the full warehouse. In PowerPlay the user is limited to a predefined set of variables. If a particular variable for an analysis is missing from any of the PowerCubes, then the researcher needs to go back to Impromptu to conduct analysis with that tool.

Lessons Learned

The most important part of the NCES prototype data warehouse is its scalability and portability. Any of a number of parts of the warehouse implementation can easily be built upon. The data model can be expanded to include other dimensions relevant to sample survey data, such as sample design, imputation, and weighting dimensions. Since ORACLE was used for the database management system, any of a number of SQL-based relational database management systems can be substituted for ORACLE. The front-end tool, Cognos, can also be substituted with any other ODBC driven OLAP tool. The prototype shows the ability of a data warehouse to quickly deliver data in a number of user-specified formats.

A key to data warehouse flexibility is the use of a star schema model. The star schema model allows for any physical implementation and is flexible enough to handle operating system and software changes. The data warehouse tools industry has progressed to the point where they realize that each individual tool is only a small part of the overall warehouse.

SMALL AREA ESTIMATION FOR THE DISTRIBUTION OF PARAMETERS

Michael P. Cohen, National Center for Education Statistics*
555 New Jersey Avenue NW, Washington DC 20208-5654

Key Words: Borrowing strength, Direct estimates, Distance, Ensemble

Abstract: If one wants to estimate a parameter for each of many small areas, one can generally improve the independent direct estimates by "borrowing strength" from the other small areas. Much research has been devoted to the situation in which one seeks to minimize the (possibly weighted) sums of the expected squared errors of the small area estimates. Thomas A. Louis, Malay Ghosh, and others have considered the contrasting situation in which the relationship among the small area parameters is of primary interest. For example, one might be interested in knowing the proportion of small areas where the high school dropout rate is above some level. The aim in such problems is to minimize the distance between the observed distribution of the "ensemble" (set) of small area estimates and the true distribution of the ensemble of parameters. In this paper we further explore the small area estimation problem when estimating the distribution of the parameters is the goal.

1. Introduction

Suppose we are investigating the values of a certain parameter (e.g. average income or an average measure of the level of literacy) for each of many small areas. If the goal is the best estimates of these parameters considered individually, then empirical and hierarchical Bayes techniques have been developed that improve upon naïve estimators. What if, though, we want to know which small areas have parameter values above a fixed cutoff C and which below? A different approach is required to treat problems of this type.

Louis (1984) was the first to study these small area estimation problems although Rubin (1981) had looked at the situation in another context. Ghosh (1992, 1994) built on the work of Louis, extending it to non-normal and multivariate situations. Our aim is to build on the work of these authors and, in particular, to investigate the use of loss functions that measure the distance between the distribution of the estimates and the distribution of the parameters.

For a general appraisal of small area estimation, Ghosh and Rao (1994) is highly recommended.

The very recent and interesting work of Shen and Louis (1998) studies and compares the different approaches to small area estimation in a two-stage hierarchical setting.

The organization of this paper is as follows: This introduction is Section 1. Section 2 provides background information. Section 3 introduces the loss functions that will be employed. In Section 4 we study a simple normal model, and in Section 5 we extend the results to more general situations. Some concluding remarks are given in Section 6.

2. Background

Consider the estimation of m parameters $\theta_1, \ldots, \theta_m$ under squared error loss. Let $\hat{\theta}_1^B, \ldots, \hat{\theta}_m^B$ denote Bayes estimates of these parameters based on data $\mathbf{X} = (X_1, \ldots, X_m)$. Let $\theta_\bullet = \frac{1}{m}\sum_{i=1}^m \theta_i$ and $\hat{\theta}_\bullet^B = \frac{1}{m}\sum_{i=1}^m \hat{\theta}_i^B$. Then

$$\mathrm{E}(\theta_\bullet|\mathbf{X}) = \hat{\theta}_\bullet^B$$

but

$$\mathrm{E}\left[\sum_{i=1}^m (\theta_i - \theta_\bullet)^2 \big| \mathbf{X}\right] > \sum_{i=1}^m \left(\hat{\theta}_i^B - \hat{\theta}_\bullet^B\right)^2.$$

This was shown by Louis (1984) under a normality assumption and, in general, by Ghosh (1992).

The point is that the Bayes estimates of the parameters (under squared error loss) have the same mean as the parameters themselves, but are on average less "spread out." If we are trying to use the collection of Bayes estimates to study the distribution of the parameters, we will have the distorted view that the parameters are more concentrated about their mean than they really are. We have been discussing Bayes estimates, but *empirical* Bayes estimates face the same problem.

In the context of small area estimation, the θ_i are parameters associated with small area i, say mean household income. If we use the $\hat{\theta}_i^B$ to study the θ_i, we will underestimate the diversity in the parameters.

Louis (1984) tackled this problem by investigating the class of estimators $\tilde{\theta}_i$ that satisfy

$$\mathrm{E}(\theta_\bullet|\mathbf{X}) = \tilde{\theta}_\bullet \qquad \text{where } \tilde{\theta}_\bullet = \frac{1}{m}\sum_{i=1}^m \tilde{\theta}_i$$

and

$$\mathrm{E}\left[\sum_{i=1}^{m}(\theta_i-\theta_\bullet)^2\big|\,\mathbf{X}\right]=\sum_{i=1}^{m}\left(\tilde{\theta}_i-\tilde{\theta}_\bullet\right)^2.$$

He still used squared error loss but it was minimized subject to these constraints. The constraints force a match on the first two moments between the distribution of the estimates and the distribution of the parameters.

In giving a theoretical basis to his work, Louis (1984, Subsection 2.2) introduced the notion of a general loss function operating on the empirical distributions of the parameter estimates and the parameters. Our investigation will be based on such loss functions; they are described in the next section.

3. Loss Functions

Given m parameters $\theta_1,\ldots,\theta_m$, define the function

$$G_m(t)=\frac{1}{m}\sum_{i=1}^{m}\mathrm{I}(\theta_i\le t) \tag{3.1}$$

where $\mathrm{I}(\cdot)$ is 1 when its argument is true and 0 otherwise. We can regard G_m as the empirical distribution function of the parameters. From a Bayesian point of view, the parameters are random variables. It should be noted, however, that the parameters will generally *not* be identically distributed and maybe not independent.

Let $\hat{G}_m$ be an estimator of G_m. For example, given m estimates $\hat{\theta}_1,\ldots,\hat{\theta}_m$ of $\theta_1,\ldots,\theta_m$ respectively, one could estimate G_m by

$$\hat{G}_m(t)=\frac{1}{m}\sum_{i=1}^{m}\mathrm{I}(\hat{\theta}_i\le t), \tag{3.2}$$

but we do not require estimators of G_m to be of the form (3.2). If we want to study the distribution of the θ_i, we would like to find an estimate $\hat{G}_m$ that is close, in some sense, to G_m. In other words, we would like $\|\hat{G}_m-G_m\|$ to be small where $\|\cdot\|$ is a distance function or metric. Examples of such distance functions include

$$\begin{aligned}&\|\hat{G}_m-G_m\|_{W,j}\\&\quad=\int_{-\infty}^{\infty}|\hat{G}_m(t)-G_m(t)|^j\,\mathrm{d}W(t),\end{aligned} \tag{3.3}$$

$$\begin{aligned}&\|\hat{G}_m-G_m\|_{\mathbf{t},\mathbf{w},j}\\&\quad=\sum_{\ell=1}^{L}w_\ell|\hat{G}_m(t_\ell)-G_m(t_\ell)|^j,\end{aligned} \tag{3.4}$$

and

$$\begin{aligned}&\|\hat{G}_m-G_m\|_\infty\\&\quad=\max_{-\infty<t<\infty}|\hat{G}_m(t)-G_m(t)|.\end{aligned} \tag{3.5}$$

In (3.3), $j>0$ and $W(t)$ is a weight function that we can choose to give more weight to ranges of parameter values in which we are especially interested. In (3.4), $j>0$ and the $\mathbf{w}=(w_1,\ldots,w_L)$ are weights attached to the points $\mathbf{t}=(t_1,\ldots,t_L)$. If we adopt a general definition of integral, the second distance function is just a special case of the first. An even more special case is

$$\|\hat{G}_m-G_m\|_{t_0,j}=|\hat{G}_m(t_0)-G_m(t_0)|^j$$

that considers only a single point in the space of parameter values. For example, if θ_i corresponds to average household income in small area i and $t_0=\$25,000$, then $|\hat{G}_m(t_0)-G_m(t_0)|$ measures how close we are in estimating the proportion of small areas with average household incomes less than or equal to $\$25,000$.

The distance function (3.5) is of great interest but difficult to work with analytically. There are, of course, other distance functions one might want to consider. In this paper, though, we concentrate on (3.3) with $j=2$. The goal is to minimize the (conditional) expected distance given the data.

If we are presented with a distribution function estimate $\hat{G}_m$ of the form (3.2), we can recover the set of values of the $\hat{\theta}_i$ from the jumps in the function $\hat{G}_m$, but we cannot determine uniquely which small area i is associated with which jump. In fact, any one-to-one assignment of the small areas to the jumps gives rise to the same value of $\hat{G}_m$. Letting $\boldsymbol{\theta}=(\theta_1,\ldots,\theta_m)$ and $\hat{\boldsymbol{\theta}}=(\hat{\theta}_1,\ldots,\hat{\theta}_m)$, Louis (1984, p. 394) suggests using a loss function of the form

$$\|\hat{G}_m-G_m\|+\epsilon\,\mathcal{L}(\hat{\boldsymbol{\theta}},\boldsymbol{\theta}) \tag{3.6}$$

for some small $\epsilon>0$ where $\mathcal{L}(\cdot,\cdot)$ is, for example, the sum of squared errors $\mathcal{L}(\hat{\boldsymbol{\theta}},\boldsymbol{\theta})=\sum_{i=1}^{m}(\hat{\theta}_i-\theta_i)^2$. The second term in the loss function is designed to force a unique assignment of the jumps in $\hat{G}_m$ of form (3.2) to the small areas i without otherwise affecting the loss function much.

Given any estimator $\hat{G}_m$ of G_m, not necessarily of form (3.2), we can estimate the ensemble $\{\theta_1,\theta_2,\ldots,\theta_m\}$ by

$$\begin{aligned}&\{\hat{\theta}_1,\hat{\theta}_2,\ldots,\hat{\theta}_m\}\\&\quad=\left\{\hat{G}_m^{-1}\left(\frac{1}{m+1}\right),\hat{G}_m^{-1}\left(\frac{2}{m+1}\right),\ldots,\right.\\&\qquad\left.\hat{G}_m^{-1}\left(\frac{m-1}{m+1}\right),\hat{G}_m^{-1}\left(\frac{m}{m+1}\right)\right\}.\end{aligned}$$

We use (3.6) to determine which $\hat{\theta}_i$ corresponds to

$$\hat{G}_m^{-1}\left(\frac{1}{m+1}\right),$$

and so forth.

In the next section, we make use of some of the loss functions described in this section to investigate a simple normal model.

4. Simple Normal Model

Suppose that each $\theta_i \sim \mathrm{N}(\mu, \tau^2)$, that is, suppose each θ_i is normally distributed with mean μ and variance τ^2. Suppose further that the θ_i are independent. Let X_i given θ_i be $\mathrm{N}(\theta_i, 1)$ and let the X_i be independent, $i = 1, \ldots, m$. We shall use this simple model as a starting point.

For known μ and τ^2, the posterior distribution of θ_i given $\mathbf{X}$ is normal with mean $\mathrm{E}(\theta_i|\mathbf{X}) = \mu + \frac{\tau^2}{1+\tau^2}(X_i - \mu)$ and variance $\mathrm{var}(\theta_i|\mathbf{X}) = \frac{\tau^2}{1+\tau^2}$. The $\theta_i|\mathbf{X}$ are independent. Letting $\gamma = \frac{\tau^2}{1+\tau^2}$, we have

$$\begin{aligned}
\mathrm{E}\{G_m(t)|\mathbf{X}\} &= \frac{1}{m}\sum_{i=1}^{m}\mathrm{E}\{\mathrm{I}(\theta_i \le t)|\mathbf{X}\} \\
&= \frac{1}{m}\sum_{i=1}^{m}\Pr(\theta_i \le t|\mathbf{X}) \\
&= \frac{1}{m}\sum_{i=1}^{m}\Phi\left(\frac{t-\mu-\gamma(X_i-\mu)}{\sqrt{\gamma}}\right)
\end{aligned} \tag{4.1}$$

where Φ is the standard normal distribution function.

Let us consider the distance function

$$\|\hat{G}_m - G_m\|_{W,2} = \int_{-\infty}^{\infty}(\hat{G}_m(t) - G_m(t))^2\,\mathrm{d}W(t)$$

where $W(t) \ge 0$ and $\int_{-\infty}^{\infty}\mathrm{d}W(t) < \infty$. The conditional expected distance given $\mathbf{X}$ is

$$\begin{aligned}
&\mathrm{E}(\|\hat{G}_m - G_m\|_{W,2} \mid \mathbf{X}) \\
&= \mathrm{E}\left\{\int_{-\infty}^{\infty}(\hat{G}_m(t) - G_m(t))^2\,\mathrm{d}W(t)\,\middle|\,\mathbf{X}\right\} \\
&= \int_{-\infty}^{\infty}\mathrm{E}\left\{(\hat{G}_m(t) - G_m(t))^2\,\middle|\,\mathbf{X}\right\}\,\mathrm{d}W(t).
\end{aligned}$$

The last step is justified because the integrand is nonnegative and bounded. But the last integral can be minimized by minimizing

$$\mathrm{E}\left\{(\hat{G}_m(t) - G_m(t))^2\,\middle|\,\mathbf{X}\right\} \tag{4.2}$$

for each t. Note that the solution does not depend on $W(t)$. It is known from standard results in Bayes estimation that (4.2) is minimized by the choice $\hat{G}_m(t) = \mathrm{E}\{G_m(t)|\mathbf{X}\}$. For the simple normal model, the latter quantity is given by (4.1).

Note: For $W(t) \equiv t$, Shen and Louis (1998) obtain

$$\hat{G}_m(t) = \mathrm{E}\{G_m(t)|\mathbf{X}\} = \frac{1}{m}\sum_{i=1}^{m}\Pr(\theta_i \le t|\mathbf{X})$$

for a two-stage hierarchical model.

It is of interest to compute the (conditional) expected loss because this provides a measure of the closeness of estimation, analogous to mean squared error. If $\hat{G}_m(t) = \mathrm{E}\{G_m(t)|\mathbf{X}\}$, then

$\mathrm{E}\left\{(\hat{G}_m(t) - G_m(t))^2\,\middle|\,\mathbf{X}\right\} = \mathrm{var}\{G_m(t)|\mathbf{X}\}$, so

$$\mathrm{E}(\|\hat{G}_m - G_m\|_{W,2} \mid \mathbf{X}) = \int_{-\infty}^{\infty}\mathrm{var}\{G_m(t)|\mathbf{X}\}\,\mathrm{d}W(t). \tag{4.3}$$

But

$$\begin{aligned}
&\mathrm{var}\{G_m(t)|\mathbf{X}\} \\
&= \frac{1}{m^2}\sum_{i=1}^{m}\mathrm{var}\{\mathrm{I}(\theta_i \le t)|\mathbf{X}\} \\
&= \frac{1}{m^2}\sum_{i=1}^{m}\Pr(\theta_i \le t|\mathbf{X})\{1 - \Pr(\theta_i \le t|\mathbf{X})\} \\
&= \frac{1}{m^2}\sum_{i=1}^{m}\left[\Phi\left(\frac{t-\mu-\gamma(X_i-\mu)}{\sqrt{\gamma}}\right)\right. \\
&\qquad \left.\times\left\{1 - \Phi\left(\frac{t-\mu-\gamma(X_i-\mu)}{\sqrt{\gamma}}\right)\right\}\right].
\end{aligned} \tag{4.4}$$

From (4.3) and (4.4), $\mathrm{E}(\|\hat{G}_m - G_m\|_{W,2} \mid \mathbf{X})$ can be computed.

5. More General Situations

5.1 Normal Model, Unequal but Known Variances

As before, let each $\theta_i \sim \mathrm{N}(\mu, \tau^2)$ and let the θ_i be independent. Now let X_i given θ_i be $\mathrm{N}(\theta_i, \sigma_i^2)$ and let the X_i be independent and $\sigma_i^2 > 0$, $i = 1, \ldots, m$.

Let

$$\gamma_i = \frac{\tau^2}{\sigma_i^2 + \tau^2}.$$

For known μ, τ^2, and σ_i^2, the posterior distribution of θ_i given $\mathbf{X}$ is normal with mean

$$\mathrm{E}(\theta_i|\mathbf{X}) = \mu + \gamma_i(X_i - \mu)$$

and variance

$$\mathrm{var}(\theta_i|\mathbf{X}) = \gamma_i\sigma_i^2.$$

All the results of Section 4 continue to hold for this more general model, with

$$\frac{t - \mu - \gamma_i(X_i - \mu)}{\sigma_i\sqrt{\gamma_i}}$$

replacing

$$\frac{t - \mu - \gamma(X_i - \mu)}{\sqrt{\gamma}}$$

in (4.1) and (4.4).

5.2 Empirical and Hierarchical Bayes Techniques

Most frequently, μ and τ^2 will be unknown and require estimation. There are standard empirical and hierarchical Bayes methods for doing this. See, for example, Ghosh and Rao (1994). The (conditional) expected loss can be estimated by means of Markov chain Monte Carlo methods.

6. Concluding Remarks

This paper has built upon the work of Louis (1984), Ghosh (1992), and others that study ways of estimating the distribution of small area parameters. Our focus has been on using loss functions that measure the distance between the distribution of the estimates of the parameters and the distribution of the parameters themselves. There are many aspects of this problem that have yet to be explored.

Acknowledgment: The author thanks Professor Malay Ghosh for acquainting him with this area of research in his 1994 Conference Board for the Mathematical Sciences lectures at the University of Connecticut, Storrs.

REFERENCES

Ghosh, M. (1992). Constrained Bayes estimation with applications, *Journal of the American Statistical Association* **87** 533–540.

Ghosh, M. (1994). Bayesian methods in survey sampling, Conference Board for the Mathematical Sciences Seminar, University of Connecticut, unpublished notes.

Ghosh, M., and Rao, J.N.K. (1994). Small area estimation: an appraisal (with discussion), *Statistical Science* **9** 55–93.

Louis, T.A. (1984). Estimating a population of parameter values using Bayes and empirical Bayes methods, *Journal of the American Statistical Association* **79** 393–398.

Rubin, D.B. (1981). Estimation in parallel randomized experiments, *Journal of Educational Statistics* **6** 377–401.

Shen, W., and Louis, T.A. (1998). Triple-goal estimates in two-stage hierarchical models, *Journal of the Royal Statistical Society,* Series B **60** 455–471.

*This paper is intended to promote the exchange of ideas among researchers and policy makers. The views are those of the author, and no official support by the U.S. Department of Education is intended or should be inferred.

1997 NCES CUSTOMER SATISFACTION SURVEY FINDINGS

Steven Fink, Lori Thurgood, Rita Bureika, Julie Scott, Sameena Salvucci, Synectics for Management Decisions, Samuel Peng, National Center for Education Statistics
Steven Fink, Synectics, 1901 North Moore St., Arlington, VA 22209

KEY WORDS: Service Standards, Performance Indicators, Customer Satisfaction Survey

INTRODUCTION

The Government Performance and Results Act (GPRA) was enacted on August 3, 1993, to "provide for the establishment of strategic planning and performance measurement in the Federal Government." One purpose of this landmark legislation is to:

> *"improve Federal program effectiveness and public accountability by promoting a new focus on results, service quality, and customer satisfaction . . . ."*

On September 11, 1993, President Clinton issued Executive Order 12862, "Setting Customer Service Standards," which called on all Federal agencies to develop plans to better serve their customers. This order requires agencies to survey customers to determine the kind and quality of services they want and their level of satisfaction with existing services.

On March 22, 1995, the President sent additional guidance to the heads of agencies in a memorandum entitled, "Improving Customer Service." The memorandum established that customer surveys are "ongoing" and "continuing" requirements. Further, it established that development and tracking of customer service measures, standards, and performance should be integrated with other performance initiatives, including strategic planning and performance measurement under GPRA.

The first Federal education agency was established in 1867 "for the purpose of collecting such statistics and facts as shall show the condition and progress of education in the several States and territories . . . ." The National Center for Education Statistics (NCES) currently performs these duties in a greatly expanded Department of Education. Its primary responsibility is to collect, analyze, and disseminate statistics relating to the status of education in the United States. This mission of NCES is, in itself, closely related to the public good. It requires that NCES collect information from and provide information to the public—its customers.

BACKGROUND

To respond to the GPRA requirements, NCES assembled a customer service team in 1997 to initiate and oversee many customer-related initiatives. These have included:

- Conducting customer focus groups with key customers
- Conducting the 1996 Customer Satisfaction Survey, targeting a broad range of current NCES customers
- Conducting the 1997 Customer Satisfaction Survey, targeting policymakers at the federal, state, and local levels, and researchers focusing on issues of educational improvement
- Making current publications and other relevant information available on the NCES web site
- Providing training seminars for NCES data customers
- Developing partnerships with education associations and broadening outreach efforts to encompass a larger segment of the NCES customer base
- Training employees about customer service

NCES has customers of many types and interests. They include:

- Policymakers (local, state, and national)
- Federal government customers
- Media, education associations
- Administrators and heads of institutions
- Researchers
- Parents, teachers, and community leaders

These customers vary in how directly NCES' work affects them and the extent to which they use (or are even aware of) the broad range of NCES products and services. Each customer group was the target of at least one focus group meeting and report that NCES completed in 1994 and 1995.

In 1996, Synectics for Management Decisions, Inc. conducted the first comprehensive customer satisfaction survey for NCES. The target population for the 1996 Customer Satisfaction Survey was *known customers across all categories of NCES customers*.

From an initial population of 11,286, NCES sent questionnaires to 4,760 customers; more than 39 percent (1,887) responded. The 1996 survey provided a broad overview of customers' use of and satisfaction with NCES products and services, established a baseline of information, and yielded useful information from which to shape subsequent surveys.

1997 NCES Customer Satisfaction Survey

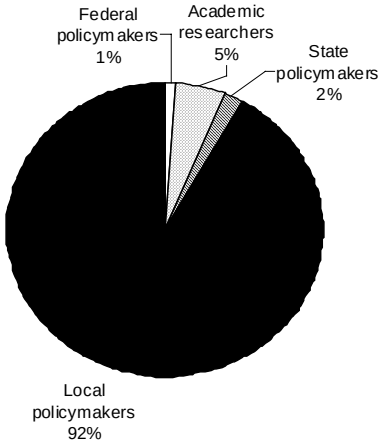
As in 1996, the 1997 survey was designed to determine information about customers and their levels of satisfaction and needs related to:

- NCES publications and reports
- NCES data files
- NCES services

Results were also used to develop and refine NCES performance objectives and indicators under GPRA.

The target population for the 1997 customer survey included two important segments of the overall NCES customer base—policymakers and researchers. The target population was divided into four groups or strata (see Figure 1).

Figure 1.—Percentage distribution of customer groups in target population: 1997



SOURCE: U.S. Department of Education, National Center for Education Statistics, 1997 Customer Satisfaction Survey.

- **Federal policymakers**: Includes U.S. Department of Education Assistant and Under Secretaries, National Science Foundation (NSF), Office of Management and Budget (OMB), Congressional Research Service (CRS), General Accounting Office (GAO), Senate and House Committees.
- **State policymakers**: Includes National Conference of State Legislators (NCSL), Council of Chief State School Officers (CCSSO), State Higher Education Executive Finance Officers (SHEEFO), Chief Officers of State Library Agencies (COSLA).
- **Local policymakers**: Includes individuals affiliated with either elementary/secondary school districts or postsecondary institutions, such as school district superintendents and higher education chief administrators (primarily directors of institutional research).
- **Academic researchers**: Includes directors of Office of Educational Research and Improvement (OERI) National Research and Development Centers, directors of Regional Education Laboratories, deans of Schools of Education, chairs of Sociology Departments.

METHODOLOGY

Frame Development

The target population for the 1997 NCES Customer Satisfaction Survey comprised 20,033 federal, state, and local policymakers and academic researchers. The population included current and potential customers.

The first step in developing a sampling frame was to create lists of names, addresses, and telephone numbers of federal and state policymakers. Next, the Common Core of Data (CCD) and the Integrated Postsecondary Education Data System (IPEDS) databases were used to develop lists of local policymakers. The 1993 CCD includes elementary/secondary school districts. The 1994 IPEDS includes postsecondary institutions.

IPEDS was also used to develop one portion of the list of academic researchers—deans of Schools of Education. The American Sociological Association was the source of names, addresses, and telephone numbers of chairs of Departments of Sociology. The Office of Educational Research and Improvement (OERI) provided this same information for directors of OERI National Research and Development Centers and directors of Regional Education Laboratories.

Final contact persons for the local policymaker and academic researcher strata were identified through phone calls to more than 2,500 institutions and school districts. The names and titles/positions of these individuals were added to the records in the frame database. The final product was a frame database with the following principal fields: names, titles, addresses, telephone numbers, and stratum and substratum identifiers.

Sample Design

For the 1997 survey, all federal and state policymakers were sampled (n=129 and n=361, respectively), and samples were drawn from the local policymaker and academic researcher groups.

For local policymakers, the two substrata—the elementary/secondary education group and the higher education group—were treated separately. The elementary/secondary education substratum was further stratified based on eight urbanicity levels. The higher education substratum was further stratified based on

Carnegie Foundation Classification Codes. Among the academic researchers, all individuals in the OERI National Research and Development Centers and Regional Education Laboratories were sampled. A sample was taken of deans of Schools of Education and chairs of Sociology Departments.

The 1997 survey design called for surveys to be sent to a subsample of customers who had also responded to the 1996 survey (the overlap sample). This enabled NCES to determine the extent of change in customer perceptions and opinions about NCES products and services.

A subsample of 294 customers was selected from the entire 1996 responding population. Of the subsample, 216 (74%) reported in both years on their use of NCES products and services. The overlap analysis compares the 1996 and 1997 responses of the same 216 respondents. The results were not weighted to the population.

Questionnaire Design and Pretest

Although the 1996 Customer Satisfaction Survey instrument was used as the basis for the redesign of the 1997 instrument, there are several differences between the two. First, two new sections were added to the 1997 instrument: Benchmarking and Non-users. The questions in the Benchmarking section were designed to capture more useful information about other sources of education data used by NCES customers and to develop measures of comparison between NCES and other organizations that can serve as benchmarks for future evaluations. The Non-users section of the 1997 survey asks non-users of NCES data to identify other organizations/associations they may have used for education data. Non-users are then asked, in an open-ended question, to identify any needs they may have for education data. Other differences between the 1996 and 1997 survey instruments include additional open-ended items and slight changes in the order of the sections, question wording, and response categories.

Survey Operations

The survey was conducted during the summer of 1997 by Westat, a survey research firm in Rockville, Maryland. The initial mailing was sent in late July. The instructions on the survey indicated that the questionnaire was a telephone survey script that respondents could complete and return. Starting two weeks later, all nonrespondents were called and data were collected over the telephone. The data collection period lasted 10 weeks.

During the data collection phase, 32 out-of-scope cases were identified, the majority being from schools/districts that had closed. Subtracting the out-of-scope cases from the sample provided a total in-scope sample of 2,948. The final survey response rate was 84 percent. Among the 2,465 responding cases, 810 surveys (33%) were completed by mail and 1,655 (67%) were completed in the telephone follow-up. The highest response rate occurred among local policymakers (86%), while the lowest occurred among federal policymakers (71%).

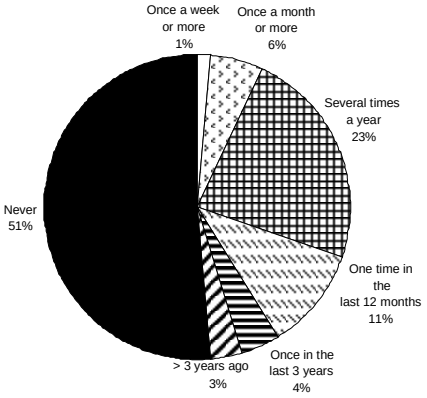
For the overlap sample, the final response rate was 75 percent. Among the 216 responding cases, 102 surveys (47%) were completed by mail and 114 (53%) were completed in the telephone follow-up.

SUBSTANTIVE FINDINGS

Questions about You, Our Customer

Overall, an almost equal number of customers had used NCES products or services (49%) as had not (51%), as shown in Figure 2. Most federal and state policymakers were users (80% and 78%, respectively), whereas about half of the local policymakers and academic researchers were users (47% and 59%, respectively).

Figure 2.—Users and non-users of NCES products and services, by customer group



SOURCE: U.S. Department of Education, National Center for Education Statistics, 1997 Customer Satisfaction Survey.

Questions about NCES Publications and Reports

More than half of the customers (56%) had not used NCES publications or reports in the past three years. Customers were asked to indicate all the reasons why they had not. The two most common responses were not being aware of NCES publications and reports (59%) and their work not requiring use of the publications and reports (40%).

Of those who had used NCES statistical compendia, a high percentage of users reported that they were very satisfied or satisfied:

- Condition of Education (91%)
- Digest of Education Statistics (90%)
- Projections of Education Statistics (86%)

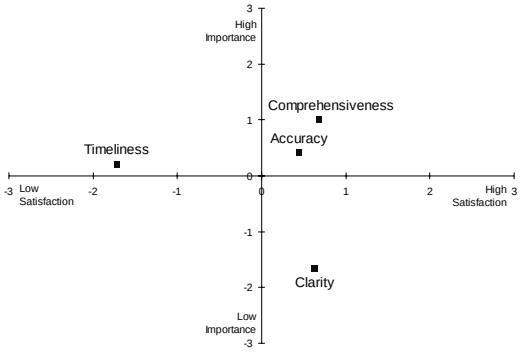
Users also reported high levels of satisfaction with publications across each of the NCES program areas. In fact, users in the overlap sample reported levels of satisfaction within several percentage points of those reported in 1996—most often above or near 90 percent very satisfied or satisfied. Among all 1997 users, publications and reports that were used more frequently showed higher levels of satisfaction.

A very high percentage of users in the 1997 survey reported being very satisfied or satisfied with most aspects of NCES publications and reports:

- Overall quality of reports (90%)
- Comprehensiveness (88%)
- Clarity of writing (87%)
- Usefulness to work (86%)

Users rated accuracy of the information (39%) as the most important aspect, followed by timeliness of the information (33%) and comprehensiveness (17%). Figure 3 displays levels of satisfaction with an aspect versus ratings of its importance.

Figure 3.—Comparison of levels of satisfaction versus importance for aspects of NCES publications and reports



SOURCE: U.S. Department of Education, National Center for Education Statistics, 1997 Customer Satisfaction Survey.

Questions about NCES Data Files

While a very low percentage (7%) of all customers had used NCES electronic data files, usage was higher among state and federal policymakers and academic researchers (31%, 18%, and 14%, respectively) than among local policymakers (6%).

The most frequently cited reason for not using NCES data files (62%) was the same as the one cited for not using publications—the customers were not aware of the products.

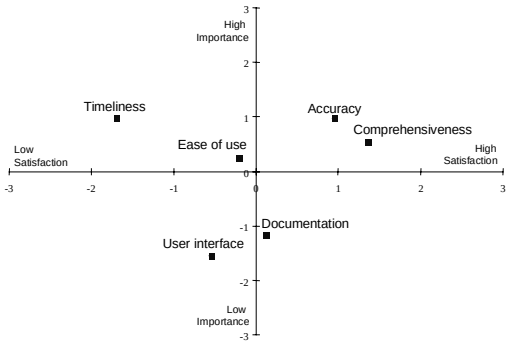
Most of the users of NCES electronic data files reported being very satisfied or satisfied with the data files, within a range of 67 percent to 95 percent. In fact, users in the overlap sample reported levels of satisfaction within several percentage points of those reported in 1996—most often above or near 85 percent very satisfied or satisfied. Among all 1997 users, electronic data files that were used more frequently showed higher levels of satisfaction than those data files with a lower frequency of use.

Users were asked how satisfied they were with six aspects of NCES electronic data files. Their levels of satisfaction are shown below:

- Comprehensiveness (82%)
- Accuracy of data in the file (74%)
- File documentation (70%)
- Ease of use (69%)
- User interface (55%)
- Timeliness of data release (52%)

Users rated accuracy of the information (40%) as the most important aspect, followed by timeliness of the information (21%) and ease of use (20%). Figure 4 displays levels of satisfaction with an aspect versus ratings of its importance.

Figure 4.—Comparison of levels of satisfaction versus importance for aspects of NCES electronic data files



SOURCE: U.S. Department of Education, National Center for Education Statistics, 1997 Customer Satisfaction Survey.

Questions about NCES Services

Customers were also asked about their use of NCES services. Overall, two-thirds (66%) of the customers did not know how to contact NCES. However, significantly more federal and state policymakers and academic researchers knew how to contact NCES (76%, 69%, and 45%, respectively) than did local policymakers (32%). Less than half of the customers were aware of each of the following NCES services:

- Internet services (49%)
- National Education Data Resource Center (48%)
- Toll-free number for education statistics (46%)
- NCES conferences (45%)
- NCES training seminars or workshops (42%)

Users of NCES services were asked about their satisfaction with aspects of the services. A high percentage of users were very satisfied or satisfied with the ease of obtaining information (92%); extent to which the information met their needs (90%); courtesy of staff (90%); and speed with which they received the information (89%).

Questions about Benchmarking

Users of NCES products or services were also asked about their experiences in obtaining education data from other organizations. Virtually all users of NCES products or services reported obtaining some education data from at least one organization other than NCES during the last three years. The organizations mentioned most often were:

- State Departments of Education (96%)
- Association for Supervision and Curriculum Development (86%)
- U.S. Bureau of the Census (84%)
- Educational Research Service (83%)
- Bureau of Labor Statistics (79%)

Users were asked to name the one organization other than NCES with which they were most favorably impressed (that is, their benchmark organization). The following nine organizations accounted for more than three-fourths of all responses:

- Association for Supervision and Curriculum Development (33%)
- State Departments of Education (19%)
- Educational Research Service (7%)
- American Association of School Administrators, National Education Association, U.S. Bureau of the Census, American Council of Education, Bureau of Labor Statistics, and Phi Delta Kappa (2-4% each)

Users were also asked to rate their benchmark organization as "better" than, the "same" as, or "worse" than NCES on each of seven dimensions. Overall, and among local policymakers, a majority of users gave "better" ratings to their benchmark organizations on all dimensions except quality of product and frequency of data collection. On these two dimensions, a majority considered their benchmark organization to be the "same" as or "worse" than NCES.

Questions for Non-Users

NCES non-users are individuals who have never used NCES products or services but who, based on their needs for education data, are potential customers. They formed 51 percent of all customers in the 1997 survey.

Non-users of NCES products or services were asked whether or not they had used education data from any other organization in the last three years, and if so, how often. About 96 percent of this group reported using data from another organization. Their usage of the various organizations is shown below:

- State Departments of Education (92%)
- Association for Supervision and Curriculum Development (69%)
- U.S. Bureau of the Census (56%)
- National Education Association (54%)
- Educational Research Service (52%)
- Bureau of Labor Statistics (51%)

Only the State Departments of Education were used frequently by a majority of non-users (72%).

In an open-ended question, 91 percent of NCES non-users described their needs for education data while 9 percent indicated that they had no needs. Responses of those reporting needs fell into the following categories: curriculum and planning/ standards (24%), institution/school governance (22%), specific education issues (17%), other (16%), local/state/regional information and comparisons (12%), and uses other sources (9%).

LIMITATIONS OF RESULTS

There are a few cautions to consider when interpreting the results of the 1997 Customer Satisfaction Survey. First, NCES does not have a comprehensive database that covers all of its customers. The 1997 survey represents four targeted groups of customers who may not comprise a representative fraction of all NCES customers.

Second, 92 percent of the targeted population were local policymakers. Furthermore, of the local policymakers, most (89%) were affiliated with elementary and secondary school districts. The survey results, therefore, will be heavily dominated by responses from this subgroup.

NEXT STEPS

In response to the findings from several focus groups and from the 1996 and 1997 Customer Satisfaction Surveys, NCES is implementing several strategies to improve awareness and satisfaction levels. First, NCES plans to enhance its outreach program, which increases public awareness and use of NCES products and services, to all user groups, with particular emphasis on local practitioners. Specific efforts include:

- creating a "tailored" practitioner web page
- increasing representation at conferences
- conducting training sessions for NCES current and potential customers
- expanding partnerships with such groups as the National School Boards Association and the American Association of School Administrators
- targeting publications for NCES current users

Timeliness of data was an area of concern indicated by customers as high in importance and low in satisfaction. NCES is addressing this concern by:

- assessing customer needs through focus groups, surveys, and web-posted queries
- examining procedures to determine if streamlining is possible at any stage prior to the release of publications and data files
- conducting methodological studies to identify reliable methods for deriving early estimates and disseminating them in a timely fashion
- managing customer expectations by communicating anticipated product release schedules through the NCES web page, training sessions, professional conferences, and association newsletters
- integrating plans for a range of NCES data products with the development of the NCES data warehouse

NCES also plans to examine benchmark organizations that customers identified as better than NCES. Strategies may include focus groups with organizations such as the Association for Supervision and Curriculum Development and State Departments of Education—two of the highest rated and most frequently used benchmark organizations.

CONCLUSIONS

Survey results from the 1996 and 1997 Customer Satisfaction Surveys indicate high levels of satisfaction with NCES publications, reports, data files, and services. With these expressions of satisfaction, however, have come some warnings about areas that NCES needs to improve, such as awareness levels of products and services and timeliness of data. NCES plans to share these survey findings with program managers to determine what can be done to improve these areas. NCES should also continue to conduct surveys and focus groups of their key customers to identify specific areas of improvement. It is up to NCES to respond positively to its customers and take actions that will continue to improve the quality, timeliness, and usability of its products and services on behalf of its customers.

REFERENCES

Thurgood, L., Fink, S., Bureika, R., Scott, J., and Salvucci S. (1998). 1997 NCES Customer Satisfaction Survey Report. Report to the National Center for Education Statistics. Arlington, VA: Synectics for Management Decisions, Inc.

Costello, A., Scott, J., Fink, S., and Salvucci S. (1997). 1996 NCES Customer Satisfaction Survey Report. Report to the National Center for Education Statistics. Arlington, VA: Synectics for Management Decisions, Inc.

Costello, A., Scott, J., Walter, E., and Salvucci S. (1998). National Center for Education Statistics Performance Report: September 1998. Report to the U.S. Department of Education. Arlington, VA: Synectics for Management Decisions, Inc.

USING THE BOOTSTRAP TO ESTIMATE THE VARIANCE FROM A SINGLE SYSTEMATIC PPS SAMPLE

Steven Kaufman, National Center for Education Statistics
Room 402b, 555 New Jersey Ave NW, Washington, DC 20208

Key Words: Balanced Half-Sample Replication

1.0 Introduction

Systematic sampling (either with equal or unequal selection probabilities) is a common sampling scheme in complex sample designs. It is used because of its simplicity of implementation and its potential increase in efficiency, given a good frame ordering, which acts as an additional stratification.

One problem with systematic sampling is that such samples can be viewed as a cluster sample of cluster sample size one. As such, unbiased variance estimation becomes impossible without additional assumptions. One common method for approximating the variance from systematic sampling is to treat the sample as a super-stratified sample. This is accomplished by placing the sample selected within a stratum into the order it was selected and pairing consecutively selected PSUs. Each pair can then be treated as a pseudo-stratum for variance estimation purposes.

There are problems using the pseudo-stratum variance approach. The main problem is that the pseudo-stratum variances still does not reflect the appropriate systematic sampling variance. As such, the variance may only reflect with-replacement sampling. By assumption, the correlation between pseudo-strata is assumed to be zero. At first glance, it seems like these drawbacks would lead to an overestimate of the variance. However, since the correlations can be negative, this need not be the case.

In Kaufman (1998), it is shown that using the pseudo-stratum approach can produce large underestimates of the variance. To reduce this problem, the 1998 paper proposes a consistent bootstrap variance estimation procedure. The advantage of the bootstrap methodology is that it becomes possible to reflect an appropriate systematic sampling variance. The problem with this procedure is that without special adjustments, the bootstrap estimator is biased. To produce an unbiased variance estimator, adjustments are based on estimates from multiple samples. Generally, this is only possible with variables on the frame. Since the required adjustment is dependent on the variable of interest, the proposed procedure can have limited utility.

In this paper, the frame will be randomized in a controlled way, so that some of the affects on efficiency of the frame ordering are maintained, while eliminating the within and between pseudo-stratum correlations. Without the correlations, it becomes possible to estimate the variance in an unbiased fashion, where the expectation is taken across all possible random orderings. With an unbiased variance estimator, the bootstrap variance estimator can be adjusted using only data from a single sample.

The organization of this paper is: 1) define the randomized systematic sampling, 2) define the bootstrap procedure, 3) describe a simulation study to test the bootstrap variance estimator, and 4) present the results and conclusions.

2.0 Systematic Sampling

Systematic probability proportionate to size sampling (PPS) is a common procedure used with complex sample designs. The procedure is described in (Wolter, 1985, pp. 283-286). The idea is to divide the frame into consecutive, exhaustive and disjoint groups of Primary Sampling Units (PSUs), called partition groups, such that the total measures of size in each group are all equal. The total measure of size in a group is called the sampling interval. For this to work, some PSUs must span multiple partition groups. The first sampled PSU is randomly selected from PSUs in the first partition group. All other PSUs are selected systematically, one per partition group, starting from the point of selection of the first PSU.

It is assumed that before sample selection, PSUs with measures of size larger than the sampling interval have been excluded from the sampling. Such units are considered certainty PSUs.

An unbiased estimate for the total of variable X ($\hat{T}_{sy}$) is $\sum_{h=1}^{H}\sum_{i=1}^{n_h} x_i / p_i$, where H is the number of stratum, n_h is the number of sampled PSUs in stratum h, x_i is the value of X for selected PSU i, and p_i is the selection probability for the PSU (i.e., p_i is the measure of size for PSU i divided by the stratum sampling interval).

To simply the development of the randomized systematic sampling procedure $\hat{T}_{sy}$ will be rewritten into an equivalent estimator by treating the PSUs split between partition groups differently. Assume a_h PSUs on the frame split between partition groups. Each of these a_h PSUs will be split into two pseudo-PSUs (j_1 and j_2). For PSU $j \in a_h$ with probability of selection p_j, the first pseudo-PSU selection probability (p_{j_1}) is the part of p_j in the first partition group containing j and the second pseudo-PSU selection probability (p_{j_2}) is $p_j - p_{j_1}$. The partitioning weights (w_{j_1} and w_{j2}) are

p_{j_1} / p_j and $p_{j_{12}} / p_j$, respectively. Without loss of generality, the PPS selection described above can be viewed as selected from this new setup. For a given $j \in a_h$, at most one j_1 or j_2 can be selected. A j_1 or j_2 actually selected will be denoted by $j_\bullet$. And let k_h be the number of $j_\bullet$ selected in stratum h.

Now,

$$\hat{T}_{sy} = \sum_{h=1}^{H} \sum_{i=1}^{n_h} x_i / p_i = \sum_{h=1}^{H} \left(\sum_{\substack{i=1 \\ i \notin a_h}}^{n_h} x_i / p_i + \sum_{\substack{j_\bullet=1 \\ j \in a_h}}^{k_h} (w_{j_\bullet} x_j / p_{j_\bullet}) \right)$$

3.0 Estimating the $V(\hat{T}_{sy})$ with a Randomized Frame

In this section, an unbiased variance estimator for $\hat{T}_{sy}$ will be derived using only the selected sample. To do this, the sampling frame must be randomized before the sample selection. The variance estimator will then be unbiased across all frame randomizations. The randomization will be done to maintain most of the affects of the implicit stratification induced by the original frame ordering. First, a general expression for $V(\hat{T}_{sy})$ is specified. Next, the frame randomization is specified. Finally, an unbiased variance estimator is derived.

3.1 General Expression for $V(\hat{T}_{sy})$

$V(\hat{T}_{sy})$ can be express as:

$$\sum_{h=1}^{H} \left[\sum_{i=1}^{n_h} V(x_i / p_i) + \sum_{i=1}^{n_h} \sum_{j \neq i}^{n_h} \rho_{hij} \sqrt{V(x_i / p_i) V(x_j / p_j)} \right] \quad (1)$$

where: ρ_{hij} is the weighted correlation between the i^{th} and j^{th} PSUs selected in the systematic selection process.

Of course, without further assumptions, none of the above quantities have unbiased estimates. With the randomized ordering, $\hat{T}_{sy}$ will be denoted as $\hat{T}_{rsy}$.

3.2 New Ordering of the Frame

To simplify the variance estimation, the original frame ordering will be modified. The first step in this process is to define pseudo-strata similar to those described in section 1.0. Within each stratum, place the frame in its original ordering. Next, determine the partition groups as described in the section 2.0. Partition groups are now consecutively paired. Each pair is considered a pseudo-stratum (ps). After, the pseudo-strata are determined, PSUs that are in multiple pseudo-strata must be spilt into two pseudo-PSUs, as described at the end of section 2.0. The final step is to randomize the PSUs and pseudo-PSUs within each pseudo-stratum. It is assumed that n_h is even, so there should be two PSU selections within each pseudo-stratum.

This methodology maintains much of the additional stratification induced by the original ordering. With the original ordering, any contiguous group on the frame would have a selected sample size within one of the expected sample size for that group. With the new ordering, the selected sample size will be within two of the expected sample size.

Another advantage of the new ordering is that across all possible frame randomization the correlation between the i^{th} and j^{th} PSUs selected is zero

(i.e., ρ_{hij} =0). (2)

One disadvantage with the new ordering is that if the sum of the covariance terms, from the original ordering, is negative for a variable X then the variance under the new ordering will be less efficient than under the old ordering. The reverse is also true, if the sum of the covariance terms is positive. Of course, with multiple purpose surveys, where many variables are measured, there may be some variables where the sum of the covariance terms is either positive or negative. In this situation, it isn't clear which ordering is overall more efficient. However, variance estimates based on the new ordering should not be negatively biased due to the covariances. A second disadvantage is that it becomes possible to select a pseudo-PSU multiple times. One way of minimizing this impact is to compute the expected number of pseudo-PSUs ($E(ps)$) selected twice and increasing the sample size by this amount: $E(ps) = \sum_h \sum_{j \in a_h} p_{j_1} p_{j_2}$, where a_h refers to the set of PSUs that span multiple pseudo-stratum.

3.3 Estimating $V(\hat{T}_{rsy})$ using the New Ordering

To estimate $V(\hat{T}_{rsy})$, the sampling must be conditioned on three things. The first, denoted by 1, represents the random ordering process described in section 3.2. The second, denoted by 2, represents the PPS systematic sampling process. The third, denoted by N_g, represents the number of PSUs/pseudo-PSUs in partition group g. There are two ways pseudo-PSUs can be formed. The first way is in the formation of the pseudo-strata described in section 3.2, which could generate a pseudo-PSU in each g. For partition group g, assume there are $m_g^{(1)}$ such units (i.e., $m_g^{(1)}$ = 0 or 1). Within a pseudo-stratum, a PSU may still span two partition groups. In this situation, the PSU would be converted into 2 pseudo-PSUs. Within g, assume there are $m_g^{(2)}$ of these units (i.e., $m_g^{(2)}$ =0 or 1). The number of PSUs and pseudo-PSUs in a partition group g is:

$N_g = N_g^{ns} + m_g^{(1)} + m_g^{(2)}$, where N_g^{ns} is the number of non-splitting PSUs.

Given this:

$$V(\hat{T}_{rsy}) = \underset{N_g}{E}\,\underset{1}{E}V(\hat{T}_{rsy}) + \underset{N_g}{E}\,V\,\underset{1}{E}\underset{2}{E}(\hat{T}_{rsy}) + V\,\underset{N_g}{E}\,\underset{1}{E}\underset{2}{E}(\hat{T}_{rsy})$$

$$= \underset{N_g}{E}\,\underset{1}{E}V(\hat{T}_{rsy}), \text{ since } \underset{2}{E}(\hat{T}_{rsy}) = T \text{ (the population total)}$$

$$= \sum_{h}^{H} \sum_{ps\in h} \sum_{g\in ps} \underset{N_g}{E}\,\underset{1}{E}V(\hat{T}_{rsy,ps,g}), \text{ from (1) and (2),}$$

where : $\hat{T}_{rsy,ps,g}$ is the total for partition group g, pseudo - stratum ps

Over all random subdivisions of PSUs and pseudo-PSUs in ps, the probability of any pair of PSUs or pseudo-PSUs being in partition group g is $N_g(N_g-1)/(N_{ps}(N_{ps}-1))$, where $N_{ps} = \sum_{g\in ps} N_g$.

By using the argument in (Cochran, 1997, pp. 266-267) for the Rao, Hartley, Cochran estimator and that the number of PSUs/ pseudo-PSUs (N_g) in a group g is a random process, an unbiased estimator for $V(T_{rsy})$ is:

$$\hat{V}(\hat{T}_{rsy}) = \sum_{h}^{H} \sum_{ps\in h} \left((\sum_{g\in ps} N_g^2 - N_{ps})/(N_{ps}(N_{ps}-1)) \right) \times \left(\sum_{i=1}^{N_{ps}} x_i^2/p_i - T_{ps}^2 \right)$$

An unbiased sample estimator for $\hat{V}(\hat{T}_{rsy})$ or $V(\hat{T}_{rsy})$, also from Cochran, $v(\hat{T}_{rsy})$, is:

$$v(\hat{T}_{rsy}) = \sum_{h}^{H} \sum_{ps\in h} \left((\sum_{g\in ps} N_g^2 - N_{ps})/(N_{ps}^2 - \sum_{g=1}^{2} N_g^2) \right) \times \left(\sum_{g=1}^{2} 1/2(2x_g/p_g - \hat{T}_{ps})^2 \right) \quad (3)$$

where x_g is the variable of interest for the sampled PSU in partition group g and p_g is its selection probability.

The second term of the product in (3) is the balanced half-sample variance estimate (BHR) for the pseudo-stratum. Therefore, any differences between (3) and BHR can be attributed to the first term in (3) (i.e., the scaling term).

The scaling term acts as a finite population correction (FPC). If the N_g 's are all equal in a stratum then this term resembles the simple random sample FPC. However, when the stratum PSUs are skewed in either direction, this term can be greater that 1. In this situation, the BHR estimator should be expected to underestimate the variance.

$v(\hat{T}_{rsy})$ will now be used to produce an unbiased bootstrap variance estimator. When computing N_g, a PSU/pseudo-PSU that spans two partition groups is included in both N_g counts.

4.0 Bootstrap Variance Estimator for $\hat{T}_{rsy}, V^*(\hat{T}_{rsy})$

The bootstrap variance estimator will be generated from a set of bootstrap samples. First, a discussion of the bootstrap sample size used in these samples, n_h^*, will be presented.

n_h^* is chosen so that $E^*(V^*(\hat{T}_{rsy})) = v(\hat{T}_{rsy})$, where E^* represents the expectation with respect to the bootstrap selection. There are two ways to do this:

The first way is to recognize that the sampling scheme proposed here, given a known set of N_g 's, has the same inclusion and joint inclusion probabilities, as well as the same estimator, as the Rao, Hartley, Cochran estimator. Hence, Sitter's (1992) solution to n_h^* can be used. One advantage here is that n_h^* will not be a function of the variable of interest. Therefore, once n_h^* is determined for one variable, across all possible randomizations, it should work for other variables, too. One disadvantage is that Sitter does not provide a closed form solution. Instead, a searching and bracketing process must be used. A possible second disadvantage is that the clustering in the selected sample is ignored.

The second solution is to use a simulation searching process to determine n_h^* that does not ignore the clustering. For this searching process, a number of stratum bootstrap variance estimates are generated, each with a different n_h^*. Each of the bootstrap variance estimates can then be compared to $v(\hat{T}_{rsy})$. A bracketing procedure can now be used to achieve an unbiased variance estimator. The disadvantage here is that this searching process is more involved than the first. However, since the cluster correlations, ρ_{hij}, across all randomizations, are zero, this solution should be reasonably close to the first solution.

In this paper, the second solution will be described and tested in a simulation. The simulation can then be used to verify that this solution, once solved for one variable, works equally well for all other variable.

$\hat{T}_{rsy}$ for stratum h will be denoted by $\hat{T}_h$.

4.1 The Bootstrap Procedure

1. Select a systematic PPS sample (s_h), as described in section 2.0, using the randomization methodology described in section 3.2.
2. Generate a bootstrap frame based on the selected sample s_h. For each selected PSU/pseudo-PSU j with sampling weight $w_j = 1/p_j$, generate bootstrap-PSUs (bj) by replicating the j^{th}

PSU/pseudo-PSU w_j times. Note w_j does not include the partitioning weight. The bj^{th} bootstrap-PSU has the following measure of size (m_{bj}):

$m_{bj} = I_{bj} \cdot 1/w_j$,

$$I_{bj} = \begin{cases} 1, \text{ if } bj \text{ is an integer component of } w_j \\ C_j, \text{ if } bj \text{ is a noninteger component of } w_j \\ \quad C_j \text{ being the noninteger component} \end{cases}$$

Associate j's pseudo-stratum with each of the bootstrap-PSUs generated from the j^{th} PSU

3. Within each stratum, define a set of bootstrap sample sizes, $n^*_{kh}, k = 1 \text{ to } K_h$:

   $n^*_{kh} = (n_h - n_{h0})(k-1)/(K_h - 1) + n_{h0}$, where n_{h0} is the lower bound for n^*_{kh}. n_{h0} must be chosen to provide a positively biased variance estimate.
4. Randomize the bootstrap-PSUs within each pseudo-stratum.
5. Choose an n^*_{kh}, say n^*_{0k} to be used to compute the first bootstrap variance.
6. The bootstrap frame, bootstrap frame ordering, measure of size (m_{bj}), and bootstrap sample size (n^*_h) have been specified. Using these quantities select B bootstrap samples using the same procedures used to select the original systematic PPS sample. The one exception to this is that a bootstrap-PSU generated from noncertainy PSUs that become certainty in the bootstrap selection should not be eliminated from the selection process and taken in sample with certainty. The bootstrap weight should properly reflect the bootstrap-PSUs selected multiple times (see 7 below). Before each selection, the bootstrap frame must be re-randomized.
7. For each bootstrap sample, compute a set of bootstrap weights, w^*_j. Compute T^*_{bh} like $\hat{T}_h$, using w^*_j instead of w_j.

   The bootstrap-PSU weight, w^*_j, is: $w^*_j = \sum_{bj \in S^B_j} w^p_{bj}$,

   S^B_j: is the set of all bj's generated from j that are selected in the B^{th} bootstrap sample, and

   $$w^p_{bj} = \begin{cases} I_{bj} \cdot M_{bj} / p_{bj}, \text{ if } bj \text{ is from a PSU} \\ I_{bj} \cdot M_{bj} / p_{bj} \times w_{j.}, \text{ if } bj \text{ is from a pseudo - PSU} \end{cases}$$

   M_{bj}: is the number of times the bj^{th} bootstrap- PSU is selected,

   p_{bj}: is the bootstrap selection probability for the bj^{th} bootstrap-PSU.

   $p_{bj} = m_{bj} / SI_h$, $SI_h = \sum_{bj \in s_h} m_{bj} / n^*_{kh}$.

   $w_{j.}$: is the partitioning weight for the selected bootstrap-PSU.

   When pseudo-PSUs are selected, a bootstrap-PSU weight can be generated by adding up the w^*_j's corresponding to the PSU.
8. The bootstrap variance for $\hat{T}_h$ given n^*_{kh} is:

   $V^*_k(\hat{T}_h | n^*_{kh}) = 1/(B-1) \sum_{b=1}^{B} (T^*_{bh} - \overline{T}^*_h)^2$,
9. Repeat steps 5-8, for each n^*_{kh}, generating $V^*_k(\hat{T}_{ih} | n^*_{kh})$ for $k = 1 \text{ to } K_h$.
10. Compute $v(\hat{T}_h)$ from sample s_h and compare it to each of the $V^*_k(\hat{T}_{ih} | n^*_{kh})$ for $k = 1 \text{ to } K_h$. Denote by M_h the stratum bootstrap variance with the smallest negative bias. Denote by P_h the stratum bootstrap variance with the smallest positive bias. Define $q_h = (v_h - M_h)/(P_h - M_h)$. Select a random number between 0 and 1. If it is less than or equal to q_h then use the replicate weights associate with P_h to produce future variances. Otherwise, use the replicate weights associated with M_h. Denote this variance estimator by $V^*(\hat{T}_h)$. This produces unbiased stratum variances because $E_q(V^*(\hat{T}_h)) = q_h P_h + (1-q_h) M_h = v(\hat{T}_h)$, where E_q represents the expectation with respect to the q_h selection. To reduce the instability introduced by the bracketing, P_h and M_h should be determined to be as close to zero bias as possible.

    Now, across all randomizations $V^*(\hat{T}_{rsy})$ is unbiased (i.e., $E(V^*(\hat{T}_{rsy})) = \sum_h Ev(\hat{T}_h) = \sum_h V(\hat{T}_h) = V(\hat{T}_{rsy})$)

5.0 Simulation

To demonstrate the advantages of the bootstrap variance estimator, a simulation study is presented comparing BHR and the bootstrap variance estimators. Five hundred simulations are generated using frame variables. In tables 1-6, estimates are computed by each stratification variable (affiliation, region and school level), as well as one of the sort variables (Urbanicity).

5.1 Comparison Statistics

Variance comparisons are based on the relative error of the standard error, relative mean square error of the variance and the 95% coverage rate.

5.2 Sample Design

Following the Schools and Staffing Survey sample design, the list frame component of NCES's Private School Survey (PSS) is stratified by detailed School Association (19 groups) by Census Region (4 levels), and by school level (3 levels). The school sample is selected systematically probability proportionate to size, using square root of the number of teachers, as the measure of size. Before sample selection, the schools are ordered by state, school highest grade, urbanicity, zip code, and school enrollment. One detailed association is simulated.

5.3 BHR Variances

The r^{th} school half-sample replicate is formed using the usual textbook methodology (Wolter, 1985) with 2 PSUs per stratum. When $n_h \geq 2$, PSUs are placed in pseudo-strata (see section 1.0), which are used as strata for estimating variances. This is the BHR without FPC variance. A second BHR variance estimate (BHR with FPC Adjustment) adjusts the first variance estimator by $1-P_h$, where P_h is the average of the selection probabilities for the selected units within stratum h.

5.4 Number of Replicates and Determining n_h^*

Forty-four and forty-five replicates have been used in the BHR and bootstrap variances, respectively. Total number of students is used to determine n_h^*.

6.0 Results

In terms of relative error, MSE, and coverage rates, tables 1-6 show that the bootstrap variance estimator is better than either of the BHR estimators.

The bootstrap and BHR variance estimates are different only in how they are scaled (see end of section 3.3). Therefore, deficiencies in the BHR estimates are due to the use of an incorrect scaling factor. Each table has examples where BHR produces a large underestimate of the variance. This shows that the correct scaling factor, used in the bootstrap, can be greater than 1 in practice.

The results indicate that the bootstrap performed well for every variable, even though the bootstrap sample size was based on a single variable (number of students). This demonstrates, as argued in the paper, that n_h^* is independent on the variable of interest.

7.0 Conclusion

In the past, the author has proposed using a bootstrap variance estimator when the PSUs are selected with a PPS systematic sampling scheme. With a non-random ordering of the frame, these bootstrap procedures can only be implemented using frame variables. To get around this problem, the PSS systematic selection scheme proposed here introduces a random element to the ordering, while maintaining much of the implicit stratification usually associated with systematic sampling. Now, an unbiased bootstrap variance estimator can be developed for any variable of interest without the dependence of the frame variables.

The simulation study presented in this paper demonstrates that the bootstrap variance estimator is better than the BHR variance estimator, even when a simple FPC adjustment is applied. This is true with respect to relative error, MSE, and coverage rates.

With respect to relative error, the bootstrap performs better than BHR because BHR is not scaled correctly. As such, BHR can provide either an underestimate or overestimate of the variance, depending on the distribution of the PSUs within the pseudo-strata.

When a survey measures one variable or a number of variables all highly correlated with each other, it is likely that a frame ordering exists for an efficient systematic sample using a non-random ordering. In this situation, the standard variance methodologies (e.g. BHR, jackknife, Taylor Series) can safely be used, since these methodologies will likely, but not necessarily, overestimate the variance. In situations where the variance overestimation is unacceptably large (e.g., when the sampling rates are high, the covariance terms are very negative or the sample sizes are small or any combination of these) then the bootstrap procedure in (Kaufman, 1998) may be useful.

When a survey measures a number of unrelated variables, determining an efficient frame ordering for all variables may be impossible. In this situation, using the standard variance methodologies, for the standard systematic sample, can be inappropriate because now variances may have a large underestimation problem. A safer alternative would be using the randomized systematic sampling procedures and bootstrap variance estimator proposed here. With these procedures, all variances can be appropriately estimated. Some estimates may be less efficient than a systematic procedure using a non-random ordering, but there will be no large systematic variance underestimation.

8.0 References

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

Kaufman, Steven (1998). "A Bootstrap Variance Estimator for Systematic PPS Sampling," *Proceedings for the Section on Survey Methods, American Statistical Association*, pp. 769-774. Alexandria, Va.: American Statistical Association.

Sitter, R. R. (1992). Comparing three bootstrap methods for survey data, *Canadian J. Statist*., 20, 135-154.

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

Table 1 -- % relative error, % relative mean square error and % coverage rates for the Bootstrap and BHR variance estimator for estimating Total Number of Students by Affiliation, Region, Level and Urbanicity

Estimate	Bootstrap			BHR without FPC Adjustment			BHR with FPC Adjustment		
	Rel. Error	Rel. MSE	Cov. Rate	Rel. Error	Rel. MSE	Cov. Rate	Rel. Error	Rel. MSE	Cov. Rate
Other Affil.	-2.3	23.8	92.7	3.2	25.8	92.9	0.2	23.4	92.7
Northeast	-1.9	46.8	91.9	-10.2	30.0	78.6	-12.4	31.9	78.6
Midwest	2.7	46.8	94.6	54.9	149.4	100.0	50.9	137.5	100.0
South	-7.8	28.3	92.3	-10.6	29.2	100.0	-14.0	32.5	99.8
West	-4.2	34.2	93.5	6.4	31.3	92.9	3.9	28.1	92.9
Elementary	0.9	37.0	93.1	2.2	30.1	99.8	0.0	28.7	99.8
Secondary	6.3	43.4	93.8	-5.7	36.6	92.9	-15.5	39.1	85.6
Combined	-2.5	25.2	93.8	5.6	36.8	99.8	2.2	33.0	99.8
Rural	3.0	31.0	96.4	-5.1	29.7	85.9	-7.9	30.3	85.9
Suburban	-2.6	27.5	92.3	17.9	58.2	92.9	14.6	51.3	92.9
Urban	4.9	33.0	95.2	6.1	37.2	93.1	3.1	33.7	93.1

Table 2 -- % relative error, % relative mean square error and % coverage rates for the Bootstrap and BHR variance estimator for estimating the Total Number of Schools by Affiliation, Region, Level and Urbanicity

Estimate	Bootstrap			BHR without FPC Adjustment			BHR with FPC Adjustment		
	Rel. Error	Rel. MSE	Cov. Rate	Rel. Error	Rel. MSE	Cov. Rate	Rel. Error	Rel. MSE	Cov. Rate
Other Affil.	-4.1	25.0	91.7	-11.1	26.8	85.7	-13.5	29.7	85.7
Northeast	-2.3	41.4	91.5	-5.2	40.7	85.7	-7.1	40.7	85.7
Midwest	4.9	44.3	94.6	27.4	71.3	100.0	24.4	64.3	100.0
South	-7.7	31.4	90.9	-1.0	23.1	85.6	-4.8	23.4	85.6
West	-8.3	36.0	90.3	3.9	32.4	93.1	1.4	30.1	93.1
Elementary	-0.8	36.9	91.7	-26.1	52.4	71.5	-27.4	53.8	71.5
Secondary	-3.0	57.8	90.1	23.1	87.5	100.0	13.0	66.8	100.0
Combined	-4.6	26.5	91.7	13.3	45.6	100.0	9.9	39.3	100.0
Rural	-7.2	27.6	91.3	35.8	93.8	100.0	32.5	84.8	100.0
Suburban	2.4	30.1	94.0	-17.1	35.4	85.9	-19.2	38.2	85.9
Urban	-2.2	25.2	93.1	-8.7	29.8	92.9	-11.1	31.6	92.9

Table 3 -- % relative error, % relative mean square error and % coverage rates for the Bootstrap and BHR variance estimator for estimating the Total Number of Teachers by Affiliation, Region, Level and Urbanicity

Estimate	Bootstrap			BHR without FPC Adjustment			BHR with FPC Adjustment		
	Rel. Error	Rel. MSE	Cov. Rate	Rel. Error	Rel. MSE	Cov. Rate	Rel. Error	Rel. MSE	Cov. Rate
Other Affil.	-4.5	24.6	93.1	-15.5	35.3	93.1	-18.0	38.2	93.1
Northeast	-2.7	39.5	91.9	-8.4	37.6	85.9	-10.8	38.3	85.9
Midwest	4.2	38.3	94.8	35.7	94.0	100.0	32.2	84.5	100.0
South	-11.7	31.7	89.7	-12.5	30.3	85.7	-15.9	34.3	85.7
West	-4.7	39.3	91.7	20.0	67.3	93.1	16.9	60.5	93.1
Elementary	-3.2	31.6	93.5	-29.7	53.7	78.4	-31.1	55.3	78.4
Secondary	7.5	45.6	94.8	7.8	42.9	100.0	-2.9	31.9	92.9
Combined	-1.7	27.4	95.0	15.5	52.5	93.1	11.9	45.5	93.1
Rural	5.6	32.2	94.0	4.9	23.5	93.1	1.7	20.4	93.1
Suburban	0.9	27.4	92.7	4.8	26.9	92.9	1.6	23.8	92.9
Urban	9.9	37.0	96.2	49.2	138.9	93.1	44.7	125.5	93.1

Table 4 -- % relative error, % relative mean square error and % coverage rates for the Bootstrap and BHR variance estimator for estimating the Number of Students per School by Affiliation , Region, Level and Urbanicity

Estimate	Bootstrap			BHR without FPC Adjustment			BHR with FPC Adjustment		
	Rel. Error	Rel. MSE	Cov. Rate	Rel. Error	Rel. MSE	Cov. Rate	Rel. Error	Rel. MSE	Cov. Rate
Other Affil.	-4.1	23.3	94.2	-7.3	27.8	85.7	-9.9	29.4	85.7
Northeast	1.2	43.1	93.5	-16.6	37.8	71.7	-18.6	40.1	71.7
Midwest	5.8	43.6	94.6	53.2	147.6	100.0	49.2	135.5	100.0
South	-8.6	28.6	92.3	-13.7	29.6	92.9	-17.0	34.0	92.4
West	-5.2	35.9	92.9	13.7	50.3	85.9	10.9	45.2	85.9
Elementary	-2.9	31.3	91.7	-13.8	40.6	92.7	-15.6	41.6	92.7
Secondary	4.1	43.8	95.2	0.4	40.9	92.9	-9.9	36.6	92.9
Combined	-2.8	25.3	93.5	19.6	63.7	100.0	15.8	55.7	100.0
Rural	2.2	41.2	95.0	-13.3	36.1	85.9	-15.8	38.0	85.9
Suburban	-2.5	30.9	92.5	-30.0	53.1	79.0	-32.0	55.4	79.0
Urban	-3.3	31.2	93.7	-32.2	57.1	71.9	-34.3	59.4	71.9

Table 5 -- % relative error, % relative mean square error and % coverage rates for the Bootstrap and BHR variance estimator for estimating the Number of Teachers per School by Affiliation , Region, Level and Urbanicity

Estimate	Bootstrap			BHR without FPC Adjustment			BHR with FPC Adjustment		
	Rel. Error	Rel. MSE	Cov. Rate	Rel. Error	Rel. MSE	Cov. Rate	Rel. Error	Rel. MSE	Cov. Rate
Other Affil.	-4.3	24.0	92.9	-16.2	34.8	93.1	-18.5	37.7	92.7
Northeast	-2.0	45.7	92.3	-15.4	38.8	85.7	-17.3	40.6	85.7
Midwest	7.9	49.0	94.2	32.7	87.3	100.0	29.3	78.6	100.0
South	-10.3	31.2	91.5	-5.8	24.4	92.7	-9.5	26.8	92.7
West	-5.8	35.9	92.5	18.9	62.7	93.1	15.9	56.1	93.1
Elementary	-2.1	35.3	93.5	-30.4	55.6	71.5	-31.7	57.1	71.5
Secondary	2.3	44.7	94.8	16.3	70.1	100.0	5.1	49.2	100.0
Combined	-3.3	26.1	94.0	11.3	41.9	93.1	7.8	36.2	93.1
Rural	2.1	39.4	93.5	16.8	57.2	100.0	13.5	50.4	100.0
Suburban	0.9	33.2	92.9	-23.5	43.8	79.0	-25.6	46.6	79.0
Urban	-2.0	31.5	94.6	-25.5	48.9	79.0	-27.8	51.5	79.0

Table 6 -- % relative error, % relative mean square error and % coverage rates for the Bootstrap and BHR variance estimator estimating the Student/Teacher Ratio by Affiliation , Region, Level and Urbanicity

Estimate	Bootstrap			BHR without FPC Adjustment			BHR with FPC Adjustment		
	Rel. Error	Rel. MSE	Cov. Rate	Rel. Error	Rel. MSE	Cov. Rate	Rel. Error	Rel. MSE	Cov. Rate
Other Affil.	-0.4	30.8	93.8	14.8	40.7	100.0	11.6	34.5	100.0
Northeast	-5.7	65.7	90.5	-8.8	47.9	85.7	-10.9	47.4	78.6
Midwest	4.4	83.8	94.0	17.2	99.1	92.9	14.4	93.9	92.9
South	-1.1	35.5	92.7	20.8	69.1	92.9	16.2	59.2	92.9
West	-6.2	35.0	91.5	14.7	56.2	93.1	12.1	51.3	93.1
Elementary	5.4	55.2	93.5	-14.4	41.1	79.0	-16.1	42.4	79.0
Secondary	0.0	36.5	90.9	12.2	45.6	92.9	1.5	29.6	85.6
Combined	-2.3	29.8	93.8	-8.1	29.6	92.9	-11.1	31.4	92.9
Rural	0.8	52.4	93.8	11.9	86.8	100.0	8.2	78.8	100.0
Suburban	1.4	40.8	95.0	91.3	296.1	100.0	86.2	276.2	100.0
Urban	0.2	37.5	94.8	-12.7	32.6	85.9	-15.2	35.0	85.9

VALIDATING ITEM RESPONSES ON SELF-REPORT TEACHER SURVEYS

John E. Mullens, Mathematica Policy Research, Inc.
Daniel Kasprzyk, National Center for Education Statistics
John E. Mullens, Mathematica, Inc., 600 Maryland Ave, SW, Suite 550, Washington, DC 20024

Key Words: Instructional Processes, Self Report, Pretesting, Validity, Reliability

This paper describes survey development work being done by the Elementary and Secondary Sample Survey Studies Program of the National Center for Education Statistics (NCES) to develop items and processes to collect accurate, valid, and meaningful data on the instructional processes teachers use with students. Instructional processes (IP) data can illuminate how teachers and students work together to approach classroom educational objectives: the emphasis teachers put on important topics within the curriculum, the learning objectives they have for their students, the activities in which students and teachers engage, and the ways in which teachers assess student learning. This information is important for national, state, and local policymakers and others interested in how school reform affects classroom practice. This article reports on the NCES Classroom Instructional Processes Study, conducted in 1997-98 and more completely described in Mullens & Gayler (1999). That work is one of a number of activities through which NCES is both collecting IP information and examining, refining, and improving the quality of data collection methods and instruments they use in national data collection programs.

NCES became interested in this line of data collection in 1994 when Commissioner Emerson Elliot authorized a comprehensive review of then-current research efforts (Leighton, Mullens, Turnbull, Weiner, & Williams, 1995), an analysis of measurement approaches (Mullens, 1995), and development of a module of items to measure IP for the Current Teacher's Questionnaire of the SASS 1994-95 Teacher Followup Survey. Following that data collection, NCES continued to fund IP item development and refinement.

Why Measure Instructional Processes

Societal demands on schools and teachers and the resulting close scrutiny of educational outcomes continue to heighten interest in how schools and teachers can better do their jobs. The desire to understand variation in student outcomes leads policymakers and researchers to seek a better understanding of how teachers and students approach the math curriculum. How do teachers approach math instruction? Do teachers use different techniques when they emphasize broad concepts compared to specific facts or processes? To what extent do teachers use newly-recommended instructional techniques? Do they integrate new techniques with their "tried-and-true" methods? To the degree that differences in teachers' instructional practices directly affect the quality of learning in the classroom, answers to such questions will inform our understanding of effective approaches to student learning.

When IP data are combined with information on student learning, policymakers, teacher trainers, and professional developers have the means to guide instructional techniques toward those that are most effective in creating desired student outcomes. Understanding how variation in student learning relates to variation in instructional methods could inform local, state, and national education policy (Burstein, Oakes, Guiton, 1992; Smith, 1988; Murnane, 1987).

Stodolsky (1996) summarized the rationale for generating a broadly representative, yet finely-textured data base of information about classroom-level IP:

> "If we are to understand, monitor, and improve our nation's schools, accurate and timely empirical, descriptive data about how schools' work must be available. The activities that take place in classrooms to engender student learning and development are the heart of any school's education efforts. It is in the transactions between and among teachers, students, materials, and tasks that deliberate efforts to educate occur. Descriptive information about how teaching and learning occur in classrooms and about what is taught provides the basis for monitoring the status of instruction in a large number of settings. Such information can provide periodic assessments of stability and change in instruction, particularly as changes relate to deliberate efforts to reform or alter curriculum and instruction."
>
> (Stodolsky, 1996)

Survey data are likely to be the major source of nationally-representative information about instructional content and practices, but there are questions about the quality of such data.

Potential Threats to the Reliability and Validity of Self-reported IP Data

While well-designed focused surveys can be cost-effective for administrators and place only limited burden on respondents, the accuracy of self-reported responses sometimes calls into question the reliability and validity of the resulting data. There are at least three reasons why this might be so (Mayer, 1999). First, teaching and learning in any context is a complex human endeavor that cannot yet be adequately represented by responses to a single survey. Second, some survey items may contain unknown phrases or ambiguous concepts that make an appropriate response difficult. Finally, for reasons beyond a survey's scope, some teachers may be sensitive to particular questions and/or the concepts they represent and therefore feel pressured to provide (perhaps socially desirable) responses that are less than accurate. These and possibly other equally serious concerns pose serious threats to using surveys to accurately portray instructional practices. Therefore, the quality of the survey items needs to be initially validated and periodically confirmed (Burstein, McDonnell, Winkle, Ormseth, Mirocha, Guiton, 1995).

The Fieldtest

To explore these possible threats to the reliability and validity of the self-reported data, this fieldtest set out to determine the accuracy of teachers' descriptions of classroom instruction when recorded on a daily basis and over one semester. It included a mail questionnaire sent to approximately 400 math teachers of eighth to twelfth grade students and a case study of 41 teachers in similar settings. Case study teachers were volunteers and received no financial incentive to participate in this study. Mail survey respondents described their instruction in one designated math course over the previous semester; case study teachers responded to the same questionnaire about a designated math course, were observed teaching, and kept logs of daily instructional activities in that course over a four week period. Mail responses were used to assess the adequacy and scope of items and response options; case study data were used to examine the reliability and validity of those same teachers' questionnaire responses. For reasons of space, this article discusses the case study fieldtest data only.

The case studies had five parts: a mail questionnaire, classroom observation, teacher interview, daily classroom logs, and a second administration of the questionnaire. At the beginning of the case study process, participating teachers completed the IP survey about the most recent semester. A researcher observed a class period in each teacher's designated class, recording on a log form the instructional objectives, classroom activities of the teacher and student, and the use and availability of instructional materials. Teachers completed a classroom log form about the same class and discussed the class and their questionnaire responses during a subsequent interview. Every day for four weeks, classroom teachers recorded their activities and those of their students. At the conclusion of the case study period, teachers completed a second questionnaire, identical to the first.

Building on previous work. The fieldtest built on the findings and recommendations from a previous NCES pilot project and on other earlier studies including the Third International Math and Science Study (1998, 1996), the UCLA/RAND Validating National Curriculum Indicators project (1995), and Reform Up Close (1993). The previous pilot project fieldtested a draft questionnaire with 111 eighth to tenth grade teachers in three districts (Mullens & Kasprzyk, 1996a). Results from that project and from subsequent experience with instructional practice items on the 1995 Teacher Followup Survey guided our questionnaire revisions and planning for this fieldtest. When refining items, we also built on the early TIMSS (1996) work developing items on IP and on Andrew Porter's (1993) work identifying effects of increased enrollments on the content and pedagogy of high school math and science courses. When designing the fieldtest, we drew heavily on the prior work of Burstein, McDonnell, et. al. (1995) developing validation procedures to improve the quality of national indicators of curriculum.

Fieldtest goal. The goal of the project was to collect information about the accuracy and reliability of self-reported data on the instructional practices of secondary math teachers and the contexts within which they occur. The items collected information on four areas of instructional practice: a) conditions for teaching and learning in the school and classroom, b) course content and emphasis, c) instructional activities, and d) the availability and use of instructional resources.

Fieldtest design. We conducted case studies during April and May 1997 in six geographic areas designed to attain some measure of dispersion yet limit travel costs: Baltimore City, Frederick, and Hagerstown, Maryland; Austin, Texas; Charleston, South Carolina; Milwaukee, Wisconsin; and Aberdeen, Bremerton, and Olympia, Washington. Fifty teachers identified one course (the "designated class") for which they were willing to be observed and to record classroom activities daily for four weeks. Together the courses covered the curriculum spectrum from eighth grade mathematics to Calculus. Forty-one of the 50 volunteers ultimately completed the case studies.

Items on instructional techniques were the core of the questionnaire. Those items asked teachers to indicate the frequency and duration with which they used various instructional methods in a single targeted class. Activities included those commonly associated with traditional teaching (such as lecture and student recitation or drill), those reflecting reform recommendations (such as student discussions of problem solving approaches), and some common to a range of styles (such as giving tests). Other items asked teachers to describe their use of student activities, which were similarly distributed among instructional approaches.

The fieldtest had two limitations. First, while the questionnaire collected information covering a full semester of instruction, the design of the case study portion included data collection on only four weeks of that semester. Ideally, the two time periods would have been identical and we could have used a semester's worth of log data with which to validate questionnaire responses. The decision to collect only four weeks of log data reflected project funding limitations. Additionally, although 41 case study teachers completed four weeks of daily logs, only 20 completed the second questionnaire. Thus the analysis of teachers' responses on the two surveys was limited to those 20 sets. We think this low response was caused by the lateness in the school year. We have no reason to believe that the teachers who returned the second questionnaire were different from the non-responding teachers in some systematic way that might bias our interpretation of their responses.

Analysis of Fieldtest Data

At the conclusion of the case studies, we used the two questionnaires, the teacher logs, and the researchers' logs to investigate the reliability and validity of the questionnaire items. Among other analyses, we examined:

- percent teacher and researcher agreement on the occurrence of student learning objectives and instructional activities to understand the extent to which teachers and researchers shared a similar understanding of the concepts in question;

- percent teacher and researcher agreement on the length of time the objective or activity occurred to understand the extent to which teachers and researchers shared a similar conception of elapsed time;

- percent agreement between case study teachers' responses to the first and second questionnaires on the frequency and duration of classroom instructional practices to understand the extent to which survey responses completed six weeks apart are the same.

Fieldtest Results

With few exceptions, fieldtest data suggest that the case study teachers interpreted key words describing instructional processes in ways that were consistent with the independent observers. Teachers also had the same sense of the passage of time as observers when recording that information. Where low rates of agreement occurred, they reflected differences of opinion between teachers and observers about what constituted "whole class discussions," "practice or drill," and "several appropriate answers or approaches."

Determining the validity of teachers' daily descriptions of classroom instruction. To assess the accuracy with which teachers described on the log form the learning activities they orchestrate on a daily basis, we compared teachers' recordings of classroom activities on the daily log to the researcher's record on the observation form. If the items, teachers, and observers were each perfect, we could expect a 100 percent match on the occurrence and duration of all student learning objectives and instructional activities.

Items on student learning objectives. Independent observers validated 79 percent of teachers' recordings of the learning objectives used that day in their class, and agreement between teachers and observers was greater than 75 percent for four of seven objectives analyzed. The lowest agreements were for memorizing facts, definitions or formulae (66 percent), recognizing and solving story problems with unfamiliar or complex structures (71 percent), and building and revising theories (73 percent). Where nonagreement occurred, teachers were more likely than observers to report that a learning objective had been part of the observed lesson.

Teacher/observer agreement appeared to vary by the degree to which the objective was observable by a classroom visitor or was explicitly stated by the teacher to the class or to the observer. For example, it was usually clear to the observer when students were doing mathematical operations, but often difficult to observe that students were memorizing or were expected to be memorizing. The learning objectives with the lowest rates of agreement, those generally less visible and more difficult to detect, may indeed have occurred but were simply not observed.

Teachers' estimates of the time spent on learning objectives were substantially verified by observers: teachers and observers strongly agreed on the minutes allocated toward the student learning objectives that occurred during that class period. In those instances where observers did not agree with teachers about the elapsed time, there was no clear pattern to the mismatches: teachers indicated either more or less time than the observers noted.

Items on teacher actions. Case study data show strong agreement between teachers and observers on the occurrence and duration of teachers' instructional activities. Teachers and observers agreed on 85 percent of all teacher activities occurring during all the lessons. In seven of the eight activities, agreement between teachers and observers about whether the activity occurred was 75 percent or greater. The highest rates of agreement between teachers and observers were for lecturing (98 percent) and providing individual or small group tutoring (95 percent); the lowest agreement was for stimulating student discussions of approaches to solving problems or explanations of their mathematical thinking (55 percent). Where there was nonagreement about an activity, teachers were more likely to report that it did happen than were observers. In 94 percent of the instances in which teachers and observers saw teacher activities differently, teachers indicated the activity had occurred and observers indicated they had not seen it.

We found a high level of agreement between teachers and observers on the minutes spent on each teacher instructional activity that occurred during the observed class period. Teachers and observers substantially agreed on the duration of all teacher activities except demonstrating a concept using two dimensional graphics.

Items on student activities. Teachers and observers agreed on 82 percent of all student activities recorded during the observed lessons. Agreement between teachers and observers on whether specific student learning activities occurred was 75 percent or greater for 13 of the 18 student activities included. High rates of agreement were recorded when students: listened to the teacher (100 percent), worked individually on exercises (93 percent), worked in small groups (93 percent), and worked on assignments due the next day (85 percent). Student activities with the lowest agreement between teacher and observer were the following: participate in whole-class discussion (56 percent), practice or drill on computational skills (63 percent), solve problems for which there are several appropriate answers or approaches (71 percent), and wait for completion of non-academic tasks (71 percent).

The low agreement rates for these activities reflect the gist of discussions following the observed classes in which teachers and observers reported differences of opinions on what constituted the first three activities. The majority of all nonagreements between teachers and observers on these items arose because the teacher saw the event as occurring but the observer did not: teachers indicated that student discussions involved the whole class, while observers were more likely to say that only a few students were actively involved; teachers thought students were drilling on basic skills, but observers saw no evidence; teachers more often said after class that they emphasized several approaches to a problem, while researchers observed only one.

There was strong agreement between teachers and observers on the length of time each student activity occurred during the observed class, ranging from 86 to 100 percent agreement per student activity. In the few instances where there was no agreement, there was also no pattern in the direction of nonagreement: observers used a clock or watch to record time as the activities occurred; teachers retrospectively over- and under-estimated elapsed time nearly equally.

Summary. Case study teachers' accounts of the student learning objectives, teachers' actions, and student activities occurring in the teachers' observed classes were substantially validated by the accounts of classroom observers on 24 of 33 items. Teachers' accounts of the length of time that student learning objectives were taught and that teachers and students engaged in activities were both substantially validated by independent observers on every item. Across the three types of items, teachers' time estimates were most accurate on those activities they used most frequently.

Determining the reliability of teachers' questionnaire responses. To assess the reliability with which teachers describe on a one-time questionnaire what they do throughout a semester, we compared teachers responses on the first questionnaire to their responses on the second questionnaire administered six weeks later. We assumed that the two sets of responses would be identical if their first responses were accurate, if their implementation of instructional practices had not changed, and their opinions about their teaching had remained the same.

Items on student learning objectives. All of the nine subitems collecting information on student learning objectives had rates of agreement within one response

option between the first and second questionnaire higher than 78 percent. There was high agreement on the frequency with which teachers' instructional objective was to have students understand concepts, relationships, theorems (100 percent); perform mathematical operations or execute algorithms (95 percent); and solve story problems with familiar structures (90 percent). The learning objective building or revising theories had the lowest agreement (79 percent) and, except for collecting data (by observing, measuring, or counting), was also the least used instructional objective during the case studies, according to log records.

Items on teacher actions. Ten of the twelve subitems assessing the frequency with which teachers use certain instructional techniques showed correspondence between teachers' responses on the two questionnaires at rates above 75 percent. Interestingly, the two items with low rates of agreement between questionnaires, leading students in recitation and drills and teacher time spent working on administrative tasks, are both forms of teaching considered to be more traditional. Additionally, teachers' responses on the typical length of time spent per class period on each instructional activity all showed more than 75 percent agreement between the two questionnaires.

Items on student activities. Of the 24 subitems assessing the frequency with which teachers have students engage in particular learning activities, 22 showed high correspondence between teachers' responses on questionnaires 1 and 2. The two items with low rates of agreement were practice or drill on computational skills (67 percent) and solving problems with more than one appropriate solution (74 percent). All 24 time-per-typical-use items had agreements greater than 75 percent.

Summary. Case study teachers' responses on the two questionnaires were substantially the same on 41 of the 45 items describing the student learning objectives and instructional activities used in the teacher's designated class.

Conclusions

These results suggest that teachers in disparate locations recognize and accurately interpret the named classroom activities, except for some glaring exceptions. Respondents' indications about whether or not certain activities occur (and for how long) coincide with those of independent observers, for the most part. Teachers' questionnaire responses about the math instructional practices they use (and for how long) are pretty reliable within one response option: teachers respond the same way to most questions on questionnaires administered six weeks apart.

The good news from this analysis is that we are confident that teachers recognize and identify most instructional activities in ways similar to the observers; they readily acknowledge their use of recitation and drill and admit to working on administrative record keeping tasks while their students wait, even though those activities may be out of favor with school reform advocates. When completing questionnaires administered six weeks apart, we know that teachers' responses to questions about instructional activities are consistent.

We suspect that three reasons may have contributed to occasional low levels of agreement between teachers and observers. Some items may have low agreement between teacher and observers because classroom observation itself is limited in its capacity to capture certain elements of classroom instruction. This is particularly true for unobservable instructional objectives such as memorization. Validation of items may also have been affected by the differences of opinion on the scale of classroom activities and was most noticeable on items that distinguish among number of participants, such as "whole-class discussion." Differences of opinion contributed to limited validation in other ways as well, such as when teachers and observers disagreed on whether problems had "more than one approach."

Recommendations

These conclusions suggest considerations for future questionnaire and fieldtest designs that may confirm and further our understanding of the accuracy and reliability with which teachers respond to self-report surveys. First, this fieldtest provided the strong basis for refining those few instructional practice items where wording that appeared unambiguous in pilot testing was subject to varying interpretation in wider use. Through fieldtesting, we identified specific issues for those select items that can now be reworded. Second, the next generation of designs might include multiple and concurrent techniques to validate the accuracy of teachers' descriptions of their daily instruction, especially information on student learning objectives whose occurrence can not be visibly or aurally confirmed by independent observers. This may entail active observation by researchers, triangulation by multiple observers, or teacher/observer/student interactions to augment first-person observation. Finally, future efforts to validate questionnaire items would be well-served to equalize case study design lengths with the item referent periods: as long as a semester (equaling the referent period of the items tested here) or as short as two weeks. Shorter referent periods (with corresponding

validation periods) are likely to result in more accurate responses by teachers, allow instructional variation across a large number of participants, and provide reliable data with which to estimate response accuracy.

We used these findings to modify items, to reduce ambiguity in problematic items, to identify particularly reliable and valid items, and to recommend a strong module of instructional process items for the 1999-2000 Schools and Staffing Survey.

References

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

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.

Leighton, M., Mullens, J., Turnbull, B., Weiner, L., & Williams, A. (1995). Measuring instruction, curriculum content, and instructional resources: The status of recent work (NCES 1995-11). U.S. Department of Education. Washington, DC: NCES Working Paper.

Mayer, D. (1999). Measuring Instructional Practice: Can Policymakers Trust Survey Data? Educational Evaluation and Policy Analysis, 21(1), 29-45.

Mullens, J. (1995). Classroom instructional processes: A review of existing measurement approaches and their applicability for the Teacher Followup Survey. (NCES 1995-15). U.S. Department of Education. Washington, DC: NCES Working Paper.

Mullens, J. & Gayler, K. (1999). Measuring classroom instructional processes: Using survey and case study fieldtest results to improve item construction (NCES 1999-08). U.S. Department of Education. Washington, DC: NCES Working Paper.

Mullens, J. & Kasprzyk, D. (1996a). Using qualitative methods to validate quantitative survey instruments. In 1996 Proceedings of the Section on Survey Research Methods. Alexandria, VA: American Statistical Association, 638-643.

Mullens, J. & Kasprzyk, D. (1996b). The Schools and Staffing Survey: Recommendations for the future (NCES 1997-596). U.S. Department of Education. Washington, DC: NCES Working Paper.

Mullis, I., Martin, M., Beaton, A., Gonzalez, E., Kelly, D., & Smith, T. (1998). Mathematics and science achievement in the final year of secondary school: IEA's third international mathematics and science study. Boston, MA: Center for the study of testing, evaluation, and educational policy, Boston College.

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

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.

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

Schmidt, W. (1996). Indicators of opportunity to learn in the Third International Mathematics and Science Study: What is the impact of this OTL information on U.S. public schools? Paper prepared for AERA Annual Meeting.

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

Stodolsky, S. (1996). Should SASS measure instructional processes and teacher effectiveness? In J. Mullens & D. Kasprzyk (Eds.) The Schools and Staffing Survey: Recommendations for the future (NCES 1997-596). U.S. Department of Education. Washington, DC: NCES Working Paper.

Acknowledgment: The authors would like to thank Daniel P. Mayer of Mathematica Policy Research, Inc. for his comments on an earlier draft.

COVERAGE IMPROVEMENT OF TERMINAL KINDERGARTEN PROGRAMS FROM ALTERNATIVE SOURCES IN THE PRIVATE SCHOOL SURVEY

Nancy R. Johnson, Charles H. Edwards, Courtney L. Daniels, Randall J. Parmer
Nancy R. Johnson, U.S. Census Bureau, Washington, DC 20233[1]

Key Words: Survey Frame Development, Coverage Improvement, Private Schools

I. Introduction

This paper discusses efforts to represent kindergarten programs in the universe of private elementary and secondary schools developed by the U.S. Census Bureau for the National Center for Education Statistics (NCES). This private school universe was initially developed in 1987, and was subsequently updated six times with the seventh update currently in progress.

The private school universe before 1993 included kindergarten programs only if they were a part of a school containing any of grades 1-12. In 1993 the NCES asked the U.S. Census Bureau to expand the private school universe to include programs where kindergarten is the highest grade. We call these programs K-terminals. (For more details on K-terminals, see the explanation of Early Childhood Care programs and K-terminals in section II.)

This paper provides details of an alternative frame approach in the Private School Survey (PSS) that significantly increased our coverage of K-terminal programs (and, to a lesser extent, the coverage of programs containing a kindergarten and one or more higher grades) during the last two universe list frame updating operations.

II. Background

Every two years the PSS (which is actually a census) collects and reports data on private schools that are comparable to public school data. It builds as accurately and completely as possible a universe list of private schools from which we can select samples for a variety of NCES surveys of private schools.

The private school universe is defined as including all private schools that provide instruction for at least one of grades 1-12, have one or more teachers, are not administered by a public agency, and are not operated in a private home. This traditional private school universe consists of two coverage improvement operations -- List Frame updating and Area Search Frame updating. List Frame updating is a national coverage improvement operation designed to locate private elementary and secondary schools not already on the existing private school universe. The updating operation uses lists from private school associations, the 50 states plus Washington, D.C., and sometimes private vendors. Area Search Frame updating is a coverage improvement operation consisting of an independent search in a nationally representative sample of counties. This operation locates private schools still missing from the private school universe after completion of the list frame updating.

In 1993-94, we began to collect information on K-terminal school programs and to build a K-terminal frame for use in NCES surveys interested in these types of private schools, specifically the Early Childhood Longitudinal Survey (ECLS). As lists were collected from the states and associations, we identified and separated those programs that contained, at most, a kindergarten. (Programs with a kindergarten and one or more higher grades went into the traditional list frame.) A kindergarten (KG) is a program providing instruction primarily for 5-year-old children who will enter first grade in the upcoming school year. This includes transitional KGs and transitional first grades if these children are expected to enter first grade upon completing these programs. Some of these K-Terminal programs may contain nursery or preschool age children.

In the 1995-96 list frame updating operation we added an alternative frame approach that collected lists of early childhood care programs from state day care licensing agencies and private child care resource and referral organizations. An Early Childhood Care program (ECC) is a center-based program for young children (generally 3- to 5-year-olds). These include preschools, prekindergartens, day care centers, nurseries, and other early childhood programs. A number of ECCs also provide a KG and, sometimes, higher grades. For the PSS, only programs containing a KG (or higher) grades are in-scope.

We felt this alternative frame would improve the coverage of K-terminal programs in the list frame private school

1 This paper reports the results of research and analysis undertaken by Census Bureau staff. It has undergone a more limited review than official Census Bureau Publications. This report is released to inform interested parties of research and to encourage discussion.

universe and, to a lesser extent, the coverage of schools containing a KG and one or more higher grades.

III. List Frame Kindergarten Coverage Rates

Table 1 shows the list frame coverage of K-terminal programs, programs with KG and a higher grade, and the overall KG coverage for three of the private school universe updating operations: 1993-94, 1995-96, and 1997-98. The coverage rate is the ratio of list frame schools to the combined list frame and area frame schools and indicates the extent of undercoverage of schools in the list frame.

Table 1: Percent Coverage of List Frame Schools with Kindergarten

	1993-94	1995-96	1997-98
K-terminal	32%	57%	87%
KG and higher grade	90%	93%	95%
Overall KG	73%	83%	93%

Our initial effort in 1993-94 produced a coverage rate of 32% for the K-terminal programs in our list frame. In the 1995-96 updating operation, we added an alternative frame that improved the coverage of K-terminal programs to about 57% for the list frame. Additional improvements in the 1997-98 list frame updating operation improved the coverage of K-terminal programs to about 87% of our area frame universe estimate.

The coverage of programs in the list frame containing KGs with one or more higher grades made a small, yet significant increase. In 1993-94 the universe had a coverage of 90% of the KGs with a higher grade. This rate increased to 93% in 1995-96 and to 95% in 1997-98.

Of course, the overall KG coverage increased as well. In 1993-94 the overall coverage of KG programs on the list frame universe was 73%. In 1995-96 the overall KG coverage rate increased to 83%. This coverage rate increased to 93% in 1997-98.

Note on Sampling Error - The standard error is a measure of the variability due to sampling when estimating a parameter. It indicates how much variability there is in the population of possible estimates of a parameter for a given sample size. For PSS, only the area frame contributes to the standard error. The list frame component of the standard error is always zero (0). We did not compute standard errors for all of the area frame statistics presented here. In 1995-96, when we first computed it, the standard error for the K-terminals was about 4%. The KG and higher standard error was about 1%. Standard errors for 1997-98 are not available.

IV. The 1993-94 PSS List Frame Updating

The 1993-1994 PSS was the initial attempt to add K-terminal programs to the private school universe. For this first effort, we made only minor modifications to our traditional procedures for collecting private schools.

The 1993-1994 strategy involved collecting both pre-kindergarten (pre-K) and KG programs in the list frame, reinstating out-of-scope programs (i.e., programs identified as pre-K or KG in 1991-92), and collecting pre-K and KG programs in the area frame. We hypothesized that some programs identified as pre-Ks may also include KG instruction so we collected pre-Ks to test this theory. An analysis of the 1993-94 results indicated that so few of the pre-Ks actually have a KG that it was not worthwhile to collect and include pre-Ks on the universe. The three parts of our strategy are described next.

When we contacted the sources (states and associations) requesting their lists of private schools, we asked them to include programs where kindergarten is the highest grade in addition to the traditional schools containing grades 1-12 that they had provided in previous PSS cycles. We then processed these pre-K/KG programs similar to the way we processed traditional list frame private schools. Using this strategy, we obtained about 1,900 KG program births from the traditional list frame.

The second part of our strategy was to reinstate the out-of-scope programs from the 1991-92 PSS that contained grades pre-K or KG. Previously, the PSS identified these programs as out-of-scope and excluded them. We believed that pre-K programs from previous years had the potential to contain a KG in the current year. By reinstating these programs, we obtained about 600 in-scope KG programs out of about 2,500 1991-92 out-of-scope cases.

The third part of our strategy was to collect pre-K/KG programs in the area frame. In our area frame operation, we identified approximately 17,000 programs as early childhood care programs containing a pre-K or KG. This unexpected workload was unmanageable from both budgetary and timing points of view so, for data collection, we sent questionnaires to a subsample of 5,979 of these programs. After tabulating the interview results, this operation added a weighted count of 5,231 pre-K and KG births.

The poor coverage rate of only 32% for K-terminals in the list frame indicated that our 1993-94 strategy was inadequate so for the next updating cycle we developed an alternative frame approach.

V. The 1995-96 PSS List Frame Updating

For the 1995-96 PSS we expanded our effort to include K-terminal programs in the universe by adding an operation parallel to our traditional list frame operation. The NCES felt that states may have lists specifically for Early Childhood Care (ECC) facilities. These facilities sometimes include KG programs. To track down these additional or "alternative" lists of ECCs, the NCES and the U.S. Census Bureau contacted our state department of education sources, as well as state health departments or social service agencies, to ask for possible sources of alternative lists containing kindergarten programs.

The results of this effort to find alternative lists for kindergarten programs achieved partial success and looked promising for future updating efforts. Most of the states either claimed that they had no lists for ECCs or no lists other than the lists they provided for the traditional list frame operation. Only eight states provided new lists of ECCs containing adds to the private school universe. Most of these new lists were from state agencies that license day care programs. Half of these lists were highly productive (i.e., they added a relatively high number of schools containing KGs to the universe.)

Overall, the coverage of K-terminal programs in the list frame increased to 57%. While this was a significant increase from the rate of 32% in 1993-94, the number of states providing separate lists was disappointing. After reviewing our experience, we felt that lists of ECCs probably did exist in many of the states but that we were not successful in contacting the appropriate sources.

VI. The 1997-98 PSS List Frame Updating

Our 1995-96 experience led us to an extensive search on the Internet for state agency contacts, such as day care licensing agencies, or other private organizations that may have lists of ECCs. We were successful in obtaining contacts which provided more lists. Unfortunately, we still had problems trying to collect lists from these kinds of contacts:

- Many states did not have a central agency that kept a list of ECCs. Sometimes county or regional agencies within a state maintained lists, which made it very hard to obtain a comprehensive list for a state. In a couple of states, private child care resource and referral organizations cooperated in providing us with lists. However, most of these private organizations were reluctant to help because they were concerned that their lists would get into competitors' hands.

- Some states could not give us a list, or if they did, some of the files were not in a compatible format and required extra time and manipulation before they were useable.

- When contacts at the agencies were asked if their lists were any different from the traditional lists the U.S. Census Bureau collects from the state departments of education, many said no. We suspect that some may have misunderstood the question and really did have lists that were different from the department of education lists.

Another difference in the alternative lists when compared to the traditional lists is that half of the alternative sources asked for money to provide their list. While this was not a problem, it did involve the extra work of setting up a system by which the U.S. Census Bureau could pay the agencies.

In the 1997-98 PSS we received lists from 30 state day care licensing agencies or private child care resource and referral organizations. Processing these ECC lists was more challenging and more costly than processing traditional lists. Some of the reasons they were more challenging are:

- The size of the ECC lists is usually much greater than the size of the traditional lists because these agencies frequently license other types of day care, such as family and adult day care. Some agencies sent files that even included all of the licensed babysitters in the state. Thus, the lists typically contain a large number of out-of-scope programs. For example, in 1997-1998 the 30 lists contained about 95,000 total ECC programs. (In comparison, the state and association lists for the 1997-98 traditional portion of the list frame contained about 54,000 total schools.)

- The nature of the lists varied widely among the states. They did not always have helpful information such as grade ranges, ages, number of children, or type of ECC (family, group, or center based) included on their lists. Not all of the agencies that sent a computer file sent it as an ASCII file. This caused some problems when trying to make it useable. (Note: In 1995-96 we automated a portion of the matching operation. Computer files in a useable format allow us to make maximum use of this automation and save money on processing.)

- To determine the true in-scope or out-of-scope status of many of the early childhood programs, it was necessary to contact the program and ask. This meant a substantial number of phone calls. In the 1997-1998 ECC operation, our clerical staff called nearly 24,000 programs. (In the 1997-98 traditional operation, they made only about 2,900 calls.) Some programs could not be reached, either because they were no longer in existence or they were not returning the calls. Also, timing constraints made contacting some of the programs impossible.

- Not all contacts could provide a list to meet our scheduled dates, so some lists came at a later date.

The above mentioned challenges resulted in a few changes in procedure. The major changes were:

- We processed the ECC lists in two waves - an 'early' ECC and a 'late' ECC. As a result, we had to do some steps twice. We also had to add an operation that compared the new schools found in each wave to eliminate duplicate schools between the waves.

- In order to obtain a realistic workload and save on processing costs, the clerical staff in the U.S. Census Bureau's National Processing Center (NPC) did a substantial amount of screening out of programs on each list. We provided them with a list of the types of programs that were not likely to contain KGs. For example, we considered all family day cares and any programs with less than 12 children as out-of-scope and screened them out.

- Due to the large number of phone calls and limited processing time, NPC was instructed to call each program only once. This saved both time and processing costs. It did increase the number of questionnaires mailed because we sent questionnaires to programs we could not contact in case they were in-scope. (Mailing a questionnaire is generally cheaper than the effort required to track down and contact each program to determine its in-scope or out-of-scope status. We discovered that ECC programs were difficult to contact. There was a larger than expected number of answering machines, disconnected numbers or numbers not in service, and wrong numbers many of which we suspected were errors on the ECC list. A large number of programs did not even have telephone numbers listed on the ECC lists.)

Otherwise, the 1997-98 ECC operation followed the same criteria specified for traditional PSS processing.

Our effort in 1997-98 improved our coverage of K-terminals from 57% in 1995-96 to 87%. In the 1997-98 ECC operation we also looked at how many of the new KG programs from the alternative lists had a KG and one or more higher grades. Of the universe adds found in the ECC operation, 18% of these contained a KG plus higher grade. We added these programs to the traditional list frame universe. This proved that these alternative lists also provided updates that improved the coverage of schools on the traditional PSS universe. The coverage of KG plus higher grades improved to 95%.

VII. Conclusions

- *Lists from day care licensing agencies or private child care resource and referral organizations did significantly improve coverage of KGs on the universe.* The initial attempt to add K-terminal programs to the universe in 1993-94 produced only a 32% coverage of these programs. The development of an alternative frame approach starting in 1995-96 improved the coverage to 57%. Additional improvements in 1997-98 increased the coverage of K-terminal programs to 87%. The coverage of programs with a KG and one or more higher grades also improved. The coverage of KGs with a higher grade increased from 90% in 1993-94 to 93% in 1995-96 and to 95% in 1997-98.

- *Traditional state and association source lists do not have complete coverage of K-terminal or KG with higher grades*. If the states and associations did have complete or nearly complete coverage, then we would expect to increase the universe by only a very small percentage. The improvement in coverage by going to an alternative frame approach indicates that state and association lists do not cover all KG programs, especially K-terminal programs.

- *Lists from day care licensing agencies (and private child care resource and referral organizations) are inefficient and costly sources to process.* They are inefficient and costly because of the following:

    1. They contain a huge number of out-of-scope programs (such as babysitters and family day care) that results in more processing effort.

    2. They do not contain all the information (age ranges, type of facility, mention of a KG, telephone number, etc.) we need to determine whether a program is a private school containing any of the grades KG-12. This also results in greater processing resources.

We have found effective methods that reduce costs and make the workloads more manageable. These methods include:

a. The staff pre-screened ECC lists to quickly remove probable out-of-scope programs from consideration. (See section VI for more information.)

b. When calling facilities to determine their in-scope or out-of-scope status, staff limited telephone calls to one per facility. (See section VI for more information.)

c. As a result of the 1997-98 operation, we realized that we could identify and target lists that appeared the most productive (i.e., lists that were the most likely to provide adds to the universe and that were easier to process). In future PSS cycles when either budgetary or timing considerations are restrictive, this targeting of lists will allow us to process only the most promising lists at a reduced cost.

It is important to note that even though they are inefficient, these sources (day care licensing and referral agencies) are the best alternative sources for finding additional kindergarten programs that we are aware of. We believe our KG frame has provided a valuable sample frame for NCES programs including the ECLS.

This page intentionally left blank.

PREPARING SASS FOR THE MILLENNIUM: PRETESTING ISSUES IN SASS 2000

Kathleen Wise Pugh (U.S. Bureau of the Census) and Andrew L. Zukerberg (Microsoft)
Kathleen Wise Pugh 4700 Silver Hill Road, Mailstop 8400, Washington, DC 20233-8400

Keywords: Schools and Staffing Survey, Cognitive Interviewing, Questionnaire Design, Behavior Coding, Pretest

This paper reports the results of research and analysis undertaken by Census Bureau staff. It has undergone a more limited review than official Census Bureau publications. This report is released to inform interested parties of research and to encourage discussion.

I. Introduction

This paper focuses on the two stages of pretesting of the Schools and Staffing Survey (SASS) questionnaires. It begins with a brief description of the history of SASS, followed by a description of the steps involved in both pretests, actions taken to resolve uncovered problems, and concludes with future issues for SASS.

II. Background

SASS is an integrated set of surveys that are sent to approximately 13,000 public and private schools nationwide. The survey is sponsored by the National Center for Education Statistics (NCES) and is administered by the United States Census Bureau. SASS was previously conducted in 1987-88, 1990-91, and 1993-1994. The next administration of SASS will occur in the 1999-2000 school year.

SASS is comprised of five standard forms: District, Principal, School, Teacher and Library. These forms are slightly modified to meet the needs of the public, private, and Bureau of Indian Affairs schools. In response to the growing population who have access the internet, an on-line survey was developed for the Library Media Center form. Due to the rapid growth of charter schools, a new questionnaire designed specifically for charter schools was developed. This new charter school questionnaire will replace the National Study of Charter Schools.

The data collected from SASS are used by several organizations affecting education. For example, Congress uses this information when considering legislation that affects education, state departments of education find it helpful when formulating state education policies, and public school districts consult the data when considering issues such as teacher pay, curriculum, professional development, recruitment and personnel policies, workplace conditions, etc.

In preparation for the 1999-2000 full-scale SASS, field tests were conducted in two stages.

III. Overall Sample Design

The 1991-92 Common Core of Data file (CCD) was used as the public sampling frame for both field tests. The 1995-96 Private School Survey data file (PSS) was used as the private school sampling frame. The field test sample schools were selected using stratified sampling. The schools were first classified by type (public/private). The public school sample was stratified by grade level and urbanity, and the private school sample was stratified by grade level and affiliation. Both types of schools were sampled with a probability proportional to the square root of the number of teachers.

IV. Stage I of the field test

Teacher Listing Form Test

The Teacher Listing Form (TLF) is the first step in the SASS data collection sequence. This list is used to select a sample of teachers to complete the Teacher Questionnaire. Jenkins and Von Thurn (1995) conducted a round of cognitive interviews that led to recommendations to reduce coverage error in the Teacher Listing Form. The recommendations included reorganizing the form to have a vertical flow, the addition of color as a navigational aid and to increase aesthetic appeal, and the addition of definitions and examples to clarify instructions. Zukerberg and Lee (1997) implemented these recommendations and conducted another round of cognitive interviews. They found that despite these changes, respondents continued to have difficulty with the matrix, indicated by incorrectly including non-teaching staff members, and by failing to read the instructions before completing the matrix. Coverage error was still an issue. Consequently the TLF was revised again. Instructions were trimmed and navigational characteristics of the form were improved. See Zukerberg and Lee (1997), for further details.

A split panel test was conducted to compare the response rate of the revised version with the original TLF. The test showed there was no statistical difference (less than three percent) in response rates between the two forms. To be significant, the difference had to be at

least eight percentage points. The revised version was adopted for the 1999-2000 SASS.

Questionnaire Content Test

- For the field test, the questionnaires were abbreviated versions that included primarily newly developed item modules as well as some of the core items that were asked on previous versions.

The field test of the revised questionnaires did not follow the usual SASS data collection procedures. The ultimate goal of normal SASS procedure is to obtain the highest possible response rate (usually ninety percent or greater). The goal of the field test was not to maximize response rate but rather to get enough responses to review (about half of the mailout) in order to determine how well the questionnaires work.

In early 1998, an advance letter was mailed to the sample Local Education Agencies (LEA). The teacher sample was selected in January from the teacher lists collected in the TLF test administered in the fall of 1997. The first mailout occurred in February, followed by a reminder postcard one week later. A second questionnaire was mailed approximately five weeks after the first mailout to all sample cases that had not returned the questionnaire. Approximately five weeks after the second mailout, Census Bureau field staff commenced telephone follow-up. Data collection was completed in May 1998. Table one shows the sample size and response rates for the field test.

Table 1: SASS Spring Field Test Sample Size and Final Response Rates

Questionnaire	Sample Size	Final Response Rate
LEA (1a)	247	79%
Public Principal (2a)	250	66%
Private Principal (2b)	250	57%
Public School (3a)	250	66%
Private School (3b)	250	59%
Public Teacher (4a)	550	70%
Private Teacher (4b)	550	58%

Behavior Coding

In previous full-scale SASS, telephone follow-up accounted for as much as half of the completed interviews. The SASS questions are written for mail administration and contain many long, complex questions. To understand how well these questions operate in telephone administration, behavior coding was conducted as part of the telephone follow-up. Behavior coding is the systematic application of codes to the interaction between the respondent and interviewer

Field test telephone follow-up interviews were tape recorded. Staff then replayed the interviews and applied codes to situations that indicated potential problems for the pretest. A total of 92 interviews were tape recorded (with respondent permission) and coded by trained Census Bureau behavior coders during Spring 1998. Table 2 shows the number of forms and type that were behavior coded. The behavior coding revealed that questions with lengthy stems and those where respondents had to choose from "any of the following" proved to be problematic for telephone administration. For example, behavior coding of the following question on form 3b, "Is your school accredited by any of the following?" uncovered that interviewers stopped reading the choices once a respondent broke in with a "yes" answer to one of the responses in the list. It is possible that some misreporting may occur because the respondents do not have the opportunity to choose from all the answers.

Upon conclusion of the spring field test, the telephone interviewers were debriefed about their interviewing experiences with the forms. The information volunteered by the interviewers was consistent with the behavior coding data. The interviewers described incidents where respondents frequently responded before interviewers could finish reading the questions and numerous requests by respondents to have questions repeated.

Table 2: Number of Forms that were Behavior Coded

Form	Number
Public Principal (2a)	17
Private Principal (2b)	20
Public School (3a)	19
Private School (3b)	16
Public Teacher (4a)	11
Private Teacher (4b)	9

Cognitive Interviews

A number of cognitive interviews were conducted in the spring of 1998 with the field test questionnaires. All interviews were conducted by trained interviewers and tape recorded (with respondent permission). The sample was drawn from schools in the Washington DC metro area. Table 3 illustrates the number of interviews conducted by type. The interviews revealed that respondents had a difficult time following skip instructions. Many respondents answered some questions that they should have skipped while other respondents missed questions that they should have answered. In reference to scaled response items (0-5 indicating no influence to a great deal of influence), respondents were reluctant to choose "0" as a response option even though many respondents indicated that they had no influence.

Table 3: Number of Cognitive Interviews by Form Type

Questionnaire	Number of Interviews
LEA (1a)	2
Public Principal (2a)	2
Private Principal (2b)	1
Public School (3a)	2
Public Teacher (4a)	3
Private Teacher (4b)	4

V. Stage II of the field test

Questionnaire Content Test

During October-January 1998, a second field test was conducted with the SASS questionnaires. This field test differed from the first field test in that all questions were included in the questionnaires. Although the second field test mirrored many features of the full-scale SASS, including prenotification letters and reminder postcards, it did so on a condensed time schedule. The questionnaires were mailed to respondents in October 1998. A second mailing was sent to non-respondents in November 1998. Telephone follow-up of non-respondents began in mid-November. The data collection period ended in January 1999. Table 4 show sample size and response rates for the second field test.

Table 4: SASS Fall Field Test Sample Size and Final Response Rates

Questionnaire	Sample Size	Final Response Rate
LEA (1a)	471	74.1%
Public Principal (2a)	474	63.1%
Private Principal (2b)	450	65.1%
Public School (3a)	474	62.9%
Private School (3b)	450	58.2%
Public Teacher (4a)	571	56.7%
Private Teacher (4b)	446	46.2%
Public Library (LS1a)	474	45.6%
Private Library (LS1b)	450	36.0%

Professional Review Panel

A professional review panel consisting of NCES and Census Bureau staff was assembled. During November and December 1998, this panel reviewed approximately fifty percent of the completed mail questionnaires. Special attention was devoted to the School Questionnaire, particularly the new items pertaining to charter schools. The Teacher Questionnaire was also reviewed very closely because of the many new items that were added. Several of the returned questionnaires (all form types) included comments from respondents that indicated they thought the questionnaires were too long.

Behavior Coding

Prior to the second field-test, training for telephone interviewers was intensified. More attention was devoted to pronunciation of unfamiliar terms and a glossary of common education terminology was provided. The Teacher Questionnaires, forms 4a and 4b, were not included in the telephone interviews because of their complexity and length. The exclusion of these forms allowed for more comprehensive training on the other forms for the telephone interviewers.

Twenty interviews with each of the 1a, 2a, 2b, 3a and 3b forms were tape recorded (with respondent permission) and coded by trained Census Bureau behavior coders during November and December 1998. The same items that were problematic on the first field test continued to prove problematic during the second field test. Additional concerns were raised. On the principal forms the question "We are interested in the importance you place on various educational goals. From the following eight goals, which do you consider the most important, the second most important, and the third most important?" had a high percentage (15%) of

respondents who requested clarification. Respondents could not remember the choices, which raised the concern that respondents may not choose their highest preferences.

Cognitive Interviews

Concurrent with the second field test, cognitive interviews were conducted by trained interviewers using traditional think-aloud and probing techniques. Respondents were recruited from personal contacts as well as from schools that met certain criteria. The schools chosen included regular elementary, middle, and high schools as well as charter schools and schools with a migrant student population. Interviews with private schools, private teachers, and private principals included both secular and non-secular schools. The interviews were conducted in urban metropolitan, suburban, and rural areas with respondents in the Pacific Northwest, South and Mid-Atlantic regions. All interviews were tape recorded (with respondent permission) and transcribed by trained Census Bureau interviewers during the period from November 1998 through April 1999. The length of the interviews varied from one half an hour to one and one half-hours. Table 5 shows the number of interviews conducted with each form.

Table 5: Number of Cognitive Interviews by Form

Questionnaire	Number of Interviews
LEA (1a)	3
Public Principal (2a)	6
Private Principal (2b)	4
Public School (3a)	8
Private School (3b)	3
Public Teacher (4a)	10
Private Teacher (4b)	11
Public Library (LS1a)	3

Several new problems were detected in the second round of cognitive interviews. Most of the problems were uncovered on forms 4a and 4b (Teacher Questionnaires) and forms 3a and 3b (School Questionnaires). Several of the teachers had trouble with the questions pertaining to teacher certification and found it difficult to distinguish between "probationary, provisional and temporary certification". A question that asked teachers to refer to a list of assignment codes also proved problematic. The response categories provided three spaces, however many of the assignments only had one or two numbers in their codes. A similar problem was detected for questions that asked teachers to report their teaching assignments and college majors. The series of questions that asked about new teacher induction also presented problems. The questions requested that anyone who began teaching before the 1994-95 school year skip out of the series. However, many teachers who should have skipped this series answered them. There was also an instruction that asked first year teachers to answer the questions for "THIS" year. Some respondents who were new teachers, but began teaching before the current school year, misunderstood this instruction and skipped the series. On forms 3a and 3b (School Questionnaires) there is a series of questions that pertains to migrant students. Before interviewing schools with migrant students, the interviewers checked with the local Migrant Education offices to find out what schools had these types of students and how many there were. However, during the interviews, some of these schools answered "no" to all the questions that pertained to migrant students.

Library Media Center Questionnaire Internet Option

In addition to the paper SASS forms, an Internet reporting option was developed for the Library Media Center Questionnaire. Usability testing was conducted with thirteen respondents from Washington DC metro area school libraries. Thirteen respondents completed the Internet questionnaire while a trained interviewer observed their navigation through the form. When a respondent completed the survey, an interviewer asked specific questions about the respondent's experience. Two navigation methods were compared in this test: a 'scroll' based method in which the questionnaire fit on one long page, similar to a word processor document, and a 'screen' based method in which sections of questions appeared on different pages and respondents used a next / previous button and menu bar to navigate through the form. In addition, two methods of providing 'edit' messages when they entered questionable data were tested: 'passive' edits that gently alerted respondents to look at their response, and 'active' edits that gave respondents the option to automatically erase their answer. A new instrument was designed based on findings from the usability test. The 'scroll' based version with 'passive' edits was adopted. See Zukerberg, 1999 for the larger study.

All respondents who were sent the library media center survey in the second field test, were encouraged to complete the form using the Internet reporting option. Potential respondents received two letters. The first letter contained their user name and the second letter contained their password. In order to ensure that all responses would remain secure a 128-bit encryption was imposed.

VI. Overall Findings and Recommendations

Upon completion of the behavior coding, cognitive interviewing, professional review, collection of item tallies and response rate data, a formal summary of the results was compiled. This summary outlined unusual findings by item and juxtaposed all aspects of the field test on a spreadsheet for easy reference.

Findings and recommendations are presented below, divided into separate sections for length issues, formatting, cognitive difficulty, and the Library Media Center Internet survey.

Length

Findings:

- Many respondents commented that the questionnaires were too long (this was reported for all survey types). This reflects the fact that real increases in content have occurred since the last administration of SASS. For example, forms 4a and 4b (Teacher) required a maximum of 377 responses for the field test. The increase in responses since the last administration of SASS was based on the addition of 110 instructional practice responses and approximately 57 professional development responses. The highest level of burden was placed on math teachers who were required to complete a section about instructional practices.
- Teachers typically do not have large blocks of time available during the day. Most teachers are contacted during breaks between classes that are typically forty minutes in length. The Teacher form (4A&B) interview requires more than one hour to administer by phone.
- Some respondents commented that the instructions and definitions for some questions seemed longer than the questions.

Recommendations:

- "Trim" questions from all forms.
- Ask those teachers with a preference for phone interview, permission to call them at their residence.
- Reduce length of instructions and/or the complexity of questions that require lengthy instructions or definitions. Highlight instructions with bullets.

Formatting

Findings:

- Respondents missed skip patterns on many occasions (across all forms). Some respondents answered questions that they were supposed to "skip" while some did not answer questions that were intended for them.
- The majority of questions on the field test forms were double banked which seemed to contribute to many of the incorrect skip patterns.
- Some respondents treated "yes or no" questions as "mark all that apply" questions. For example, a respondent would answer six out of ten questions as "yes" and leave the remaining four questions blank.
- Several teachers became confused on a series of questions that required them to fill in a code number that corresponded to a table of pre-coded choices. The pre-codes that they had to choose from had two and three digits. The response option allowed space for a three digit code.

Recommendations:

- Reformat skip patterns so that respondents may navigate through the form more easily. Replace the term "skip" with "go to" and combine with the appropriate question and page number.
- Revise double banking format to a single column. There is a bit of a trade off with this recommendation since switching to a single column format would increase the number of pages in the questionnaires. However, we believe that this measure will simplify the flow of the questions and will outweigh any negative impact associated with the minor increase in length.
- Enlarge response boxes for "yes, no" series questions
- Re-code categories from the tables and corresponding response options to two-digit codes to provide consistency and avoid confusion.

Cognitive Difficulty

Findings:

- A few respondents were confused by questions containing professional terms or "jargon". One such example was the use of the term "pedagogy" which was thought to be a fairly common term in the teaching profession but was found to be problematic.
- Some question stems with many sub-items were continued across several pages causing respondents to forget what the question was asking.
- Many respondents had difficulty answering questions on the School form (3A&B) that pertained

to migrant students. The respondents commented that individuals at the District office or the Migrant Education office would have access to more accurate information.

Recommendations:

- Eliminate or define "jargon."
- Repeat stem so that responses are not disconnected from the main focus of the question. Eliminate or combine redundant response categories.
- Evaluate whom in the establishment would be best suited (most knowledgeable) to answer specific questions and place those questions on the appropriate questionnaire.

Library Media Center Internet Survey

Findings:

- Only one percent of the respondents completed the survey over the internet. Many respondents' internet browsers could not support the high level of encryption required to access the survey (128-bit encryption).
- Two letters containing a password and user name respectively were sent to all potential respondents. This proved problematic because some respondents lost one of the letters and could not access the internet version.

Recommendations:

- Reduce the level of encryption required.
- Send one letter containing both the password and user name rather than two separate letters.

VII. Future Issues for SASS

As new ideas develop and policy makers require access to educational information, it is likely that the demand for SASS data will continue to grow. The scope of information that is currently available from SASS is already quite large. Future administrators of SASS will have to consider narrowing the scope or otherwise run the risk of receiving invalid data. This undoubtedly will prove problematic because the number of stakeholders involved with SASS is numerous. Many of these stakeholders will be reluctant to delete questions of interest to them.

Another issue to be considered is the negative impact that proposed changes could pose to time series data. Some policy makers and school organizations track data from each SASS administration. Every effort must be made to protect these questions but ultimately the integrity of the questionnaires had to be upheld. This is a delicate balancing act that will continue to be a struggle as SASS grows.

The National Processing Center (NPC) will print all of the forms for SASS on site, with Docuprint. This machine has many features that will help to usher SASS into the new millennium. Docuprint's most notable feature is its capacity for customization of individual forms. For example, teacher questionnaires can be personalized with the teacher and school name throughout the questionnaire.

Finally, as more people become Internet literate, adding an internet option to the administration of all SASS questionnaires should be explored as a way to further increase response rates.

VIII. References

Jenkins, C.R. and D. Von Thurn. (1995) "Cognitive Research on the Teacher Listing Form for the Schools and Staffing Survey." Washington: U.S. Census Bureau

Zukerberg, A.L. and Lee, M. (1997) "Report on Cognitive Research on the Teacher Listing Form." Washington: U.S. Census Bureau

Zukerberg, A.L. (1999) "1998 Pretest and Questionnaire Revisions." Washington: U.S. Census Bureau

CREATING LINKED CCD DATA TO IMPROVE THE QUALITY OF ELEMENTARY/SECONDARY SAMPLE SURVEYS

Stephen R. Wenck, Albert C.E. Parker, Sameena M. Salvucci, Synectics for Management Decisions,
Carl Schmitt, National Center for Education Statistics
Stephen R. Wenck, Synectics, Suite 900, 1901 N. Moore St., Arlington, VA 22209

Introduction

The National Center for Education Statistics (NCES) has been collecting data about the approximately 15,000 public school agencies (school districts) and the approximately 85,000 public schools as part of the Common Core of Data (CCD) for the past 15 years. These data form the sample frame for all of NCES' elementary/secondary school sample surveys. In order to improve the coverage and efficiency of their elementary/secondary school sample surveys, the quality of the sample frame—the CCD—must be improved and made more consistent. NCES also wishes to make CCD school data for the 1986–87 to 1997–98 academic years available to researchers, public policy analysts, public officials, and the general public in an on-line electronic format in which comparisons across schools, districts, or years can be made easily.

At the school level, the data include:

- Number of teachers (full time equivalent);
- Number of students by grade level;
- Grade levels served by the school;
- Number of students by racial or ethnic (Hispanic) category;
- Number of students eligible for Federally subsidized free lunches.

To be of maximum usefulness, the data should be presented in a consistent manner and should be as correct as possible, without obvious errors arising from misreporting or incorrect data entry. Data for a large set of school districts (about 90 percent of all districts) have been extensively edited. Imputation has been done for districts where data are missing or appear unreasonable. At the school level, however, no editing or imputation has so far been done. In an earlier investigation of the quality and consistency of the data on numbers of students and teachers the preparatory work was carried out on a dataset limited to four types of school districts within the 50 states and the District of Columbia:

1. Local school districts that are not components of supervisory unions;
2. Local school district components of supervisory unions sharing a superintendent and administrative services with other local school districts;
3. Supervisory union administrative centers or county superintendents serving the same purposes as supervisory unions; and
4. Regional education service agencies or county superintendents serving the same purposes as supervisory unions.

The remainder of the paper will discuss the steps we have taken, and those we are still in the process of completing, in developing a multi-year linked CCD dataset.

Step 1. Preparation of Data for Linked Datasets

This activity included acquiring all the necessary data and converting the data to a common format. These activities were completed before any changing of the data by editing or imputation.

TECHNICAL APPROACH

Review what schools to include

Our next step was to determine the schools to include in the time series data. We agreed that these data would cover the 1986 to 1997 school years. Over time, schools do occasionally change school districts. Therefore it is possible that a school in a district in 1986 would not still be in a district in 1995.

In addition to reviewing the inclusion and exclusion of school districts with all of their schools, we added and deleted individual schools as necessary. In our preparatory work, we determined that 662 schools were candidates for deletion. These schools appeared only once in the ten-year time span. There was an interruption of reporting of one year or more at another 80 schools. These 80 schools were candidates for addition in the skipped years. In all 742 cases, we worked with NCES to determine which schools to add or delete.

Add master school ID field to all school records

We then created a consistent ID number for linking across all 10 years. This ID is a composite ID (state FIPS number, state-assigned district ID, and state-assigned school ID) for the school in the year that the school was first reported. We maintain this ID for all subsequent years. However, each annual record also includes the ID assigned to the school for that year. Our

preparatory research found numerous instances of changes in district and school ID. The current annual identifying information is maintained for matching to outside data that use it as an identifier.

Make variable names consistent
Next, the variable names were made consistent across all years. Since we are creating a rectangular dataset to compare school characteristics across time, the variable names must be made the same. Currently there is a year identifier on each variable (e.g., "member89" for school enrollment in 1989).

Make special codes consistent
Related to consistent variable names are consistent "special" codes. These codes designate data that are missing, not reported, or not applicable. It is beneficial for the numeric codes for these situations to be consistent across data fields, to facilitate identification of codes that must be treated specially during tabulation or other statistical manipulation of the data. We developed a consistent scheme for these codes and converted all existing codes to the scheme. The scheme was based on existing codes to minimize the amount of conversion required.

Determine which schools have ID number shifts and settle on a consistent ID to use throughout
Once the schools to be included were determined, we also reviewed the consistency of their ID numbers. During the preceding task we found cases that appeared to be the same school (same name, city, and state) with different ID numbers. We identified all such cases and made sure that they have the same ID number in all years. The initial ID number was carried across all years even if the state-assigned ID number changed during the coverage period.

Add fields to all school records for imputation flags
Once the master file was created, we added fields to all school records in which to place flags indicating that data have been imputed. The flags differentiate between data that were imputed because they were missing or not reported and data that were imputed to replace numbers that were implausible or unreasonable. All flags were set to "not imputed." Flags for cases that were imputed would be changed in later steps.

Step 2. Editing and Imputing of School Data

At this point, we had a consistent set of schools, with consistently labeled data, across a period of 11 years. This dataset would be sufficient for a data warehouse or any time series analysis, but we knew from our earlier investigation that there were anomalies in the data that suggested reporting or calculation errors.

TECHNICAL APPROACH

Find student total anomalies, FTE teacher anomalies, and review
We first examined the number of full-time equivalent (FTE) teachers. We looked at total enrollment and FTE teachers together because large accompanying changes are more plausible than large changes in the number of teachers only. Preparatory work included development of a mathematical definition of longitudinal anomalies —what we called a difference measure.

$$D_n = (S_{n+1}/T_{n+1} - S_n/T_n)/S_n/T_n)$$

where D_n = the difference measure for two years, n and $n + 1$, S = the number of students in year n or $n + 1$, and T = the number of teachers in year n or $n + 1$. We investigated the distribution of the difference measure for some of the years and found that about one percent of all the schools in the 1986–97 files showed anomalies over three-year periods that warranted further investigation.

Where the number of students or teachers was inconsistent with other data or with adjacent years, we calculated district-wide totals to determine the likelihood of a data entry error as the cause. For example, suppose that:

1. A school is reported as having 25, 62, and 27 teachers in three successive years;
2. Adding the number of teachers in all schools across the district produces a total that exceeds the total in the district record by 36; and
3. Reducing the middle year total for the school in question by 36 would produce an FTE teacher series of 25, 26, 27.

In this case we assumed that "62" was a transposition in data entry for "26". For situations where a data point appears, we will impute a replacement number.

We discussed marginal or puzzling cases with McLaughlin to take advantage of his experience with district-level data and knowledge of unusual developments in the organizations and enrollments of particular districts.

Impute teacher anomalies due to data entry error, reporting errors, missing/implausible data
We imputed a replacement number using PROC IMPUTE. Our regression equation used the adjacent years of FTE data (year prior and year after target year),

three years of enrollment data (prior, target, and subsequent), school locale, and school size as the predictor variables for the FTE value to be imputed.

Find and review grade total anomalies
We are now in the process of reviewing individual grade enrollments. We will perform a series of tests on the data. After each test a flag will be assigned as to whether the school (or grade) passed that test. After all tests have been performed we will look at the distribution of these flags and from that distribution decide which combination of passes and fails we will chose to impute.

The first test is the "District Imputation Test." This test determines if a school is contained within a district where the district enrollment was imputed. This test is helpful as it indicates a data quality problem at the district level, which may mean something is peculiar at the school, and possibly grade level.

The second test is the "District Sum Test." This test determines if the sum of the grade enrollments for all schools in a district equal the reported enrollment (unimputed) on the district file. If this test passes (sum of school data equals district data), then it is a pretty strong indicator that the data at the school level are okay. Cases which pass the District Sum Test will be set aside; no further testing will be done on grades within these schools, and they will not be considered for imputation.

The next two tests are related. They are the "Cohort Test" and the "Grade Test." In the cohort test, we compare the enrollment in each numbered grade within each school (excluding UG, PK, and KG) with the enrollment of the next lower grade in the preceding year and the next higher grade in the following year. If the grade is the highest in the school, we compare it to the next lower grade in the preceding year and the next lower grade than that in the year before that. If the grade is the lowest in the school, we compare it to the next higher grade in the following year and the next higher grade than that in the year after that. Examples for a school with a grade span of grade 1 to grade 6:

Target		Comparison 1		Comparison 2	
Grade	Year	Grade	Year	Grade	Year
4	1988	3	1987	5	1989
1	1988	2	1989	3	1990
6	1988	5	1987	4	1986

The grade fails if all or one of the following sets of conditions (for a V or a Λ) is true:

General Cohort Test for V:

(1) the previous year's enrollment in the next lower grade is at least twice the current year's enrollment in the grade under examination;
(2) the next year's enrollment in the next higher grade is at least twice the current year's enrollment;
(3) the difference between the previous year's enrollment in the next lower grade and the current year's enrollment is $\geq$15;
(4) the difference between the next year's enrollment in the next higher grade and the current year's enrollment is $\geq$15.

General Cohort Test for Λ:

(1) the previous year's enrollment in the next lower grade is no more than half the current year's enrollment in the grade under examination;
(2) the next year's enrollment in the next higher grade is no more than half the current year's enrollment;
(3) the difference between the previous year's enrollment in the next lower grade and the current year's enrollment is $\geq$15;
(4) the difference between the next year's enrollment in the next higher grade and the current year's enrollment is $\geq$15.

The ratios, 2 and .5, are greater than the 1st and 99th percentiles for within-grade cross-year ratios for 1994. Percentiles vary from grade to grade and year to year. A consistent pair of ratios is easier to understand, explain, and apply than the annual grade-specific ratios that we used in the trial run and that are comparable to the annual percentiles that we used for the earlier school-level task.

V/Λ Cohort Test for Lowest Grades. If the grade is the lowest grade in the school, it fails if it includes $\geq$15 students and the current year's enrollment is at least twice or less than half of the enrollment for both of the successively higher grades in the next two years. For example, the 7th grade enrollment in a school with a grade span of 7-12 would have to be half of the 8th grade enrollment in the next year and half of the 9th grade enrollment in the year after that or half of the 8th and 9th grade enrollments in the same years.

V/Λ Cohort Test for Highest Grades. If the grade is the highest grade in the school, it fails if it includes $\geq$15 students and the current year's enrollment is at least twice or less than half of the enrollment for both of the successively lower grades in the preceding two years. For example, the 6th grade enrollment in a school with a grade span of kindergarten-6 would have to be half of

or twice the 5th grade enrollment in the previous year and half of or twice the 4th grade enrollment in the year before that.

V/Λ Cohort Test for End Years. The Cohort Tests cannot be performed as described above for the first and last academic years covered by this project (1986–87 and 1997–98, respectively). Where the grade range is sufficient, the Cohort Test can be performed instead on the next two years for 1986 and the previous two years for 1997. In these cases, the target year enrollment should be significantly different from successively lower grades in both of the preceding two years, or significantly different from successively higher grades in both of the following two years. In these tests, the enrollment if graphed would describe not a V or a Λ but a sharp increase or decrease (doubling or halving) followed or preceded by a plateau (the asterisk indicates the target enrollment):

1986: */‾‾‾ *__
1997: __/* ‾‾‾*

The 1986 tests are not applicable to the two highest grades in a school; that is, if (HIGR – *g*) < 2, where HIGR is the highest grade with enrollment in the school in 1986 and *g* is the target grade. The 1997 tests are not applicable to the two lowest grades in a school; that is, if (*g* – LOGR) < 2 where LOGR is the lowest grade with enrollment in the school in 1997.

Some children do not attend public pre-kindergarten classes, some public schools do not offer kindergarten, and ungraded students by definition can not be expected to progress annually from one "grade" to another. Therefore, the V/Λ Cohort Test is not applicable to the three enrollment categories designated as "UG," "PK," and "KG." Only the V/Λ Grade Test applies to enrollments in these categories.

V/Λ Grade Test. The Grade Test compares the current year's enrollment with enrollment in the previous and subsequent years.

V/Λ Grade Test for End Years. The Grade Tests cannot be performed as described above for the first and last academic years covered by this project (1986–87 and 1997–98, respectively). The Grade Test can be performed instead on the next two years for 1986 and the previous two years for 1997. In these cases, the target year enrollment should be significantly different from enrollment in the same grade in both of the preceding or following two years. Thus, the enrollment if graphed would describe not a V or a Λ but a sharp increase or decrease (doubling or halving) followed by a plateau (the asterisk indicates the target enrollment):

1986: */‾‾‾ ‾‾‾*
1997: __/* *__

Impute remaining missing/implausible grade totals
We will review the combinations of passes and fails for the above tests, and decide to impute cases after reviewing that distribution.

We will impute a replacement number using PROC IMPUTE. Our regression equation uses enrollment in the same grade for the prior and subsequent years; enrollment in the next lower grade, if available, in the previous year; enrollment in the next higher grade, if available, in the following year; and school size as the predictor variables for the grade enrollment to be imputed. Note that we have to split the dataset into a series of grade datasets (e.g., all schools offering grade 1, all schools offering grade 2, etc.). This is necessary as a school not offering a particular grade will have a missing value for that grade. PROC IMPUTE uses missing values as the values to be imputed. Therefore a school not offering a grade, but included in that grade's imputation dataset would be assigned a value for that grade when in fact it never could have had students in that grade.

Compare grade totals by school to grade ranges and identify discrepancies, review discrepancies between grade totals and grade ranges
As we edit individual grade enrollment, we will be able to edit the grade span variable. A difficulty with "grade span" is that a small school may be organized and intended to serve a span of grades but some grades might not have any students in a particular year. For each school, we will compare reported grade span to the grades for which there is positive enrollment. If this grade span differs from the reported grade span, we will retain the reported grade span if the missing grades had enrollment within the previous two years and the subsequent two years and the coefficient of variation of annual grade-specific enrollment changes within the school is close to 1.0. We will have to investigate some actual cases to determine a rule that can be applied to large numbers of schools. When a grade disappears from a school, we will also look at the next lower grade in the previous year and the next higher grade in the subsequent year to see whether the absence of the grade is explainable by an anomaly in the distribution of students in the community across grades.

If a grade at the top or bottom of the grade range disappears for only one year, we will look at the enrollments of other schools in the district to try to

confirm a temporary reorganization of the district. If the "missing" grades can be accounted for by increases in enrollments in the corresponding grades in other schools in the district, we will assume that there was a reorganization of the grade structure of the district schools that was reversed after only one year. If the grade disappears from more than one school in the district, we will assume that it was the result of a deliberate local decision about the grade structures of the various schools. If there are several schools to which the "missing" grade could have been assigned, so that the students could have been dispersed without causing anomalies in the grade structures of other schools, we will also assume that there was a deliberate decision to change the service grade span of the school in question.

We will compare the sum of the individual grade enrollments to the total enrollment for a school. This type of comparison allows for the data to be consistent across years as well as within year. We will complete our data "cleaning" process by conducting a similar comparison on a larger scale. We will sum the enrollment and number of teacher fields for all the schools in a district and compare those against the data in the time series district dataset. This will ensure that school data within a district are consistent between the two time series datasets.

Correct grade range discrepancies
Some grade-range anomalies may be obvious enough for us to decide to replace the reported data without further consideration. However, we will discuss marginal or puzzling cases with McLaughlin to take advantage of his experience with district-level data and knowledge of unusual developments in the organizations and enrollments of particular districts. After we have discussed these cases with him, we will decide what corrections to make and replace what we have good reason to believe are erroneous data. PROC IMPUTE does not work for grade service ranges as the grade range variable is a character variable and PROC IMPUTE is only capable of imputing numeric data.

RACE AND ETHNICITY

We will look for year-to-year anomalies as we did for students, teachers, and grade-specific enrollments, and we will compare anomalous data to data from the Office of Civil Rights. Because of possible district policies regarding the racial composition of schools, such as the institution of busing or the abandonment of busing plans effected before the coverage period, and the creation of magnet schools, we can expect to find some sudden but legitimate shifts in the number and proportion of students from a particular race in a particular school. Therefore, it will be necessary to accept abrupt changes in race data as long as the students can be accounted for elsewhere in the same school district.

Find race/ethnicity data discrepancies
We should be able to detect anomalies within each racial and ethnic group by using the same method as for total students and teachers, that is, by comparing each year's data to the preceding year's (except for the first year) and the following year's (except for the last year). The difference measure formula should be applied to each race within each school. It is not clear whether it should be applied to the *number* of students of each race within the school or to the *percentage* of students of each race within the school. The number could change for all races if the school is expanded or part of it is closed, changing the total enrollment dramatically. The percentage should not change dramatically unless there is a change in attendance zone boundaries or busing policy or the school becomes a magnet school. We should calculate a difference measure for both number of students and percentage of students for one race for one year and compare the results.

Review race/ethnicity data discrepancies
Discrepancies will have to be reviewed carefully to determine whether they are plausibly related to changing ethnic patterns within a district rather than to a reporting or data entry error. For example, if there is an abrupt change in a school but the same shift occurs within the district, then the shift should be accepted as plausible and not changed on the dataset.

Correct race/ethnicity discrepancies
Some cases may be obvious enough for us to decide to replace the reported data without further consideration. However, we will discuss marginal or puzzling cases with McLaughlin to take advantage of his experience with district-level data and knowledge of unusual developments in the organizations and enrollments of particular districts. After we have discussed these cases with him, we will decide what corrections to make and replace data due to data entry and reporting errors for which we can identify an obvious correction. We will impute data for missing cases or other cases for which a single replacement number is not apparent.

Step 3. Documenting Edits and Imputations

Documentation of the editing and imputation will be important for future users to understand the strengths and limitations of the data when they are made available for public use. We will prepare documentation

of all the steps we took in creating the time series datasets.

TECHNICAL APPROACH

Prepare draft documentation

The documentation should include the following materials:

1. A description of the criteria used to include and exclude schools from the datasets;
2. A table showing the number of schools, and the number of districts they represent, in the database by year;
3. A description of the methodology for linking records across all years and of the policy followed in assigning and maintaining school ID numbers;
4. A codebook listing all variables names, imputation flags, and status codes, and the meaning of all codes that are not representation of numeric values of continuous variables;
5. The record layout used for all years;
6. A description of the criteria used to identify anomalous data that was replaced by imputation; and
7. A description of the imputation procedures used.

IMPROVING RESPONDENT SELECTION PROCEDURES IN SCHOOL ESTABLISHMENT SURVEYS: IMPLICATIONS FROM COGNITIVE INTERVIEWS FROM THE SCHOOLS AND STAFFING SURVEY (SASS)

Benjamin Cohen (ESSI), Andrew Zukerberg and Kathleen Wise Pugh (U.S. Bureau of the Census)

This paper reports the results of research and analysis undertaken by Census Bureau staff. It has undergone a more limited review than official Census Bureau publications. This report is released to inform interested parties of research and to encourage discussion.

This paper may not be quoted without permission of the author.

Complex organizations, such as schools, often serve as a unit of analysis in establishment surveys. One inherent difficulty in selecting the most appropriate respondents from complex organizations is that their structure may mean different types of respondents—teachers and principals, for example—rely on different information to answer survey questions. This circumstance makes schools unlike simpler organizations such as households, establishments where all adults are assumed to have enough knowledge of the household to be eligible to report for the entire household. Thus, in establishments that segment themselves into specialty areas or departments, as many schools do, identifying and gaining the cooperation of the most knowledgeable respondents requires an understanding of the complexities that affect the data these respondents can provide.

This paper reports on respondent selection research based on cognitive interviews using questions from the Schools and Staffing Survey (SASS). SASS is a national- and state-level representative survey, previously conducted in school years 1987-88, 1990-91, and 1993-94. Approximately 13,000 public and private schools are included in each SASS sample administration. The survey consists of several questionnaires,[1] including a teacher and principal form. For this study, we analyzed teacher and principal responses to attitudinal questions each of them answered about (1) school problems and (2) the extent of influence teachers have in the school. Implications for respondent selection and the analysis of data from these questions are discussed.

Three research questions circumscribe this framework:

1. How do teachers' perceptions compare to principals'?
2. Why are their perceptions different?
3. Are teachers or principals in more advantageous positions to answer certain questions? How does this impact data quality and analysis of data?

The 1993-1994 administration of SASS revealed differences between principal and teacher responses to these questions. Out of the six dimensions of teacher influence and 24 types of school problems that SASS measures, we focused on those "influence" and "school problem" items where the greatest differences between principals and teachers were evident.

Methodology

This study utilized cognitive interviews to capture the salient operative aspects of establishments. In particular, the study focused on the ways teachers influence different components of the establishment operation, such as teacher evaluations, discipline policies, and curriculum. Perceptions of problems that affect the school, such as student apathy or parental involvement, were also examined.

Cognitive think-aloud interviews were conducted by trained Census Bureau and Education Statistics Services Institute (ESSI) researchers following a protocol, although interviewers were free to deviate from the protocol as necessary. Thirty-one respondents were recruited from the Washington, DC, metropolitan area through fliers and personal contacts. Respondents were offered a $30 incentive to participate in this study. Interviews were conducted at schools and in respondents' homes from October 1998 through April 1999 and were tape recorded (with the respondent's consent). Interviews lasted from approximately 1 to 2½ hours. In roughly half of the schools we interviewed both a teacher and the principal. The remaining participants worked in unrelated schools. Table 1 shows the number of respondents interviewed and type of school where they were employed.

[1] The questionnaires examined in this study include the public and private teacher questionnaire, and the public and private principal questionnaire. In addition to these, SASS has a public and private school questionnaire, a district questionnaire, and a public and private library questionnaire. In addition to these SASS collects data from Bureau of Indian Affairs schools, Department of Defense Schools, and Charter Schools.

Table 1 Number of respondents interviewed, by type of school

Form	Principals	Teachers
Public School	6	10
Private School	4	11
Total	10	21

Interviewers queried each respondent about the source of information used to compute an answer, the length of recall time, other generic answer computation processes (i.e., "How did you arrive at that answer?"), and what respondents meant by their answers (e.g., "What does a 'serious' student apathy problem mean?"). Interviews were tape recorded and transcripts were organized in spreadsheets so that responses could be easily compared in different dimensions (e.g., across schools, within respondent, within question items). Once the data were available in this form, all responses were analyzed and common themes were observed. Essentially, the themes were coded and then analyzed independent of the items themselves. More objective analyses were possible by comparing actual numeric responses to the questions.

This study's findings are limited by the small sample size. However, while the findings are not representative, they shed light on a useful method of respondent selection. Moreover, since this study includes respondents from urban, rural, and suburban schools—teachers and principals with varying degrees of experience—even though no sampling methodology was applied, the results point to important factors to consider in the design of establishment surveys.

To answer the research questions, two main sections of results follow:

- Teachers and principals are compared on the varying extent of influence each perceives that teachers have.
- Teachers and principals are compared on how they view school problems.

RESULTS

Cognitive Tests of Teacher Influence Questions: Implications for Respondent Selection

The data revealed four themes that categorize teacher and principal answers to questions about teacher influence. These themes also categorize the operational aspects of the school, which in turn can help survey researchers design questionnaires and analyze data. The themes deal with: (1) the teacher's official position or role at the school; (2) non-institutional phenomena or the teacher's personal motivations; (3) extra-school phenomena (such as the State and LEA); and (4) intra-school policy, what we sometimes refer to as school "workings."

(1) Teachers' Official Position and Their Multiple Roles While teachers were found to provide valuable information for their school, their various roles must be known if data are to be validly understood. Teachers described numerous roles in addition to their main teaching assignment that brought them more influence. These teachers were often department heads. As a department head, teachers explained that they had more influence on decisions about hiring new teachers, evaluating current teachers, and/or operating a budget at the department level. However, not being in these roles does not necessarily limit teacher influence. For example, in influencing the budget, department chairs in each case had more influence than other teachers. However other teachers in this study explained that budgets are sometimes developed by teams of teachers or in collaboration with department chairs. Teachers mentioned other roles beyond their main teaching assignment that increased the extent of their influence. These roles were associated with committee membership, program involvement (e.g., a drug/alcohol prevention program led by teachers), being a curriculum or reading specialist or other coordinator, and being a lead or master teacher. These circumstances do not invalidate teacher responses to establishment surveys; however, if data are to be more useful, myriad teacher roles must be known to evaluate the data that teachers provide.

Principals are clearly advantaged by their position in the school and their responses provide important information for researchers. One principal in this study remarked about his influence on professional development relative to teacher influence, claiming, "I see the full picture and sometimes [teachers] only see their needs in a classroom. So if I see a [school need] I will go for that..." Another principal's position was more tenuous, given his school's reconstitution eligibility; due to the additional State oversight, he had less power than other principals we interviewed. Like teachers, the principal should be consulted to provide additional information about teacher influence.

Type of Teaching Assignment Some respondents explained that teacher influence on curriculum or budget is related to the type of course they teach. For example, a teacher's influence on an Advanced Placement class is limited by standards set by organizations that are often beyond the teacher's control. Teacher influence on budget may depend on what items the budget includes; teachers may have more influence on the purchase of certain books and instructional materials than on items not directly related

to instruction. Survey researchers must adjust their analysis with consideration of such details if teachers' responses are to be correctly understood.

Tenure of Other School Officials One teacher remarked that her influence was greater because the principal was new to the school. In this particular case, the principal's status resulted in teachers taking greater authority. While only one instance of this was evident in the interview data, researchers should pay attention to how tenure issues might affect teacher influence, particularly for school surveys that sample new or emergent institutions. In this school, one teacher seems to be a better respondent since the principal's role was reduced. More research would be useful to determine how data quality is affected by the relative tenure of teachers and principals; this finding has implications for large-scale surveys in which several schools in sample are likely to have new principals.

From these results it is clear that both teachers and principals provide valuable data to establishment surveys. However, their responses must be understood in terms of teachers' auxiliary roles, teachers' status relative to the principal, teachers' or principals' involvement in different school programs, and teaching assignments. Similarly, a principal's tenure, degree of authority, and the school's status are variables related to the answers principals provide to school surveys. Establishment surveys must collect such information if a more fully determined understanding of the establishment is desired. Teachers who are members of various committees may serve as better respondents to questions that address the substantive work of those committees. It follows that principal membership on district-level- or community-based committees also affects their quality as a respondent.[2]

(2) Teacher's Personal Status and Other Non-institutional Phenomena The value of teacher and principal responses was related to personal motives and unofficial phenomena that shaped day-to-day school activities.

Motivations of the Teacher While not directly related to school operations, teachers' motivations were important in determining their influence on various operational activities at the school. Both proxy and direct responses, as well as evidence from pairs of principals and teachers at the same school, revealed how motivation affected teachers' influence. For example, a principal described the opportunities teachers at his school had to participate in curriculum development. Two teachers at his school acknowledged this opportunity yet explained that their motivation prevented them from participating in curriculum development. Second, teachers made proxy responses for their colleagues, explaining that their motivations affected their influence on various matters (such as curriculum development or setting discipline policies). Third, teachers acknowledged that their own motivations affected how much influence they had. Since a teacher's motivation or morale affects the quality of data they provide, measures of morale may highlight suspect data.[3]

Time Available If teachers are chosen as respondents, the time available to them may alter their responses, particularly when they answer questions about which school operations they can influence. Teachers indicated that the time available to them at school affected their ability to participate in the evaluations of their peers. That is, teachers were not able to participate instead of choosing not to participate. Other studies have shown that the time demands placed upon teachers inhibit them in more than one way.[4] Therefore, it comes as no surprise that the ability of teachers to influence school operations is small when day-to-day classroom work prevents some teachers from participating in other school activities.

Establishment surveys must consider these "non-official" structures (time) and personal aspects (motivations) of the school establishment's central character, the teacher. Without an accounting of the teachers' motivation and available time, researchers will not have a full understanding of the meaning of

[2] Currently, the SASS Teacher Listing Form (TLF), sent to schools prior to the full scale survey administration to obtain a complete list of potential teacher respondents, may be improved by collecting data on such committee memberships.

[3] The SASS teacher questionnaire currently asks, "If you could go back to your college days and start over again, would you become a teacher or not?" Responses to this question might indicate low morale, possibly related to answers about teacher influences on school policies and programs.

[4] Indicator 42, "Teaching workload of full-time teachers," in the *Condition of Education 1997* states: "While full-time public school teachers were required to be at school 33 hours per week on average in the 1993–94 school year, they reported working 45 hours per week. Private school teachers were required to be at school an average of 34 hours per week, but reported working 47 hours per week. Also, public and private full-time teachers reported spending extra hours (12 and 13 hours, respectively) before and after school and on weekends; of these extra hours, about one-fourth were spent in activities involving students." (nces.ed.gov/nces/pubs/ce/c9742a01.html) Such working hours leave little time for teachers to be involved in planning school policies, or other areas they might influence.

their answers. However, even though many teachers have considerable demands placed upon them, this aspect of schools does not indicate that teachers should be ignored.

(3) External Influences on Teachers

Official Management or Governing Bodies The interviews indicated many factors operating outside the school which were related to the ability of teachers to influence various school-establishment operations (e.g., professional development planning, budget planning, etc.). Often, teachers and principals alike cited the State, county, district, school education board, parents, and for private schools, religious authorities or the Diocesan board, as affecting their influence. For example, one private school principal described religious documents affecting discipline policy beyond teachers, since "students are asked to look through the book of Psalms and see what [disciplinary action] might be pertinent to their offense."[5] In one public school, a principal explained that, "Teachers govern by what State and local school boards say. They don't set the standards, they implement them." This principal's comment suggests his belief that teachers implement policies according to their intent. Information about the official intent or design of policies may be known to principals and teachers alike; however, the actual effect of a policy, or its implementation, in many cases may be better known by teachers (discussed below in section (4)).

The overall performance status of the school also relates to respondent selection procedures. In this study one school district was reconstituting some of the schools under its authority. Under these circumstances, teacher influence was impeded by stricter LEA control of school budget, performance standards, and teacher evaluations and its use of strategic plans (which require more precisely defined rules for the school's staff and the school's operation). Other external influences included site-based management teams, which include various actors in and out of the school. Such teams interact with the different roles teachers have, complicating the flow of influence a teacher has on any particular matter the management teams address. The implication for respondent selection is to contextualize teacher and principal answers according to a school's degree of independence from higher authorities and the roles of mid-level management entities, such as site-based management.

Teachers and principals both provide useful data on the effect of external factors, but their experiences may be quite different. This is evident when comparing measures of the intent and implementation of school or district policies. Establishment surveys must develop research designs that account for this circumstance if they are to effectively use data from principals and teachers.

(4) Intra-school Policies and Operations (How does the actually school work?)

When respondents answered questions about teacher influence, many mentioned day-to-day school operations in practical terms without referencing official policies. In this vein, respondents described how committees function, the difference between formal and informal procedures, and other details that shape the "true" nature of school operations versus what is intended by formal rules.

Intent versus Practice When teachers responded to questions about their influence on school policies, many explained that their formal influence on the policy differed from their influence in the classroom. Variations on policies were evident in classrooms. To understand teacher influence on policies requires a multi-level analysis; formal versus informal and school versus classroom. Such complexity does not eliminate the teacher as a respondent, but their answers must be understood in context with these circumstances.

One teacher's description of her influence on discipline policy captures the flavor of such responses. She distinguished the school's "overall policy" from the policy she applies in her classroom. While major transgressions of acceptable student behavior could not be ignored, she explained that minor infractions could be dealt with according to the teacher's discretion, whether or not such infractions were covered in the discipline policy.

Another private school teacher's influence on professional development captured many aspects of the "actual" workings of the school-establishment, "…it is out of our hands in general, although there are specifics if you happen to be a member of the committee…But for me on a day to day basis we do have some input but certainly less than 50% in my mind, but some input. What does [a "great deal of influence"] look like? It would be democratic process, ...I think it would be almost impossible to do because you would not get anybody in here to reach a consensus." This response underscores the complexity of who has authority in schools.

What is "Final Authority?" Some principals said they were the ultimate authority in many areas and others acknowledged how teachers (individually or as teams/groups) were taken seriously as principals make

[5] Approximate quote.

decisions. Some principals only acknowledged teachers' ideas while others permitted teachers to influence school decisions. Obviously, the principal's relationship to the faculty varies from school to school. Establishment surveys that attempt to measure the influence of teachers in school decisions must account for how principals view teachers.

Cognitive Tests of School Problems Questions: Implications for Respondent Selection

This study examined a second group of questions that measured teachers' and principals' perceptions of school problems. On the Schools and Staffing Survey, absenteeism, disrespect for teachers, student apathy, parental involvement and student unpreparedness were measured with a 4-point scaled response option: "Not a problem," "minor" "moderate," or "serious" (see attachment 1). Cognitive interviews conducted using these questions provided data that discern whether teachers or principals differ in their responses and how these responses are formulated.

Teachers' Information Sources The information sources teachers used to judge the extent of problems indicated that they have detailed knowledge of classroom and student-level phenomena directly related to school problems. While teachers' knowledge at the student- and classroom-level may be more complete, it also seems more subject to bias. For instance, in describing absenteeism problems, teachers typically referred to daily reports. However, some teachers based their views of absenteeism on certain students. Less formal assessments were used as well to measure student disrespect. For example, one teacher described conversations or meetings in which she heard about student disrespect. Another teacher's position on a school drug-alcohol treatment program gave her additional insight into this problem, and another based her answer on the same students she has had in her class several years in a row. Teachers described other informal information sources as they computed answers to a question about student unpreparedness.

Principals' Information Sources Principals did not focus on individual students, as some teachers did, when they described school problems. For instance, they tracked absenteeism with reports, but they also had more access to district and state information to assess the school's performance. However, the availability of more information is not necessarily beneficial to these respondents. For example, one principal compared his school absentee rates to those in another metropolitan area afflicted with much greater absentee problems. This process considerably reduced his concern about the nature of the absentee problem.

Student grades were used to understand apathy. However, this source is subject to variation among schools, due to different grading procedures. In addition to information in official reports, principals described referrals from other teachers that impacted their perceptions of school problems.

Official policies and structures in the school, such as grade level, "tracking" of students, and tuition payments also affect perceptions of school problems. One private school required parental involvement, which directly affected the extent of this "problem" at one school.

Teachers and principals both provide valuable information about school problems. Teachers seem to have a clearer view of problems at the classroom level, while principals more often rely on formal reports that describe the entire school. Both points of view are valuable; one is broader while the other more able to describe the incidents in greater detail.

CONCLUSIONS

Essentially, this study sought to determine whether teachers or principals may provide better information to establishment surveys. The situation is not simple, since in general neither respondent is advantaged over the other. Understanding the complexity of the response process is the key to understanding the data. Cognitive interviews revealed that in most cases both teachers and principals supply valuable information for establishment surveys. Each of the four areas described above must be attended to by survey researchers as they construct questionnaires that measure teacher perceptions of influence on school operations. In effect the themes serve as mediating variables that should be measured when teachers or principals are selected.

♦ *Teachers' and Principals' Role(s)* Principal and teacher answers were both valuable. However, the participation of the respondents on different kinds of programs or committees, their teaching assignment, or other roles in the school, affected the quality of their answers. While all answers about teacher influence were valid, respondents who were members of committees had higher ratings of their influence. In large-scale establishment surveys such as SASS, where several teachers from one school may be sampled, such circumstances must be accounted for to understand the quality of the responses.

♦ *Non-Institutional Phenomena and Personal Motivation* Teachers and principals are viable respondents, however their answers must be understood in terms of the time available to them and the motivations or morale of individuals. SASS includes

some variables that may provide insights into the morale of respondents; the time demands on teachers are well documented.

♦ *External Phenomena* Teachers may know about the actual implementation of policy while principals are more likely to know the official intent. Other authorities, such as district, state, and community organizations, can either restrain teachers and principals or provide them additional information if they are members.

♦ *Intra-school Policies and Operations* Teachers' and principals' responses to survey questions will be well-grounded when researchers consider the intent of school programs and policies versus their actual implementation, and when notions of "final authority" are more clearly addressed in question items that measure perceptions of influence.

Measuring Perceptions of Problems The measurement of school problems appears to be a less complex endeavor, particularly if survey researchers seek to understand only teachers' perceptions. Clearly, questions should be directed to the respondents who have the best information. While some principals in this study may have a bird's eye view of schools, teachers are no less able to answer for their classes, which are certainly central units of a school. The cognitive interviews improved our understanding of the sources of information used by respondents at different levels of an organization. The research also provided some explanation for the significant differences in descriptions of an establishment based on the type of respondent answering.

Survey researchers must realize and attempt to account for the complexity of micro-political phenomena operating in schools if they wish to construct valid and reliable questionnaires for schools. Currently, the Schools and Staffing Survey accounts for several intervening factors that impact on the formulation of answers pertaining to teacher influence.[6] In future administrations of SASS and other establishment surveys, questionnaires which measure the inter-relationships between roles, extra-school policies, personal motivations, and school-level workings will enhance the quality of data.

SUMMARY

The research questions helped to organize the conclusions of this study:

1. How do teachers' perceptions compare to principals'?
   Attachment 2 (below) shows that teachers and principals differ considerably when describing teachers' influence on various school operations. Differences are also evident when each respondent describes school problems.

2. Why are their perceptions different?
   Teacher and principal perceptions of influence and school problems are different due to several intervening factors. These factors include principal and teacher roles, personal motivations, external phenomena (e.g., community authorities), and unofficial day-to-day school operations within and outside of classrooms.

3. Are teachers or principals in more advantageous positions to answer certain questions? How does this impact data quality and analysis of data?
   This study cannot conclude that it is more advantageous to query one type of respondent than another. In general, both teachers and principals provide useful data. There are some exceptions, such as teachers who are advantaged in observing the actual implementation of policies as they vary from their intended effect.

References

Ball, Stephen J. (1987). The micro-politics of the school: towards a theory of school organization. New York, NY: Methuen.

Blase, Joseph, ed. (1991). The Politics of life in schools: power, conflict, and cooperation. Newbury Park, CA: Sage.

Christianson, A., and Tortora, R. (1995). "Issues in Surveying Business: an International Survey." Chapter 14 in Business Survey Methods, edited by Cox, Binder, Chinappa, Christianson, Colledge, Kott. New York: John Wiley and Sons.

Edwards, W.S., and Cantor, D. (1991). Toward a response model in establishment surveys. In P. Biemer, R. Groves, G. Forsman, L. Lyberg, N. Mathiowetz, and S. Sudman (Eds.), *Measurement Error in Surveys* (pp. 211-236). New York: John Wiley.

Sudman, S., Bradburn, N.M., and Schwarz, N. (1996). *Thinking About Answers: The Application of Cognitive Processes to Survey Methodology* (pp. 227-243). San Francisco: Jossey-Bass. (chapter on proxy reporting)

[6] E.g., the school form asks about strategic plans, multiple teaching assignments, the availability of different programs in schools.

ATTACHMENT 1: SASS Field Test Items Examined

The following items were asked of both teachers and principals:

1. To what extent is each of the following matters a problem in this school? Indicate whether it is a serious problem, a moderate problem, a minor problem, or not a problem in this school.

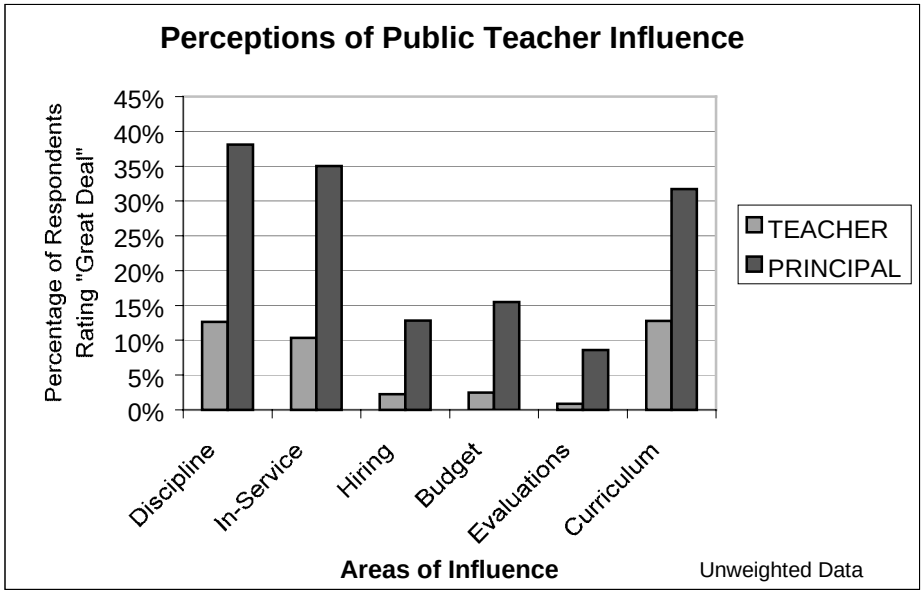
a. Student tardiness
b. Student absenteeism
c. Student disrespect for teachers
d. Lack of parent involvement
e. Student apathy
f. Students coming to school unprepared to learn

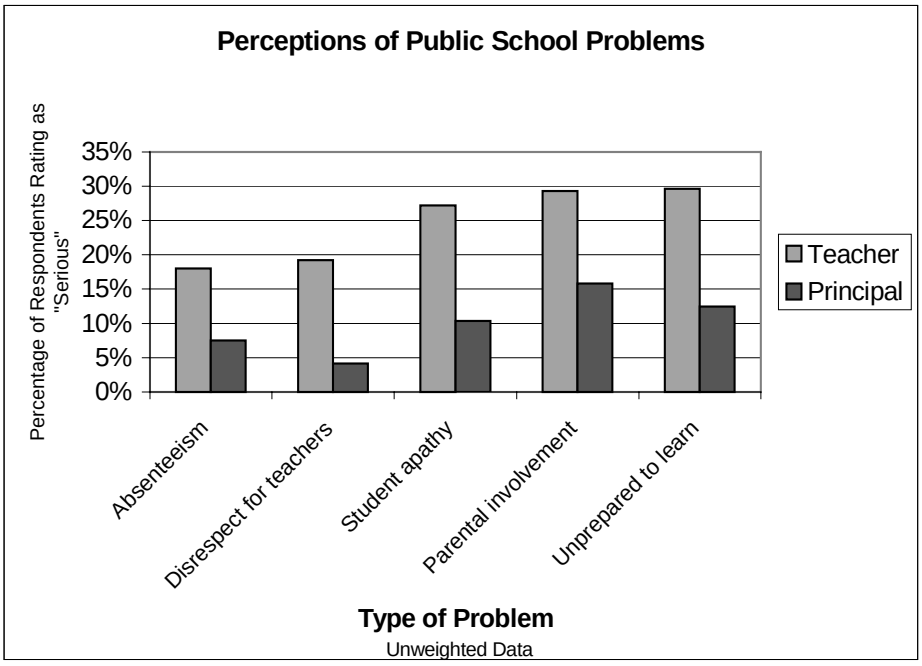
---

2. Using the scale 0-5, where 0 is "No influence" and 5 is "A great deal," how much ACTUAL influence do you think teachers have on decisions concerning the following activities?

	No Influence 1	2	3	4	A great deal 5
a. Setting performance standards for students of this school					
b. Establishing curriculum at this school					
c. Determining the content of in-service professional development programs for teachers in this school					
d. Evaluating teachers at this school					
e. Hiring new full-time teachers at this school					
f. Setting discipline policy at this school					
g. Deciding how your school budget will be spent					

ATTACHMENT 2: Summary Results from 1993-94 Schools and Staffing Survey
Unweighted Data





DESIGNING SURVEYS FOR THE NEXT MILLENNIUM: INTERNET QUESTIONNAIRE DESIGN ISSUES

Andrew Zukerberg, Elizabeth Nichols, Heather Tedesco
Elizabeth Nichols, U.S. Bureau of the Census, Washington, DC 20233

Introduction

As survey forms have changed from paper to computer administered modes, research has focused on the effect of computers on survey design. In recent years technology has allowed the creation of computerized self-administered questionnaires (CSAQs). Functionality programmed into CSAQs such as skip patterns, sound, and additional probing for questionable answers, allows researchers to bridge the interactivity gap between paper self-administered and interviewer-administered questionnaires. These changes can improve data quality and decrease respondent burden, especially in complex questionnaires (e.g., Turner et al., 1998; Bloom, 1998) In addition to these benefits, Pilon and Craig (1988) documented a "novelty effect" with CSAQs, suggesting increased response rates over that of a traditional self-administered paper questionnaire.

The growth and acceptance of the Internet as a mode of communication and commerce has opened new possibilities for the implementation of CSAQs. However, little guidance is provided on how to best harness this power to improve survey instruments. The amount and type of functionality to use in a CSAQ depends on a number of issues including the complexity of the survey, the respondent's computer experience and hardware and software limitations. The Internet opens more possibilities for functionality and may generate a new novelty effect as households and businesses connect for the first time. It also brings new complexities in browser, speed of access, and operating system configurations.

This paper presents the questionnaire design challenges we faced adapting a paper self-administered questionnaire to an Internet self-administered questionnaire (ISAQ). We start with a brief discussion of the survey on which this research is based. This is followed by a review of previous research on ISAQ design. The inconclusive and sometimes contradictory findings provided little guidance for proceeding with our instrument design. We report our design findings from our usability testing and the modest experiment we conducted to test alternative design options. The paper concludes with rudimentary recommendations for designing an ISAQ.

The Library Media Center Survey

The research reported here is based on a pretest of the Library Media Center (LMC) Survey, one of seven components of the Schools and Staffing Survey (SASS). The U.S. Bureau of the Census conducts SASS on behalf of the National Center for Education Statistics. The survey was last conducted in 1993-1994 and will be conducted again in 1999–2000 with approximately 13,000 schools. In the 1993–94 administration of SASS, the LMC component had one of the lowest response rates of the seven components. For the 1998–99 LMC pretest, an ISAQ was designed as an alternative reporting option to the paper questionnaire. This was done in an attempt to boost response rates by providing a more convenient reporting option. The LMC appeared to be a good candidate for an ISAQ because many school libraries have Internet access and because the questions in the LMC are complex and involve lookup from library records. In many libraries these records are stored on a computer. For respondents in these libraries, an Internet option brings the questionnaire to the place where the records are kept.

Literature Review of ISAQ Design Issues

Unlike a paper form where respondents determine the navigational path as they complete the questionnaire, researchers using an ISAQ must determine how they want respondents to navigate before the instrument is fielded. An ISAQ can be programmed to exhibit more or less control over the path respondents take. For instance, some ISAQs are programmed so that the respondent completes one question, submits his/her answer and receives the next question. Other ISAQs are designed more like paper questionnaires where questions can be answered in any order and at any time. Still others use a combination of the two approaches. Choosing the optimal navigation for an ISAQ is important because it becomes the foundation of the instrument design.

Some researchers have recently begun to consider the impact of different navigational strategies. One typical ISAQ design uses a scroll bar for navigation. This design allows respondents to navigate as they would through a word processor. A variation of this design employs a menu bar that allows respondents to jump between sections of the form. Interestingly, in an expert review of a scroll form with a menu frame and a scroll form without a menu frame, Sweet et al. (1997b) documented a strong reviewer preference for the menu version even though reviewers did not use the menu to navigate. Sweet et al. (1997a) also conducted an expert review of two versions of an ISAQ. In one version, respondents proceeded question by question and could not quickly skip questions. The other version used a scroll bar and "hot links" which allowed reviewers to jump within the form. Reviewers in the study suggested a main menu bar on the left side of the form would improve navigation for respondents because it would allow quick access to any question. These

reviewers also felt that combining "hot links" within the frame containing the questions using scroll-based navigation would impede users' understanding of navigation through the form. Bates and Nichols (1998) examined a Web application for hiring new employees. Unlike the previous research that used expert reviewers, this test was conducted with respondents. This application contained a scroll bar with a menu frame. Usability testing of the instrument showed that the scroll bar, not the menu in the left frame, was the primary means of navigation for subjects. However, this test did not assess subject's preferences for the menu frame, leaving open the possibility that the menu provides some benefit beyond navigation. Dillman et al. (1998) suggested principles of Web questionnaire layout and design including presenting questions in the conventional format for paper questionnaires; providing specific instructions regarding the necessary computer actions; and designing scroll-based rather than screen-based questionnaires. They reported the results of a preliminary study examining two different CSAQ designs and found that the design adhering more closely to their set of guidelines had a higher completion rate and better overall performance. However, they did not systematically examine the effect of adhering to these individual principles.

Implementation of edit checks is another key design decision that a researcher conducting an ISAQ must face. Edit checks are messages informing the respondent that an entry may be out of scope (called a range check), an answer is missing, there is an inconsistency between two or more answers (called an inter-item check), or the format of the answer is incorrect (called an alpha/numeric check). The placement, wording and action taken by the edit check could have implications for data quality. There has been very little research examining the implementation of edit messages in CSAQs. One exception is a study by Rosen et al. (1998), in which the number of edit messages received by respondents was tallied. However, this research was solely descriptive and did not examine different types of edit messages or strategies for edit checks, so the cause of these edit failures (i.e., overly restrictive range, confusing message, etc.) could not be ascertained. In addition, some reviewers in Sweet et al. (1997a) suggested inclusion of a button that would run edit checks before submission of the survey.

Methodology

The lack of systematic research with respondents left us unsure how to design our ISAQ. We decided to compare two navigational approaches and two types of edit messages experimentally before selecting the design for the LMC pretest instrument. Four different LMC ISAQs were created. These versions shared the same questions, help content, and basic attributes, but differed in the combination of navigation and edit checks used.

Navigational Designs

Two different navigational designs were implemented. Figure 1 is called the scroll-based design, where all of the questions are on one page. Respondents can tab between questions or use the scroll bar on the right to move through the form.

Figure 1: Scroll-based design

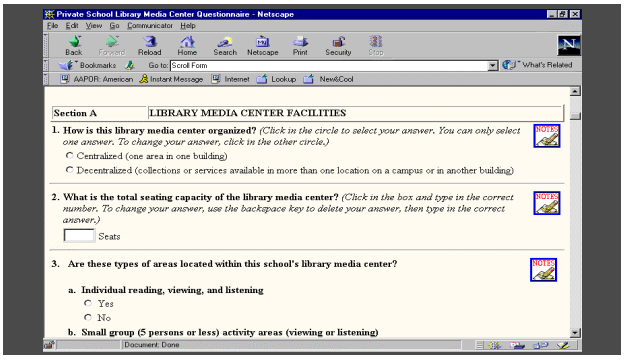
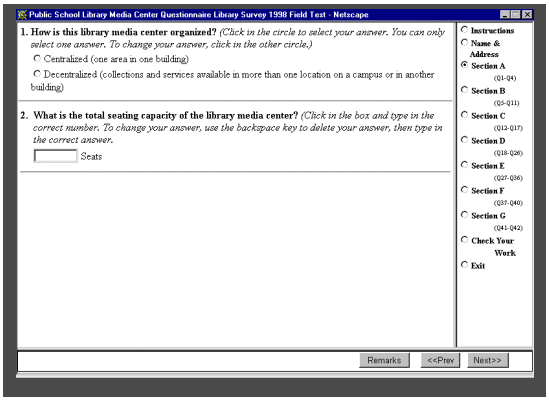


Figure 2 is called the screen-based design. This design allowed respondents to use Next and Previous buttons and/or a menu bar to navigate. A filled circle in the bar indicates the respondent's location in the questionnaire. Respondents could click on the circles to move between sections, but not individual questions within the section. The next and previous buttons allow the respondent to navigate between screens of questions.

Figure 2: Screen-based design



Edit Message Designs

There were two item-level edit procedures implemented. These edits were run when the respondent pressed Tab, Enter, or clicked into the next response box. The edit message appeared in a small box on the screen and prevented respondents from entering data until the box was cleared. Figure 3 is the item-level edit referred to as passive, which displayed the error message in a small box with an OK button. Clicking OK would clear the edit message but did not change any data on the survey form.

Figure 3: Example of a Passive Edit Message

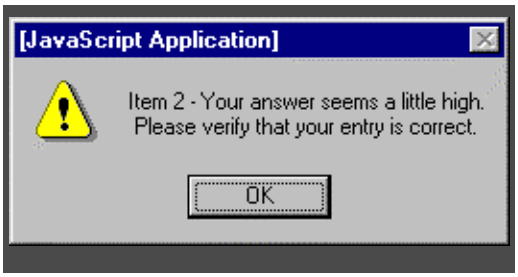
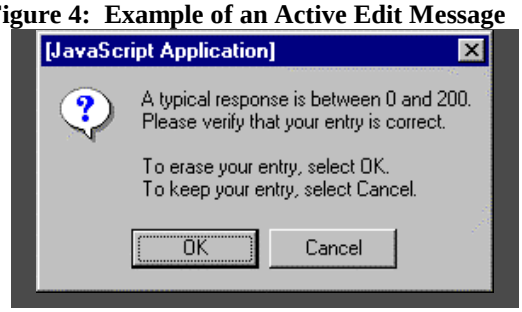


Figure 4 is the item-level edit referred to as active. Within the message box, respondents could choose either OK which would clear the suspicious entry or Cancel which functioned like the passive edit and did not change their entry. The active edits provided more interaction with the respondent. We expected respondents to be more likely to change their data in response to the active edits.

Figure 4: Example of an Active Edit Message



For both edit procedures, a response to the edit message box was necessary in order to clear it and return to the questionnaire. The instrument was programmed to accept an answer after the edit message was cleared. However, if the respondent erased and retyped the same answer, the same edit message was triggered again.

Each of the four ISAQ versions contained a different combination of the two designs. The scroll and screen designs using the passive edits were developed and tested first. We discovered some problems during the first round and made several modifications to the forms before the next round of testing. In the next round, we tested the modified scroll and screen designs using active edits.

To compare these different forms, we conducted usability tests with thirteen school librarians from the metropolitan Washington D.C. area. Interviews were conducted at the Census Bureau's usability lab and in respondents' libraries. Respondents completed an ISAQ version in the presence of a trained usability interviewer. All interviews were videotaped with respondent's consent. After completing the questionnaire, respondents were asked to rate their experience on a number of subjective scales. The interviewer then conducted a retrospective interview, focusing on content and questionnaire design issues as well as addressing any potential problems they observed during the interview. The videotapes were reviewed following an observation protocol to characterize the respondent's navigational approach and to count the number of times s/he utilized instrument features. We were thus able to combine performance data from the actual questionnaire with subjective satisfaction ratings and cognitive interviews regarding question content. This usability testing should be viewed as exploratory research. The small sample size greatly limits our ability to generalize based on this research. Table 1 shows the distribution of respondents by treatment.

Table 1: Distribution of Respondents by Treatment

Navigation	Passive	Active	Total
Scroll	3	4	7
Screen	3	3	6
Total	6	7	13

Results

Navigational Approach

Table 2 compares the overall reaction from the subjective satisfaction scales for those who used the scroll design to those who completed the screen design.

Table 2: Mean subjective rating for the overall reaction to the system (combined rounds)

Overall reaction to the system:	Scroll (n=7)	Screen (n=6)
1=Dissatisfied...9=Satisfied	7.14	6.83
1=Difficult...9=Easy	7.17	7.33
1=Frustrating...9=Satisfying	6.71	6.17
1=Dull...9=Stimulating	6.71	7.00
1=Rigid...9=Flexible	7.14	6.83
Average	6.97	6.83

The overall satisfaction on each of the points was high for both designs. The average rating for the scroll was 6.97 versus 6.83 for the screen design. Based on these subjective ratings, we conclude that respondents found

both ISAQ relatively easy to use. Due to the testing methodology, respondents did not compare designs, which might have produced different results.

In addition to their overall reactions to the system, respondents were asked to rate various aspects of the navigation as shown in Table 3. These results again demonstrate that respondents had positive reactions to the navigation in both designs.

Table 3: Mean subjective rating for navigation

How would you rate	Scroll (n=7)	Screen (n=6)
Screen organization?..... 1=Confusing... 9=Clear	7.71	7.00
Sequence of screens?..... 1=Confusing...9=Clear	8.14	7.00
Screen appearance?..... 1=Poor...9=Excellent	7.29	7.33
Moving within a screen?..... 1=Difficult...9=Easy	7.43	8.50
Moving between screens?..... 1=Difficult...9=Easy	7.71	8.67
Backing up?..... 1=Difficult...9=Easy	7.43	8.60

Many of the metrics captured on the videotape failed to demonstrate differences between the two designs. No glaring differences were found between the two designs for the number of missing responses or the number of questions skipped incorrectly, nor for the time it took to complete the forms. On average, it took 32.38 minutes to complete the scroll form and 35.10 minutes to complete the screen design. Table 4 shows similar ratings regarding 'learning to operate the system' or 'time to learn to use the system' for both navigation designs.

Table 4: Mean subjective rating for learning the system

How would you rate	Scroll (n=7)	Screen (n=6)
Learning to operate the system? 1=Difficult...9=Easy	7.71	7.83
Time to learn to use the system? 1=Slow...9=Fast	7.29	7.60

However, the small sample size may have masked differences between the two approaches. Our observations highlighted differences in how respondents interacted with the forms. For example, respondents sometimes got lost in the forms. Four of the seven respondents in the scroll version got lost at some point. In 3 of the 4 cases, they became lost when they clicked in the wrong place on the scroll bar causing the screen to jump to another location. However, they appeared to know what happened and recovered quickly, since the average time lost was only 29.25 seconds. Four of the six screen-based respondents also became lost, but with an average time lost of 7.14 minutes. Two of these four cases used the menu bar to move between sections, but were unaware that each section contained multiple questions on different screens. (They were considered lost the entire time they were unaware each section contained multiple questions.) Thus, they responded to the first question of each section (a total of 9 questions out of 42), which meant around 73 items were not completed. A first round respondent eventually realized the error and used the Next/Previous buttons to access the remaining questions. Because of her problems, we modified the ISAQ in the second round and moved the Next and Previous buttons to the right side of the bottom frame, directly below the menu bar. Despite these changes a respondent in the second round made the same navigational error. This person claimed during the retrospective interview to see the Next and Previous buttons, but believed that they were equivalent to the menu bar function. She stated that the menu bar moved between sections, as did the Next and Previous. It was unclear to her that each section contained multiple questions.

Most respondents had very little trouble with the scroll version. All seven respondents primarily used the scroll bar arrows to navigate through the survey. Some respondents preferred to answer all the questions on one screen and then scroll down to bring up a new set of questions. Others tended to answer one question, scroll down to the next question, and so on. We discovered during the usability tests that navigating with the Tab key in a scroll form is not ideal because it moves the respondent directly to the next entry field. When respondents are at the bottom of the screen, using the Tab key repositions the display so the next field is at the top of the screen and the question text doesn't display. Thus, with a scroll design, the researcher loses some control over the display of questions.

We noticed that all respondents completed their answers in a sequential manner. Again, this is what we would typically expect, but it is difficult to say how much of this strategy was affected by the usability test itself. It was apparent that some of the responses would have required record look-up. Since they were instructed to make their best guess, we cannot say exactly what they would have done outside the testing environment. Perhaps they would have skipped questions, or utilized another navigational scheme. This limits our conclusions to some extent, but we assume this would have affected both screen and scroll respondents equally.

Edit Messages

Our study attempted to understand the implications of different item-level edit messages on respondent burden and data quality. We compared edit messages that only state the possible error (passive) to edit messages that allow the respondent to automatically change data on the form (active). Table 5 shows the average number of messages displayed per interview and the average number of messages ignored (we considered a message ignored if the respondent did not change the entry based on the edit message or write any explanation of the failed entry in the Remarks section). All respondents received at least one edit message.

Table 5: Average # of edit messages received and ignored

	Passive (n=6)	Active (n=7)
Avg. # of messages	4.00	5.14
Avg. # of messages ignored	1.5	2.86

Based on these numbers alone, the differences in the active and passive edits do not appear to have an effect on the action taken by the respondent. As in the navigation comparison, our observations highlighted differences in the respondents' interactions with these edits.

Interestingly, we found that some respondents never read the item-level edit messages. This happened once during the passive edit testing. When presented with an edit failure, the respondent quickly clicked OK. When retrospectively probed, she had no recollection of any messages appearing on the screen. We hypothesized that during active edit testing, respondents would be more likely to read the message since there would be two choices, OK and Cancel, and a selection was required Unfortunately, a few respondents continued to ignore the active edit messages.

For those who read the active messages, many did not fully understand the difference between OK and Cancel. Some acknowledged this during the retrospective interview; we concluded this for others when they retyped their original answer after erasing it by clicking OK. Prior experiences with the words 'OK' and 'Cancel' may have influenced their understanding of these messages. For example, one respondent's definition of Cancel was closing the application and canceling all of her responses. (We had earlier thought that asking a question with a Yes/No response might be more straightforward.)

Some of our edit messages were designed to prevent out-of-range answers. The goal of the range check was to inform respondents about possible errors, while allowing them to keep the data if it was correct. Respondents had surprising reactions to the range checks. Two of the six respondents during the passive edit testing changed their answer to reflect the upper bound of a range check when their responses failed. We found out during the retrospective interview that they did not think they could continue with the questionnaire until they resolved the message. For example, one message stated, 'Enter a number between 1 and 10. Please verify your entry.' We modified the range message for the active edit test. The revised range check message stated, 'A typical response is between 1 and 10. Please verify your entry.' We also modified one frequently failed message to read 'Your response seems [FILL high/low]. Please verify your response.' These revisions did not alleviate the problem of respondents conforming to the range checks. Active edit respondents changed answers in the range checks with both the ranges specified and the high/low range message. One respondent noted during the retrospective interview that the edit messages gave her the feeling that her answer was 'unreasonable' and that she had to change her response. Clearly, these messages were having a negative impact on data quality.

A number of respondents erased and retyped their original answer after receiving an error message. This reactivated the error message causing respondents to become frustrated and leave subsequent answers blank. Another frustration for respondents occurred in the scroll instrument. Edits were activated when the respondent clicked on the next response. So, if a respondent answered question 1 and then decided to scroll down and answer question 16 next, the edit for item 1 would be run when they entered the answer to item 16 and any messages would appear at item 16. In two cases, the respondent entered an answer in error, but scrolled well past the field to perform the next action. Since the screen did not jump back to the item in question, and the item number was not contained within the message text, the respondent did not make a connection to the item in error.

Discussion

As survey researchers, a number of the observation measures caused us great concern about using the screen based design or active edits for the upcoming LMC survey. The fact that respondents missed questions in the screen based form and did not realize that their data were being changed by the active edits led us to conclude that these approaches should not be instituted as they were tested. However, we do not advise abandoning these approaches without further research. Our testing demonstrated that minor changes in wording or formatting can have a large impact in how respondents interact with an ISAQ. A screen-based approach can generate questions on the fly, allowing skip patterns to be fully automated. Active edits can provide more interactive feedback to respondents about their answers. It is possible that a

different manifestation of these approaches will produce better results.

This testing experience highlighted some pitfalls to avoid in designing a Web questionnaire. The navigation used in the screen based design was not consistent. Respondents could use the Next and Previous buttons to access all the questions, but if they chose to use the menu bar, they also had to use those buttons to access the remaining questions in each section. The instructions were not sufficient for overcoming this navigational inconsistency. Whatever approach is utilized, the instructions must clearly explain how to navigate the instrument, or more ideally, the navigation within the instrument must be self-evident.

The necessary iterative modification and retesting of active edit messages would have been time consuming. While still very important, the messages become less critical with the passive edit, since data are not changed. With the active edit, the questionable data disappeared once the respondent clicked OK on the edit message. In the test some respondents did not reenter data. This could lead to higher missing data rates in production. We noticed that when data entries were erased, respondents were sometimes reluctant to guess again, or to enter data for other items for which they were unsure. For example, after receiving repeated error messages on a range field, one respondent left the field blank during the passive edit testing. After two repeated error messages on a range field during the active edit testing, another respondent left many questions blank. Given the reaction that some had to the edit messages, we suspect that automatically erasing data could increase the item nonresponse rate.

There is a broader question about the placement of edit messages that our research did not address. Some ISAQs run edits on the question, while others run them at the end of the survey as part of a submit feature. We imbedded both features in the designs, but cannot conclude whether one is better (in terms of data quality and/or respondent burden) since both seemed to have limitations. It was apparent that the wording of edits at the end of the questionnaire is equally as important as when they are run on the question. In addition, respondents need to be able to navigate quickly back to the item in question to resolve any issue, regardless of the placement of the edit. They also need to quickly identify which item a message refers to. This is critical in a scroll design since the respondent could be past the item when the edit is invoked.

Survey researchers have a vast amount of research from mail and interviewer-administered surveys to draw on in designing surveys. Internet surveys provide design opportunities that are unavailable in these older methods. However, these surveys also create new challenges and necessitate constant re-evaluation as computer programming capacity increases and hardware improves. Regardless of the navigational approach selected or types of interactivity built in, we cannot emphasize enough the importance of iterative testing of the instrument with respondents. These tests provide valuable insights on how respondents interact with the questionnaire and are useful for identifying programming errors.

Note: This paper reports the results of research and analysis undertaken by Census Bureau staff. It has undergone a more limited review than official Census Bureau publications. This report is released to inform interested parties of research and to encourage discussion.

The authors thank Steve Tourkin, Chris Cooper, Robin King, Robin Goldberg, Adrien Hnat, Barbara Sedivi and Howard Kanarek for their assistance in the design and testing of the instrument and Nancy Bates, Susan Ciochetto and Cleo Redline for their helpful comments in the preparation of this paper.

References

Bates, N. and E. Nichols. 1998. "The Census Bureau WWW Hiring Questionnaire: A Case Study for Usability Testing." 1998 Proceedings of the Section on Survey Research Methods of the American Statistical Association, p. 360-365.

Bloom, D. 1998. "Technology, Experimentation, and the Quality of Survey Data." *Science* 280:847-848.

Dillman, D.A., R.D. Tortora, and D. Bowker. 1998. "Principles for Constructing Web Surveys." http://survey.sesrc.wsu.edu/dillman/papers.htm

Pilon, T.L. and Craig, N.C. 1988. "Disks-by-Mail: a New Survey Modality," Proceedings of the 1988 Sawtooth Software Conference on Perpetual Mapping, Conjoint Analysis and Computer Interviewing, Sun Valley, ID: Sawtooth Software.

Rosen, R., C. Manning, and L. Harrell. 1998. "Web-based Data Collection in the Current Employment Statistics Survey." 1998 Proceedings of the Section on Survey Research Methods of the American Statistical Association, p. 354-59.

Sweet, E., B. Sedivi, K. Marquis, and F. Nash. 1997a. "Results of Expert Review of Two Internet R&D Questionnaires." Human-Computer Interaction Report Series #1. Center for Survey Methods Research.

Sweet, E., B. Sedivi, N. VanDerveer, E. Soper and Z. Zhang. 1997b. "Results from the expert review of the defunct Decennial Census electronic short form for the Web." Human-Computer Interaction Report Series #3. Center for Survey Methods Research.

Turner, C.F., L. Ku, S.M. Rogers, L.D. Lindberg, J.H. Pleck, and F.L. Sonenstein. 1998. "Adolescent Sexual Behavior, Drug Use, and Violence: Increased

Reporting with Computer Survey Technology." *Science* 280:867-873.

This page intentionally left blank.

Listing of NCES Working Papers to Date

Working papers can be downloaded as pdf files from the NCES Electronic Catalog (http://nces.ed.gov/pubsearch/). You can also contact Sheilah Jupiter at (202) 502-7444 (sheilah_jupiter@ed.gov) if you are interested in any of the following papers.

Listing of NCES Working Papers by Program Area

No.	Title	NCES contact
Baccalaureate and Beyond (B&B)		
98-15	Development of a Prototype System for Accessing Linked NCES Data	Steven Kaufman
Beginning Postsecondary Students (BPS) Longitudinal Study		
98-11	Beginning Postsecondary Students Longitudinal Study First Follow-up (BPS:96-98) Field Test Report	Aurora D'Amico
98-15	Development of a Prototype System for Accessing Linked NCES Data	Steven Kaufman
1999-15	Projected Postsecondary Outcomes of 1992 High School Graduates	Aurora D'Amico
Common Core of Data (CCD)		
95-12	Rural Education Data User's Guide	Samuel Peng
96-19	Assessment and Analysis of School-Level Expenditures	William J. Fowler, Jr.
97-15	Customer Service Survey: Common Core of Data Coordinators	Lee Hoffman
97-43	Measuring Inflation in Public School Costs	William J. Fowler, Jr.
98-15	Development of a Prototype System for Accessing Linked NCES Data	Steven Kaufman
1999-03	Evaluation of the 1996-97 Nonfiscal Common Core of Data Surveys Data Collection, Processing, and Editing Cycle	Beth Young
Decennial Census School District Project		
95-12	Rural Education Data User's Guide	Samuel Peng
96-04	Census Mapping Project/School District Data Book	Tai Phan
98-07	Decennial Census School District Project Planning Report	Tai Phan
Early Childhood Longitudinal Study (ECLS)		
96-08	How Accurate are Teacher Judgments of Students' Academic Performance?	Jerry West
96-18	Assessment of Social Competence, Adaptive Behaviors, and Approaches to Learning with Young Children	Jerry West
97-24	Formulating a Design for the ECLS: A Review of Longitudinal Studies	Jerry West
97-36	Measuring the Quality of Program Environments in Head Start and Other Early Childhood Programs: A Review and Recommendations for Future Research	Jerry West
1999-01	A Birth Cohort Study: Conceptual and Design Considerations and Rationale	Jerry West
2000-04	Selected Papers on Education Surveys: Papers Presented at the 1998 and 1999 ASA and 1999 AAPOR Meetings	Dan Kasprzyk
Education Finance Statistics Center (EDFIN)		
94-05	Cost-of-Education Differentials Across the States	William J. Fowler, Jr.
96-19	Assessment and Analysis of School-Level Expenditures	William J. Fowler, Jr.
97-43	Measuring Inflation in Public School Costs	William J. Fowler, Jr.
98-04	Geographic Variations in Public Schools' Costs	William J. Fowler, Jr.
1999-16	Measuring Resources in Education: From Accounting to the Resource Cost Model Approach	William J. Fowler, Jr.
High School and Beyond (HS&B)		
95-12	Rural Education Data User's Guide	Samuel Peng
1999-05	Procedures Guide for Transcript Studies	Dawn Nelson
1999-06	1998 Revision of the Secondary School Taxonomy	Dawn Nelson
HS Transcript Studies		
1999-05	Procedures Guide for Transcript Studies	Dawn Nelson
1999-06	1998 Revision of the Secondary School Taxonomy	Dawn Nelson

No.	Title	NCES contact
International Adult Literacy Survey (IALS)		
97-33	Adult Literacy: An International Perspective	Marilyn Binkley
Integrated Postsecondary Education Data System (IPEDS)		
97-27	Pilot Test of IPEDS Finance Survey	Peter Stowe
98-15	Development of a Prototype System for Accessing Linked NCES Data	Steven Kaufman
National Assessment of Adult Literacy (NAAL)		
98-17	Developing the National Assessment of Adult Literacy: Recommendations from Stakeholders	Sheida White
1999-09a	1992 National Adult Literacy Survey: An Overview	Alex Sedlacek
1999-09b	1992 National Adult Literacy Survey: Sample Design	Alex Sedlacek
1999-09c	1992 National Adult Literacy Survey: Weighting and Population Estimates	Alex Sedlacek
1999-09d	1992 National Adult Literacy Survey: Development of the Survey Instruments	Alex Sedlacek
1999-09e	1992 National Adult Literacy Survey: Scaling and Proficiency Estimates	Alex Sedlacek
1999-09f	1992 National Adult Literacy Survey: Interpreting the Adult Literacy Scales and Literacy Levels	Alex Sedlacek
1999-09g	1992 National Adult Literacy Survey: Literacy Levels and the Response Probability Convention	Alex Sedlacek
National Assessment of Educational Progress (NAEP)		
95-12	Rural Education Data User's Guide	Samuel Peng
97-29	Can State Assessment Data be Used to Reduce State NAEP Sample Sizes?	Steven Gorman
97-30	ACT's NAEP Redesign Project: Assessment Design is the Key to Useful and Stable Assessment Results	Steven Gorman
97-31	NAEP Reconfigured: An Integrated Redesign of the National Assessment of Educational Progress	Steven Gorman
97-32	Innovative Solutions to Intractable Large Scale Assessment (Problem 2: Background Questionnaires)	Steven Gorman
97-37	Optimal Rating Procedures and Methodology for NAEP Open-ended Items	Steven Gorman
97-44	Development of a SASS 1993-94 School-Level Student Achievement Subfile: Using State Assessments and State NAEP, Feasibility Study	Michael Ross
98-15	Development of a Prototype System for Accessing Linked NCES Data	Steven Kaufman
1999-05	Procedures Guide for Transcript Studies	Dawn Nelson
1999-06	1998 Revision of the Secondary School Taxonomy	Dawn Nelson
National Education Longitudinal Study of 1988 (NELS:88)		
95-04	National Education Longitudinal Study of 1988: Second Follow-up Questionnaire Content Areas and Research Issues	Jeffrey Owings
95-05	National Education Longitudinal Study of 1988: Conducting Trend Analyses of NLS-72, HS&B, and NELS:88 Seniors	Jeffrey Owings
95-06	National Education Longitudinal Study of 1988: Conducting Cross-Cohort Comparisons Using HS&B, NAEP, and NELS:88 Academic Transcript Data	Jeffrey Owings
95-07	National Education Longitudinal Study of 1988: Conducting Trend Analyses HS&B and NELS:88 Sophomore Cohort Dropouts	Jeffrey Owings
95-12	Rural Education Data User's Guide	Samuel Peng
95-14	Empirical Evaluation of Social, Psychological, & Educational Construct Variables Used in NCES Surveys	Samuel Peng
96-03	National Education Longitudinal Study of 1988 (NELS:88) Research Framework and Issues	Jeffrey Owings
98-06	National Education Longitudinal Study of 1988 (NELS:88) Base Year through Second Follow-Up: Final Methodology Report	Ralph Lee
98-09	High School Curriculum Structure: Effects on Coursetaking and Achievement in Mathematics for High School Graduates—An Examination of Data from the National Education Longitudinal Study of 1988	Jeffrey Owings
98-15	Development of a Prototype System for Accessing Linked NCES Data	Steven Kaufman
1999-05	Procedures Guide for Transcript Studies	Dawn Nelson
1999-06	1998 Revision of the Secondary School Taxonomy	Dawn Nelson
1999-15	Projected Postsecondary Outcomes of 1992 High School Graduates	Aurora D'Amico

No.	Title	NCES contact
National Household Education Survey (NHES)		
95-12	Rural Education Data User's Guide	Samuel Peng
96-13	Estimation of Response Bias in the NHES:95 Adult Education Survey	Steven Kaufman
96-14	The 1995 National Household Education Survey: Reinterview Results for the Adult Education Component	Steven Kaufman
96-20	1991 National Household Education Survey (NHES:91) Questionnaires: Screener, Early Childhood Education, and Adult Education	Kathryn Chandler
96-21	1993 National Household Education Survey (NHES:93) Questionnaires: Screener, School Readiness, and School Safety and Discipline	Kathryn Chandler
96-22	1995 National Household Education Survey (NHES:95) Questionnaires: Screener, Early Childhood Program Participation, and Adult Education	Kathryn Chandler
96-29	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	Comparison of Estimates from the 1995 National Household Education Survey (NHES:95)	Kathryn Chandler
97-02	Telephone Coverage Bias and Recorded Interviews in the 1993 National Household Education Survey (NHES:93)	Kathryn Chandler
97-03	1991 and 1995 National Household Education Survey Questionnaires: NHES:91 Screener, NHES:91 Adult Education, NHES:95 Basic Screener, and NHES:95 Adult Education	Kathryn Chandler
97-04	Design, Data Collection, Monitoring, Interview Administration Time, and Data Editing in the 1993 National Household Education Survey (NHES:93)	Kathryn Chandler
97-05	Unit and Item Response, Weighting, and Imputation Procedures in the 1993 National Household Education Survey (NHES:93)	Kathryn Chandler
97-06	Unit and Item Response, Weighting, and Imputation Procedures in the 1995 National Household Education Survey (NHES:95)	Kathryn Chandler
97-08	Design, Data Collection, Interview Timing, and Data Editing in the 1995 National Household Education Survey	Kathryn Chandler
97-19	National Household Education Survey of 1995: Adult Education Course Coding Manual	Peter Stowe
97-20	National Household Education Survey of 1995: Adult Education Course Code Merge Files User's Guide	Peter Stowe
97-25	1996 National Household Education Survey (NHES:96) Questionnaires: Screener/Household and Library, Parent and Family Involvement in Education and Civic Involvement, Youth Civic Involvement, and Adult Civic Involvement	Kathryn Chandler
97-28	Comparison of Estimates in the 1996 National Household Education Survey	Kathryn Chandler
97-34	Comparison of Estimates from the 1993 National Household Education Survey	Kathryn Chandler
97-35	Design, Data Collection, Interview Administration Time, and Data Editing in the 1996 National Household Education Survey	Kathryn Chandler
97-38	Reinterview Results for the Parent and Youth Components of the 1996 National Household Education Survey	Kathryn Chandler
97-39	Undercoverage Bias in Estimates of Characteristics of Households and Adults in the 1996 National Household Education Survey	Kathryn Chandler
97-40	Unit and Item Response Rates, Weighting, and Imputation Procedures in the 1996 National Household Education Survey	Kathryn Chandler
98-03	Adult Education in the 1990s: A Report on the 1991 National Household Education Survey	Peter Stowe
98-10	Adult Education Participation Decisions and Barriers: Review of Conceptual Frameworks and Empirical Studies	Peter Stowe
National Longitudinal Study of the High School Class of 1972 (NLS-72)		
95-12	Rural Education Data User's Guide	Samuel Peng
National Postsecondary Student Aid Study (NPSAS)		
96-17	National Postsecondary Student Aid Study: 1996 Field Test Methodology Report	Andrew G. Malizio
National Study of Postsecondary Faculty (NSOPF)		
97-26	Strategies for Improving Accuracy of Postsecondary Faculty Lists	Linda Zimbler
98-15	Development of a Prototype System for Accessing Linked NCES Data	Steven Kaufman
2000-01	1999 National Study of Postsecondary Faculty (NSOPF:99) Field Test Report	Linda Zimbler

No.	Title	NCES contact
Private School Universe Survey (PSS)		
95-16	Intersurvey Consistency in NCES Private School Surveys	Steven Kaufman
95-17	Estimates of Expenditures for Private K-12 Schools	Stephen Broughman
96-16	Strategies for Collecting Finance Data from Private Schools	Stephen Broughman
96-26	Improving the Coverage of Private Elementary-Secondary Schools	Steven Kaufman
96-27	Intersurvey Consistency in NCES Private School Surveys for 1993-94	Steven Kaufman
97-07	The Determinants of Per-Pupil Expenditures in Private Elementary and Secondary Schools: An Exploratory Analysis	Stephen Broughman
97-22	Collection of Private School Finance Data: Development of a Questionnaire	Stephen Broughman
98-15	Development of a Prototype System for Accessing Linked NCES Data	Steven Kaufman
2000-04	Selected Papers on Education Surveys: Papers Presented at the 1998 and 1999 ASA and 1999 AAPOR Meetings	Dan Kasprzyk
Recent College Graduates (RCG)		
98-15	Development of a Prototype System for Accessing Linked NCES Data	Steven Kaufman
Schools and Staffing Survey (SASS)		
94-01	Schools and Staffing Survey (SASS) Papers Presented at Meetings of the American Statistical Association	Dan Kasprzyk
94-02	Generalized Variance Estimate for Schools and Staffing Survey (SASS)	Dan Kasprzyk
94-03	1991 Schools and Staffing Survey (SASS) Reinterview Response Variance Report	Dan Kasprzyk
94-04	The Accuracy of Teachers' Self-reports on their Postsecondary Education: Teacher Transcript Study, Schools and Staffing Survey	Dan Kasprzyk
94-06	Six Papers on Teachers from the 1990-91 Schools and Staffing Survey and Other Related Surveys	Dan Kasprzyk
95-01	Schools and Staffing Survey: 1994 Papers Presented at the 1994 Meeting of the American Statistical Association	Dan Kasprzyk
95-02	QED Estimates of the 1990-91 Schools and Staffing Survey: Deriving and Comparing QED School Estimates with CCD Estimates	Dan Kasprzyk
95-03	Schools and Staffing Survey: 1990-91 SASS Cross-Questionnaire Analysis	Dan Kasprzyk
95-08	CCD Adjustment to the 1990-91 SASS: A Comparison of Estimates	Dan Kasprzyk
95-09	The Results of the 1993 Teacher List Validation Study (TLVS)	Dan Kasprzyk
95-10	The Results of the 1991-92 Teacher Follow-up Survey (TFS) Reinterview and Extensive Reconciliation	Dan Kasprzyk
95-11	Measuring Instruction, Curriculum Content, and Instructional Resources: The Status of Recent Work	Sharon Bobbitt & John Ralph
95-12	Rural Education Data User's Guide	Samuel Peng
95-14	Empirical Evaluation of Social, Psychological, & Educational Construct Variables Used in NCES Surveys	Samuel Peng
95-15	Classroom Instructional Processes: A Review of Existing Measurement Approaches and Their Applicability for the Teacher Follow-up Survey	Sharon Bobbitt
95-16	Intersurvey Consistency in NCES Private School Surveys	Steven Kaufman
95-18	An Agenda for Research on Teachers and Schools: Revisiting NCES' Schools and Staffing Survey	Dan Kasprzyk
96-01	Methodological Issues in the Study of Teachers' Careers: Critical Features of a Truly Longitudinal Study	Dan Kasprzyk
96-02	Schools and Staffing Survey (SASS): 1995 Selected papers presented at the 1995 Meeting of the American Statistical Association	Dan Kasprzyk
96-05	Cognitive Research on the Teacher Listing Form for the Schools and Staffing Survey	Dan Kasprzyk
96-06	The Schools and Staffing Survey (SASS) for 1998-99: Design Recommendations to Inform Broad Education Policy	Dan Kasprzyk
96-07	Should SASS Measure Instructional Processes and Teacher Effectiveness?	Dan Kasprzyk
96-09	Making Data Relevant for Policy Discussions: Redesigning the School Administrator Questionnaire for the 1998-99 SASS	Dan Kasprzyk
96-10	1998-99 Schools and Staffing Survey: Issues Related to Survey Depth	Dan Kasprzyk
96-11	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	Predictors of Retention, Transfer, and Attrition of Special and General Education Teachers: Data from the 1989 Teacher Followup Survey	Dan Kasprzyk

No.	Title	NCES contact
96-15	Nested Structures: District-Level Data in the Schools and Staffing Survey	Dan Kasprzyk
96-23	Linking Student Data to SASS: Why, When, How	Dan Kasprzyk
96-24	National Assessments of Teacher Quality	Dan Kasprzyk
96-25	Measures of Inservice Professional Development: Suggested Items for the 1998-1999 Schools and Staffing Survey	Dan Kasprzyk
96-28	Student Learning, Teaching Quality, and Professional Development: Theoretical Linkages, Current Measurement, and Recommendations for Future Data Collection	Mary Rollefson
97-01	Selected Papers on Education Surveys: Papers Presented at the 1996 Meeting of the American Statistical Association	Dan Kasprzyk
97-07	The Determinants of Per-Pupil Expenditures in Private Elementary and Secondary Schools: An Exploratory Analysis	Stephen Broughman
97-09	Status of Data on Crime and Violence in Schools: Final Report	Lee Hoffman
97-10	Report of Cognitive Research on the Public and Private School Teacher Questionnaires for the Schools and Staffing Survey 1993-94 School Year	Dan Kasprzyk
97-11	International Comparisons of Inservice Professional Development	Dan Kasprzyk
97-12	Measuring School Reform: Recommendations for Future SASS Data Collection	Mary Rollefson
97-14	Optimal Choice of Periodicities for the Schools and Staffing Survey: Modeling and Analysis	Steven Kaufman
97-18	Improving the Mail Return Rates of SASS Surveys: A Review of the Literature	Steven Kaufman
97-22	Collection of Private School Finance Data: Development of a Questionnaire	Stephen Broughman
97-23	Further Cognitive Research on the Schools and Staffing Survey (SASS) Teacher Listing Form	Dan Kasprzyk
97-41	Selected Papers on the Schools and Staffing Survey: Papers Presented at the 1997 Meeting of the American Statistical Association	Steve Kaufman
97-42	Improving the Measurement of Staffing Resources at the School Level: The Development of Recommendations for NCES for the Schools and Staffing Survey (SASS)	Mary Rollefson
97-44	Development of a SASS 1993-94 School-Level Student Achievement Subfile: Using State Assessments and State NAEP, Feasibility Study	Michael Ross
98-01	Collection of Public School Expenditure Data: Development of a Questionnaire	Stephen Broughman
98-02	Response Variance in the 1993-94 Schools and Staffing Survey: A Reinterview Report	Steven Kaufman
98-04	Geographic Variations in Public Schools' Costs	William J. Fowler, Jr.
98-05	SASS Documentation: 1993-94 SASS Student Sampling Problems; Solutions for Determining the Numerators for the SASS Private School (3B) Second-Stage Factors	Steven Kaufman
98-08	The Redesign of the Schools and Staffing Survey for 1999-2000: A Position Paper	Dan Kasprzyk
98-12	A Bootstrap Variance Estimator for Systematic PPS Sampling	Steven Kaufman
98-13	Response Variance in the 1994-95 Teacher Follow-up Survey	Steven Kaufman
98-14	Variance Estimation of Imputed Survey Data	Steven Kaufman
98-15	Development of a Prototype System for Accessing Linked NCES Data	Steven Kaufman
98-16	A Feasibility Study of Longitudinal Design for Schools and Staffing Survey	Stephen Broughman
1999-02	Tracking Secondary Use of the Schools and Staffing Survey Data: Preliminary Results	Dan Kasprzyk
1999-04	Measuring Teacher Qualifications	Dan Kasprzyk
1999-07	Collection of Resource and Expenditure Data on the Schools and Staffing Survey	Stephen Broughman
1999-08	Measuring Classroom Instructional Processes: Using Survey and Case Study Fieldtest Results to Improve Item Construction	Dan Kasprzyk
1999-10	What Users Say About Schools and Staffing Survey Publications	Dan Kasprzyk
1999-12	1993-94 Schools and Staffing Survey: Data File User's Manual, Volume III: Public-Use Codebook	Kerry Gruber
1999-13	1993-94 Schools and Staffing Survey: Data File User's Manual, Volume IV: Bureau of Indian Affairs (BIA) Restricted-Use Codebook	Kerry Gruber
1999-14	1994-95 Teacher Followup Survey: Data File User's Manual, Restricted-Use Codebook	Kerry Gruber
1999-17	Secondary Use of the Schools and Staffing Survey Data	Susan Wiley
2000-04	Selected Papers on Education Surveys: Papers Presented at the 1998 and 1999 ASA and 1999 AAPOR Meetings	Dan Kasprzyk

Listing of NCES Working Papers by Subject

No.	Title	NCES contact
Adult education		
96-14	The 1995 National Household Education Survey: Reinterview Results for the Adult Education Component	Steven Kaufman
96-20	1991 National Household Education Survey (NHES:91) Questionnaires: Screener, Early Childhood Education, and Adult Education	Kathryn Chandler
96-22	1995 National Household Education Survey (NHES:95) Questionnaires: Screener, Early Childhood Program Participation, and Adult Education	Kathryn Chandler
98-03	Adult Education in the 1990s: A Report on the 1991 National Household Education Survey	Peter Stowe
98-10	Adult Education Participation Decisions and Barriers: Review of Conceptual Frameworks and Empirical Studies	Peter Stowe
1999-11	Data Sources on Lifelong Learning Available from the National Center for Education Statistics	Lisa Hudson
American Indian – education		
1999-13	1993-94 Schools and Staffing Survey: Data File User's Manual, Volume IV: Bureau of Indian Affairs (BIA) Restricted-Use Codebook	Kerry Gruber
Assessment/achievement		
95-12	Rural Education Data User's Guide	Samuel Peng
95-13	Assessing Students with Disabilities and Limited English Proficiency	James Houser
97-29	Can State Assessment Data be Used to Reduce State NAEP Sample Sizes?	Larry Ogle
97-30	ACT's NAEP Redesign Project: Assessment Design is the Key to Useful and Stable Assessment Results	Larry Ogle
97-31	NAEP Reconfigured: An Integrated Redesign of the National Assessment of Educational Progress	Larry Ogle
97-32	Innovative Solutions to Intractable Large Scale Assessment (Problem 2: Background Questions)	Larry Ogle
97-37	Optimal Rating Procedures and Methodology for NAEP Open-ended Items	Larry Ogle
97-44	Development of a SASS 1993-94 School-Level Student Achievement Subfile: Using State Assessments and State NAEP, Feasibility Study	Michael Ross
98-09	High School Curriculum Structure: Effects on Coursetaking and Achievement in Mathematics for High School Graduates—An Examination of Data from the National Education Longitudinal Study of 1988	Jeffrey Owings
Beginning students in postsecondary education		
98-11	Beginning Postsecondary Students Longitudinal Study First Follow-up (BPS:96-98) Field Test Report	Aurora D'Amico
Civic participation		
97-25	1996 National Household Education Survey (NHES:96) Questionnaires: Screener/Household and Library, Parent and Family Involvement in Education and Civic Involvement, Youth Civic Involvement, and Adult Civic Involvement	Kathryn Chandler
Climate of schools		
95-14	Empirical Evaluation of Social, Psychological, & Educational Construct Variables Used in NCES Surveys	Samuel Peng
Cost of education indices		
94-05	Cost-of-Education Differentials Across the States	William J. Fowler, Jr.
Course-taking		
95-12	Rural Education Data User's Guide	Samuel Peng
98-09	High School Curriculum Structure: Effects on Coursetaking and Achievement in Mathematics for High School Graduates—An Examination of Data from the National Education Longitudinal Study of 1988	Jeffrey Owings

No.	Title	NCES contact
1999-05	Procedures Guide for Transcript Studies	Dawn Nelson
1999-06	1998 Revision of the Secondary School Taxonomy	Dawn Nelson
Crime		
97-09	Status of Data on Crime and Violence in Schools: Final Report	Lee Hoffman
Curriculum		
95-11	Measuring Instruction, Curriculum Content, and Instructional Resources: The Status of Recent Work	Sharon Bobbitt & John Ralph
98-09	High School Curriculum Structure: Effects on Coursetaking and Achievement in Mathematics for High School Graduates—An Examination of Data from the National Education Longitudinal Study of 1988	Jeffrey Owings
Customer service		
1999-10	What Users Say About Schools and Staffing Survey Publications	Dan Kasprzyk
2000-02	Coordinating NCES Surveys: Options, Issues, Challenges, and Next Steps	Valena Plisko
2000-04	Selected Papers on Education Surveys: Papers Presented at the 1998 and 1999 ASA and 1999 AAPOR Meetings	Dan Kasprzyk
Data quality		
97-13	Improving Data Quality in NCES: Database-to-Report Process	Susan Ahmed
Data warehouse		
2000-04	Selected Papers on Education Surveys: Papers Presented at the 1998 and 1999 ASA and 1999 AAPOR Meetings	Dan Kasprzyk
Design effects		
2000-03	Strengths and Limitations of Using SUDAAN, Stata, and WesVarPC for Computing Variances from NCES Data Sets	Ralph Lee
Dropout rates, high school		
95-07	National Education Longitudinal Study of 1988: Conducting Trend Analyses HS&B and NELS:88 Sophomore Cohort Dropouts	Jeffrey Owings
Early childhood education		
96-20	1991 National Household Education Survey (NHES:91) Questionnaires: Screener, Early Childhood Education, and Adult Education	Kathryn Chandler
96-22	1995 National Household Education Survey (NHES:95) Questionnaires: Screener, Early Childhood Program Participation, and Adult Education	Kathryn Chandler
97-24	Formulating a Design for the ECLS: A Review of Longitudinal Studies	Jerry West
97-36	Measuring the Quality of Program Environments in Head Start and Other Early Childhood Programs: A Review and Recommendations for Future Research	Jerry West
1999-01	A Birth Cohort Study: Conceptual and Design Considerations and Rationale	Jerry West
Educational attainment		
98-11	Beginning Postsecondary Students Longitudinal Study First Follow-up (BPS:96-98) Field Test Report	Aurora D'Amico
Educational research		
2000-02	Coordinating NCES Surveys: Options, Issues, Challenges, and Next Steps	Valena Plisko
Employment		
96-03	National Education Longitudinal Study of 1988 (NELS:88) Research Framework and Issues	Jeffrey Owings
98-11	Beginning Postsecondary Students Longitudinal Study First Follow-up (BPS:96-98) Field Test Report	Aurora D'Amico
Faculty – higher education		
97-26	Strategies for Improving Accuracy of Postsecondary Faculty Lists	Linda Zimbler

No.	Title	NCES contact
2000-01	1999 National Study of Postsecondary Faculty (NSOPF:99) Field Test Report	Linda Zimbler
Finance – elementary and secondary schools		
94-05	Cost-of-Education Differentials Across the States	William J. Fowler, Jr.
96-19	Assessment and Analysis of School-Level Expenditures	William J. Fowler, Jr.
98-01	Collection of Public School Expenditure Data: Development of a Questionnaire	Stephen Broughman
1999-07	Collection of Resource and Expenditure Data on the Schools and Staffing Survey	Stephen Broughman
1999-16	Measuring Resources in Education: From Accounting to the Resource Cost Model Approach	William J. Fowler, Jr.
Finance – postsecondary		
97-27	Pilot Test of IPEDS Finance Survey	Peter Stowe
Finance – private schools		
95-17	Estimates of Expenditures for Private K-12 Schools	Stephen Broughman
96-16	Strategies for Collecting Finance Data from Private Schools	Stephen Broughman
97-07	The Determinants of Per-Pupil Expenditures in Private Elementary and Secondary Schools: An Exploratory Analysis	Stephen Broughman
97-22	Collection of Private School Finance Data: Development of a Questionnaire	Stephen Broughman
1999-07	Collection of Resource and Expenditure Data on the Schools and Staffing Survey	Stephen Broughman
Geography		
98-04	Geographic Variations in Public Schools' Costs	William J. Fowler, Jr.
Imputation		
2000-04	Selected Papers on Education Surveys: Papers Presented at the 1998 and 1999 ASA and 1999 AAPOR Meetings	Dan Kasprzyk
Inflation		
97-43	Measuring Inflation in Public School Costs	William J. Fowler, Jr.
Institution data		
2000-01	1999 National Study of Postsecondary Faculty (NSOPF:99) Field Test Report	Linda Zimbler
Instructional resources and practices		
95-11	Measuring Instruction, Curriculum Content, and Instructional Resources: The Status of Recent Work	Sharon Bobbitt & John Ralph
1999-08	Measuring Classroom Instructional Processes: Using Survey and Case Study Field Test Results to Improve Item Construction	Dan Kasprzyk
International comparisons		
97-11	International Comparisons of Inservice Professional Development	Dan Kasprzyk
97-16	International Education Expenditure Comparability Study: Final Report, Volume I	Shelley Burns
97-17	International Education Expenditure Comparability Study: Final Report, Volume II, Quantitative Analysis of Expenditure Comparability	Shelley Burns
Libraries		
94-07	Data Comparability and Public Policy: New Interest in Public Library Data Papers Presented at Meetings of the American Statistical Association	Carrol Kindel
97-25	1996 National Household Education Survey (NHES:96) Questionnaires: Screener/Household and Library, Parent and Family Involvement in Education and Civic Involvement, Youth Civic Involvement, and Adult Civic Involvement	Kathryn Chandler
Limited English Proficiency		
95-13	Assessing Students with Disabilities and Limited English Proficiency	James Houser
Literacy of adults		
98-17	Developing the National Assessment of Adult Literacy: Recommendations from Stakeholders	Sheida White

No.	Title	NCES contact
1999-09a	1992 National Adult Literacy Survey: An Overview	Alex Sedlacek
1999-09b	1992 National Adult Literacy Survey: Sample Design	Alex Sedlacek
1999-09c	1992 National Adult Literacy Survey: Weighting and Population Estimates	Alex Sedlacek
1999-09d	1992 National Adult Literacy Survey: Development of the Survey Instruments	Alex Sedlacek
1999-09e	1992 National Adult Literacy Survey: Scaling and Proficiency Estimates	Alex Sedlacek
1999-09f	1992 National Adult Literacy Survey: Interpreting the Adult Literacy Scales and Literacy Levels	Alex Sedlacek
1999-09g	1992 National Adult Literacy Survey: Literacy Levels and the Response Probability Convention	Alex Sedlacek
1999-11	Data Sources on Lifelong Learning Available from the National Center for Education Statistics	Lisa Hudson
Literacy of adults – international		
97-33	Adult Literacy: An International Perspective	Marilyn Binkley
Mathematics		
98-09	High School Curriculum Structure: Effects on Coursetaking and Achievement in Mathematics for High School Graduates—An Examination of Data from the National Education Longitudinal Study of 1988	Jeffrey Owings
1999-08	Measuring Classroom Instructional Processes: Using Survey and Case Study Field Test Results to Improve Item Construction	Dan Kasprzyk
Parental involvement in education		
96-03	National Education Longitudinal Study of 1988 (NELS:88) Research Framework and Issues	Jeffrey Owings
97-25	1996 National Household Education Survey (NHES:96) Questionnaires: Screener/Household and Library, Parent and Family Involvement in Education and Civic Involvement, Youth Civic Involvement, and Adult Civic Involvement	Kathryn Chandler
1999-01	A Birth Cohort Study: Conceptual and Design Considerations and Rationale	Jerry West
Participation rates		
98-10	Adult Education Participation Decisions and Barriers: Review of Conceptual Frameworks and Empirical Studies	Peter Stowe
Postsecondary education		
1999-11	Data Sources on Lifelong Learning Available from the National Center for Education Statistics	Lisa Hudson
Postsecondary education – persistence and attainment		
98-11	Beginning Postsecondary Students Longitudinal Study First Follow-up (BPS:96-98) Field Test Report	Aurora D'Amico
1999-15	Projected Postsecondary Outcomes of 1992 High School Graduates	Aurora D'Amico
Postsecondary education – staff		
97-26	Strategies for Improving Accuracy of Postsecondary Faculty Lists	Linda Zimbler
2000-01	1999 National Study of Postsecondary Faculty (NSOPF:99) Field Test Report	Linda Zimbler
Private schools		
96-16	Strategies for Collecting Finance Data from Private Schools	Stephen Broughman
97-07	The Determinants of Per-Pupil Expenditures in Private Elementary and Secondary Schools: An Exploratory Analysis	Stephen Broughman
97-22	Collection of Private School Finance Data: Development of a Questionnaire	Stephen Broughman
Projections of education statistics		
1999-15	Projected Postsecondary Outcomes of 1992 High School Graduates	Aurora D'Amico
Public school finance		
1999-16	Measuring Resources in Education: From Accounting to the Resource Cost Model Approach	William J. Fowler, Jr.

No.	Title	NCES contact
Public schools		
97-43	Measuring Inflation in Public School Costs	William J. Fowler, Jr.
98-01	Collection of Public School Expenditure Data: Development of a Questionnaire	Stephen Broughman
98-04	Geographic Variations in Public Schools' Costs	William J. Fowler, Jr.
1999-02	Tracking Secondary Use of the Schools and Staffing Survey Data: Preliminary Results	Dan Kasprzyk
Public schools – secondary		
98-09	High School Curriculum Structure: Effects on Coursetaking and Achievement in Mathematics for High School Graduates—An Examination of Data from the National Education Longitudinal Study of 1988	Jeffrey Owings
Reform, educational		
96-03	National Education Longitudinal Study of 1988 (NELS:88) Research Framework and Issues	Jeffrey Owings
Response rates		
98-02	Response Variance in the 1993-94 Schools and Staffing Survey: A Reinterview Report	Steven Kaufman
School districts, public		
98-07	Decennial Census School District Project Planning Report	Tai Phan
1999-03	Evaluation of the 1996-97 Nonfiscal Common Core of Data Surveys Data Collection, Processing, and Editing Cycle	Beth Young
School districts, public – demographics of		
96-04	Census Mapping Project/School District Data Book	Tai Phan
Schools		
97-42	Improving the Measurement of Staffing Resources at the School Level: The Development of Recommendations for NCES for the Schools and Staffing Survey (SASS)	Mary Rollefson
98-08	The Redesign of the Schools and Staffing Survey for 1999-2000: A Position Paper	Dan Kasprzyk
1999-03	Evaluation of the 1996-97 Nonfiscal Common Core of Data Surveys Data Collection, Processing, and Editing Cycle	Beth Young
Schools – safety and discipline		
97-09	Status of Data on Crime and Violence in Schools: Final Report	Lee Hoffman
Software evaluation		
2000-03	Strengths and Limitations of Using SUDAAN, Stata, and WesVarPC for Computing Variances from NCES Data Sets	Ralph Lee
Staff		
97-42	Improving the Measurement of Staffing Resources at the School Level: The Development of Recommendations for NCES for the Schools and Staffing Survey (SASS)	Mary Rollefson
98-08	The Redesign of the Schools and Staffing Survey for 1999-2000: A Position Paper	Dan Kasprzyk
Staff – higher education institutions		
97-26	Strategies for Improving Accuracy of Postsecondary Faculty Lists	Linda Zimbler
State		
1999-03	Evaluation of the 1996-97 Nonfiscal Common Core of Data Surveys Data Collection, Processing, and Editing Cycle	Beth Young
Statistical methodology		
97-21	Statistics for Policymakers or Everything You Wanted to Know About Statistics But Thought You Could Never Understand	Susan Ahmed
Students with disabilities		
95-13	Assessing Students with Disabilities and Limited English Proficiency	James Houser

No.	Title	NCES contact
Survey methodology		
96-17	National Postsecondary Student Aid Study: 1996 Field Test Methodology Report	Andrew G. Malizio
97-15	Customer Service Survey: Common Core of Data Coordinators	Lee Hoffman
97-35	Design, Data Collection, Interview Administration Time, and Data Editing in the 1996 National Household Education Survey	Kathryn Chandler
98-06	National Education Longitudinal Study of 1988 (NELS:88) Base Year through Second Follow-Up: Final Methodology Report	Ralph Lee
98-11	Beginning Postsecondary Students Longitudinal Study First Follow-up (BPS:96-98) Field Test Report	Aurora D'Amico
98-16	A Feasibility Study of Longitudinal Design for Schools and Staffing Survey	Stephen Broughman
1999-07	Collection of Resource and Expenditure Data on the Schools and Staffing Survey	Stephen Broughman
1999-17	Secondary Use of the Schools and Staffing Survey Data	Susan Wiley
2000-01	1999 National Study of Postsecondary Faculty (NSOPF:99) Field Test Report	Linda Zimbler
2000-02	Coordinating NCES Surveys: Options, Issues, Challenges, and Next Steps	Valena Plisko
2000-04	Selected Papers on Education Surveys: Papers Presented at the 1998 and 1999 ASA and 1999 AAPOR Meetings	Dan Kasprzyk
Teachers		
98-13	Response Variance in the 1994-95 Teacher Follow-up Survey	Steven Kaufman
1999-14	1994-95 Teacher Followup Survey: Data File User's Manual, Restricted-Use Codebook	Kerry Gruber
Teachers – instructional practices of		
98-08	The Redesign of the Schools and Staffing Survey for 1999-2000: A Position Paper	Dan Kasprzyk
Teachers – opinions regarding safety		
98-08	The Redesign of the Schools and Staffing Survey for 1999-2000: A Position Paper	Dan Kasprzyk
Teachers – performance evaluations		
1999-04	Measuring Teacher Qualifications	Dan Kasprzyk
Teachers – qualifications of		
1999-04	Measuring Teacher Qualifications	Dan Kasprzyk
Teachers – salaries of		
94-05	Cost-of-Education Differentials Across the States	William J. Fowler, Jr.
Variance estimation		
2000-03	Strengths and Limitations of Using SUDAAN, Stata, and WesVarPC for Computing Variances from NCES Data Sets	Ralph Lee
2000-04	Selected Papers on Education Surveys: Papers Presented at the 1998 and 1999 ASA and 1999 AAPOR Meetings	Dan Kasprzyk
Violence		
97-09	Status of Data on Crime and Violence in Schools: Final Report	Lee Hoffman
Vocational education		
95-12	Rural Education Data User's Guide	Samuel Peng
1999-05	Procedures Guide for Transcript Studies	Dawn Nelson
1999-06	1998 Revision of the Secondary School Taxonomy	Dawn Nelson