DOCUMENT RESUME

ED 146 189 TM 006 553

TITLE	National Longitudinal Study of the High School Class of 1972. Sample Design Efficiency Study: Effects of Stratification, Clustering, and Unequal Weighting on the Variances of NLS Statistics.
INSTITUTION	National Center for Education Statistics (DHEW), Washington, D.C.
SPONS AGENCY	Office of the Assistant Secretary for Education (DHEW), Washington, D.C.
REPORT NO	NCES-77-258
PUB DATE	77
CONTRACT	OEC-0-73-6666
NOTE	26p.
AVAILABLE FROM	Superintendent of Documents, U.S. Government Printing Office, Washington, D.C. 20402 (Stock Number 017-080-01692-3, $.65); there is a minimum charge of $1.00 for each mail order
EDRS PRICE	MF-$0.83 HC-$2.06 Plus Postage.
DESCRIPTORS	*Analysis of Variance; Cluster Analysis; Data Analysis; *High School Students; *Longitudinal Studies; *National Surveys; Predictor Variables; Research Design; *Sampling; Secondary Education; *Statistical Analysis; Weight
IDENTIFIERS	National Longitudinal Study High School Class 1972; Stratification

ABSTRACT

A complex two-stage sample selection process was used in designing the National Longitudinal Study of the High School Class of 1972. The first-stage sampling frame used in the selection of schools was stratified by the following seven variables: public vs. private control, geographic region, grade 12 enrollment, proximity to institutions of higher education, percentage of minority group enrollment, community income level, and degree of urbanization. Six hundred strata determined the initial sample of 1,200 schools; later, a random sample of 18 seniors per high school was selected. This report considers the effects of stratification, oversampling of schools by percentage of minority group enrollment and community income level, clustering of students within a school and unequal weighting on the variances of the resulting statistics and therefore, on the precision of the sample statistics. Results suggest that school stratification variables reduced the variances of national estimates to twenty percent below what would have been expected with unstratified cluster sampling. Of the five major stratification variables: socioeconomic status, size of school, type of control, geographic region, and proximity to college or university; region is the strongest variable and type of control is the weakest. (Author/HV)

SPONSORED REPORTS SERIES

TP1-

NATIONAL LONGITUDINAL STUDY OF THE HIGH SCHOOL CLASS OF 1972

SAMPLE DESIGN EFFICIENCY STUDY: EFFECTS OF STRATIFICATION, CLUSTERING, AND UNEQUAL WEIGHTING ON THE VARIANCES OF NLS STATISTICS

NCES 77-258

NATIONAL LONGITUDINAL STUDY OF THE HIGH SCHOOL CLASS OF 1972

SAMPLE DESIGN EFFICIENCY STUDY: EFFECTS OF STRATIFICATION, CLUSTERING, AND UNEQUAL WEIGHTING ON THE VARIANCES OF NLS STATISTICS

Project Officer
William B. Fetters
National Center for Education Statistics

U.S. DEPARTMENT OF HEALTH, EDUCATION, AND WELFARE
Joseph A. Califano, Jr., *Secretary*

Education Division
Philip E. Austin, *Acting Assistant Secretary for Education*

National Center for Education Statistics
Marie D. Eldridge, *Administrator*

NATIONAL CENTER FOR EDUCATION STATISTICS

"The purpose of the Center shall be to collect and disseminate statistics and other data related to education in the United States and in other nations. The Center shall . . . collect, collate, and, from time to time, report full and complete statistics on the conditions of education in the United States; conduct and publish reports on specialized analyses of the meaning and significance of such statistics; . . . and review and report on education activities in foreign countries."--Section 406(b) of the General Education Provisions Act, as amended (20 U.S.C. 1221e-1).

Prepared under contract No. OEC-0-73-6666 with the U.S. Department of Health, Education, and Welfare. Contractors undertaking such projects are encouraged to express freely their professional judgment. This report, therefore, does not necessarily represent positions or policies of the Education Division, and no official endorsement should be inferred.

U.S. GOVERNMENT PRINTING OFFICE
WASHINGTON: 1977

For sale by the Superintendent of Documents, U S Government Printing Office
Washington, D C 20402 - Price 65 cents

Stock No. 017-080-01692-3

There is a minimum charge of $1 00 for each mail order

FOREWORD

A complex two-stage sample selection process was used in designing the National Longitudinal Study (NLS) of the High School Class of 1972. The first-stage sampling frame used in the selection of schools was stratified by the following seven variables:

- Type of control (public or private)
- Geographic region (Northeast, North Central, South, and West)
- Grade-12 enrollment (less than 300, 300 to 599, and 600 or more)
- Proximity to institutions of higher learning (3 categories)
- Percent minority group enrollment (8 categories, public schools only)
- Income level of the community (11 categories, public schools; 8 categories, Catholic schools)
- Degree of urbanization (10 categories)

Both priority considerations and judgment were used in consolidating the various classes to produce the 600 final strata from which a sample of 1,200 schools was chosen. The second stage of the sample selection involved choosing a simple random sample of 18 seniors per high school. This report considers the effects of stratification, oversampling of schools by percent minority group enrollment and income level of the community, clustering of students within a school, and unequal weighting on the variances of the resulting statistics and hence the precision of the sample statistics.

The results suggest that the school stratification variables reduced the variances of national estimates by 20 percent below what would have been expected with unstratified cluster sampling. Variances of subpopulation were reduced by lesser amounts, from 6 to 20 percent, depending upon the subpopulation. Clustering the sample of students increased variances of national estimates by an estimated 83.5 percent over simple random sampling with smaller increases for various subgroups. In general, the increase in variance due to cluster sampling is only partly offset by the reduction due to stratification.

Of the five major stratification variables, SES (socioeconomic status), size of school, type of control, geographic region, and proximity to college or university, region is perhaps the strongest; type of control is the weakest; and the other three lie somewhere between.

The final section of the report describes a limited and approximate analysis to secure rough indications of the effects of unequal weightings due to oversampling, nonresponse adjustments, unequal stratum sizes, and imprecise school size measures.

This study was conducted by R.P. Moore and B.V. Shah, of Research Triangle Institute, under contract with the U.S. Department of Health, Education, and Welfare for the National Center for Education Statistics.

Francis V. Corrigan, *Acting Director*
Division of Multilevel Education Statistics

Elmer F. Collins, *Chief*
Statistical Analysis Branch

CONTENTS

LIST OF TABLES

I. INTRODUCTION

The efficiency of the 1972 National Longitudinal Study (NLS) sample design for a base-year survey was analyzed previously using variance component estimates and estimated efficiencies [1]. In this report, average design effects for statistics estimated from the base-year data are presented. Attempts to partition the design effect into effects due to stratification, clustering, and unequal weighting are discussed. The expected increase in subpopulation sample sizes due to oversampling is calculated and compared with the actual increases observed in the base-year survey. The effects on variances of oversampling and other factors which lead to unequal weighting are approximated and the optimum oversampling rates for several subpopulations are estimated. Several of the stratification variables are ranked from most effective to least effective in reducing the variances of survey estimates.

NOTE --References indicated in brackets are listed on page 22.

II. PARTITIONING THE DESIGN EFFECT

A. Estimated Design Effects

The design effect [2] or "Deff", defined as the ratio of the actual variance of a survey estimate to the variance for a simple random sample of the same size, is useful in evaluating a sample design. The Deff measures the combined effects of clustering, stratification, and unequal weighting on the variances of survey estimates.

Variance component estimates computed for 357 statistics in the study of NLS design efficiency were used to calculate estimated design effects. For each statistic, the components estimated were:

σ_0^2 = variation among final strata,

σ_1^2 = variation among schools within final strata, and

σ_2^2 = variation among students within schools.

The variance component estimates were used to model the variance of each statistic with the NLS design,

$$\Sigma_1^2 = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_1 n_2} \quad (1)$$

where

n_1 = the number of sample schools, and

n_2 = the number of sample students per school.

The approximate variance of each statistic for a simple random sample of $n_1 n_2$ students was calculated as

$$\Sigma_3^2 = \frac{\sigma_0^2 + \sigma_1^2 + \sigma_2^2}{n_1 n_2} \quad (2)$$

Then the design effect, D, for each statistic was estimated as

$$D = \frac{\Sigma_1^2}{\Sigma_3^2} \quad (3)$$

using n_1 = 1,043 and n_2 = 17, the approximate numbers of responding sample schools and students per school in the NLS base-year survey.

Table 1 shows the average values of the design effects and root design effects calculated, by type of statistic. We note that the estimated design effects tend to be largest for national means and tend to vary with the average cluster size (number of respondents per school) for subgroup means. The design effects for subgroup, or domain, means tend to be larger than those for the differences between subgroup and national means.

Table 1.—Average number of respondents and average estimated design effects

Type of statistic	Number of respondents per school	Number of statistics	Design[1] effect D	Square root of design effect $\sqrt{D}$
National means	15.363	21	1.463	1.203
Subgroup means				
White	11.711	42	1.327	1.147
Females	7.690	42	1.213	1.097
Males	7.552	42	1.173	1.081
Father high school graduate	6.399	42	1.156	1.074
Father less than high school	4.651	42	1.117	1.056
Father college graduate	2.440	42	1.119	1.057
Black	1.888	42	1.219	1.100
Other races	1.465	42	1.182	1.085
All domain means	5.475	168	1.233	1.107
Differences of domain and national means	5.475	168	1.143	1.067
All statistics	6.056	357	1.204	1.094

[1] Assumes $n_2 = 17$

The root design effects computed using variance component estimates (table 1) are 10 to 15 percent higher than comparable ones tabulated by William B. Letters [3] using the conventional between-PSU-within-stratum variance summed over strata. This is not surprising recalling that the variance component estimates are thought to be overestimates [1] and realizing that equation 3 may be rewritten as

$$D = \frac{n_2 \sigma_1^2}{\sigma_0^2 + \sigma_1^2 + \sigma_2^2} + \frac{\sigma_2^2}{\sigma_0^2 + \sigma_1^2 + \sigma_2^2} \quad (4)$$

From the above, we see that if σ_1^2 and/or σ_2^2 were overestimated to a greater extent than the remaining components, then D would be overestimated.

B. Effects of Stratification and Clustering

We can also use the variance component estimates to approximate the effect of clustering the sample of students by school and the effect of stratifying schools. The effects on the variances of survey estimates are of interest in studying the efficiency of the sample design. Recalling equation 3, $D = \Sigma_1^2/\Sigma_3^2$, the estimated design effect may be rewritten as

$$D = 1 + (n_2 - 1)\,\delta_{c/w} - n_2\,\delta_{rs/w} \quad (5)$$

or

$$D = C_{rw} - n_2\,\delta_{rs/w} \quad (6)$$

where

$$\delta_{c/w} = \frac{\sigma_0^2 + \sigma_1^2}{\sigma_0^2 + \sigma_1^2 + \sigma_2^2}, \quad (7)$$

and

$$\delta_{rs/w} = \frac{\sigma_0^2}{\sigma_0^2 + \sigma_1^2 + \sigma_2^2} \quad (8)$$

The first term of equations 5 and 6 represents the effect of clustering the sample of students by school attended, where $\delta_{c/w}$ is the intraschool cluster correlation for an unstratified selection of schools and students, given the unequal weighting of the NLS base year sample design. The last term in equations 5 and 6 represents the reduction in the variances of survey estimates obtained from school stratification, where $\delta_{rs/w}$ is the intrastratum cluster correlation for a random selection of students from an unstratified frame.

If we introduce Σ_2^2, the variance of a survey estimate for an unstratified cluster sample,

$$\Sigma_2^2 = \frac{\sigma_0^2 + \sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_1 n_2}, \quad (9)$$

then we can write

$$C_{rw} = \frac{\Sigma_2^2}{\Sigma_3^2} = 1 + (n_2 - 1)\,\delta_{c/w}$$

as before. Using equation 9, we can also write the effect of stratification, S_{cw}, in a multiplicative model as

$$S_{cw} = \frac{\Sigma_1^2}{\Sigma_2^2}$$

$$= \frac{C_{rw} - n_2 \delta_{rs/w}}{C_{rw}} \quad (10)$$

Now we can write

$$D = \frac{\Sigma_2^2}{\Sigma_3^2} \cdot \frac{\Sigma_1^2}{\Sigma_2^2} = C_{rw} \; S_{cw} \quad (11)$$

and the design effect has been partitioned into C_{rw}, the effect of clustering, and S_{cw}, the effect of stratification.

Table 2 shows the average values of design factors calculated for the NLS sample design using the variance component estimates described earlier. (The S_{cw} values shown were derived from the average C_{rw} and D values.) Clustering the sample of students increased variances of national estimates by an estimated 83.5 percent over simple random sampling. Stratifying the clusters using the NLS school stratification scheme reduced the variances of national estimates by an estimated 20.3 percent or $100(1 - S_{cw})$, on the average, below what they would have been with unstratified cluster sampling. Both effects are reduced for subgroups and there appears to be a tendency for both effects to approach 1.0 as the subpopulation size gets small. In general, the increase in variance due to cluster sampling is only partly offset by the reduction due to stratification. Table 3 shows average values of the ratios of variance components $\delta_{rs/w}$ and $\delta_{c/w}$ used in the modeling.

Having estimated the reduction in variance from the stratification variables used in the NLS design, one also would like to compare the effectiveness of the individual stratification variables. Knowledge of which variables were most effective in reducing the variance of survey estimates would be useful in designing future NLS samples and also in the design of similar samples. The results of comparing the individual stratification variables are shown in section III of this report.

Design effects (and variances of estimates) are also affected by the unequal weighting of the individual elements of the sample. The effect of unequal weighting is discussed in the next section.

Table 2.—Average effects of clustering and stratification for the NLS design[1]

Statistic	C_{rw}	$n_2\delta_{rs/w}$	D	S_{cw}
National means	1.835	0.372	1.463	.797
Subgroup means				
White	1.655	.328	1.327	.802
Females	1.364	.151	1.213	.889
Males	1.356	.183	1.173	.865
Father high school graduate	1.302	.146	1.156	.888
Father less than high school	1.274	.157	1.117	.877
Father college graduate	1.188	.069	1.119	.942
Black	1.458	.239	1.219	.836
Other races	1.311	.129	1.182	.902
All domain means	1.480	.197	1.233	.862
Differences of domain and national means	1.296	.153	1.142	.882
All statistics	1.391	.187	1.204	.857

[1] Assumes $n_2 = 17$.

Table 3.—Average ratios of variance component estimates

Domain	$\delta_{rs/w}$	$\delta_{c/w}$	Number of statistics
National means	0.022	0.052	21
Subgroup means			
White	.019	.041	42
Females	.009	.023	42
Males	.011	.022	42
Father high school graduate	.009	.019	42
Father less than high school	.009	.017	42
Father college graduate	.004	.012	42
Black	.014	.029	42
Other races	.008	.019	42
All domain means	.012	.027	168
Differences of domain and national means	.009	.019	168
All statistics	.011	.024	357

C. Effect of Unequal Weighting

The variances of survey estimates are increased when the sample elements (students) have unequal weights. Unequal weights arise from oversampling certain subpopulations, from using imprecise size measures to select sample schools, and from nonresponse adjustments. The estimated design effects presented in the previous section include the effect of unequal weighting, as do the estimated effects of stratification and/or clustering. That is, in the previous section the design effect was partitioned into

$$D = C_{rw} \; S_{cw},$$

whereas it would be possible to partition the design effect into

$$D = WSC.$$

Folsom [4] discusses the methodology which could be used to estimate the effect of unequal weighting and other finer partitionings of the design effect (see equation 62 in reference 4). Unfortunately, completing the analysis described by Folsom was beyond the scope of the project as it would have required the development of several new computer programs, estimation of an additional set of variance components, and additional analysis time.

In order to obtain some information about the effects of unequal weighting in the NLS design, a more limited and approximate analysis was conducted. The analysis involved estimating the approximate effect of unequal weighting on the variances of survey estimates. A portion of the unequal weighting is due to oversampling a part of the population and the effect of this oversampling is estimated. The remainder of the unequal weighting, aside from oversampling, is caused by nonresponse adjustments, unequal stratum sizes, and imprecise school size measures. Estimates of the combined effect of these factors were also computed. The reader should be cautioned that the analyses presented here are based on oversimplifications and far-reaching assumptions, and the results should be regarded as rough indications of the effects rather than precise estimates.

1. Extent of Oversampling

The school sampling frame for the 1972 NLS was divided into two socioeconomic (SES) strata. The low SES stratum (type A schools) was formed by grouping schools with high percentages of minority students and/or schools located in low income areas. The high SES stratum (type B schools) consisted of all other schools in the sampling frame. Students from the low SES stratum were sampled at approximately twice the sampling rate used in the high SES stratum, in order to increase the number of sample students who belonged to critical subpopulations—the minorities, the poor, and the poorly educated. (Additional details are given in the Westat report [5] on the sample design.)

Data needed to complete this analysis included sample counts and estimated subpopulation sizes for the low SES and high SES strata separately. These data are shown in tables 4 and 5 for subpopulations defined by sex, race, and father's education. Also shown, for general interest, are "adjusted" estimates where the "not reported" estimates and sample sizes were proportionally added to the remaining categories for each subpopulation-defining variable. The estimated totals for the low and high SES strata are close to the estimated numbers of seniors (983,240 and 2,064,647) used in designing the sample [5], considering that the latter were estimates based on enrollments in earlier school years and that some of the schools in the sampling frame had closed by the time the survey was conducted.

The "not reported" categories for father's education include both students who answered the question as "not applicable" and those who left the question blank. The estimated subpopulation size estimates indicate, as might be expected, that students

Table 4.—Estimated subpopulation sizes for low and high SES adjusted for missing subpopulation classifier variables

Subpopulation	Low SES (type A)		High SES (type B)		Total	
	Number	Percent	Number	Percent	Number	Percent
Unadjusted estimates						
Sex						
Male	426,902	45 5	927,144	46 0	1,354,046	45 8
Female	438,887	46 8	921,729	45 7	1,360,616	46 1
Not reported	72,509	7 7	166,257	8 3	238,767	8 1
Race						
White	537,321	57 3	1,655,621	82 2	2,192,942	74 3
Black	197,227	21 0	55,398	2 7	252,624	8.6
Other	118,103	12 6	115,763	5 7	233,866	7 9
Not reported	85,648	9 1	188,348	9 3	273,997	9 3
Father's education						
Less than high school graduate	281,679	30 0	426,640	21 2	708,319	24 0
High school graduate	212,265	22 6	529,020	26 3	741 286	25 1
College graudate	197,064	21 0	705,395	35 0	902,459	30 6
Not reported	247,291	26 4	254,074	17 6	601,365	20 4
Total	**938,299**	**100.0**	**2,015,130**	**100.0**	**2,953,429**	**100.0**
Adjusted estimates [1]						
Sex						
Male	462,655	49·3	1,010,516	50 1	1,473,171	49 9
Female	475,543	50 7	1,004,614	49 9	1,480,257	50 1
Race						
White	591,294	63 0	1,826,322	90.6	2,417,616	81 9
Black	217,038	23 1	61,110	3 0	278,148	9 4
Other	129,966	13 9	127,699	6 3	257,665	8 7
Father's education						
Less than high school graduate	382,483	40 7	517,584	25 7	900,067	30 5
High school graduate	288,228	30 7	641,787	31 8	930,015	31 5
College graduate	267,587	28 5	855,759	42.5	1,123,346	38.0
Total	**938,299**	**100.0**	**2,015,130**	**100.0**	**2,953,429**	**100.0**

[1] Adjusted estimates computed by proportionately allocating "not reported" estimate to other categories.

Table 5.—Subpopulation sample sizes for low and high SES adjusted for missing subpopulation classifier variables

Subpopulation	Low SES (type A)		High SES (type B)		Total	
	Number	Percent	Number	Percent	Number	Percent
Unadjusted counts:						
Sex:						
Male	3,788	45.2	4,289	45.9	8,075	45.6
Female	3,946	47.1	4,256	45.5	8,202	46.3
Not reported	640	7.6	809	8.6	1,449	8.2
Race:						
White	4,775	57.0	7,652	81.8	12,427	70.1
Black	1,807	21.6	252	2.7	2,059	11.6
Other	1,036	12.4	542	5.8	1,578	8.9
Not reported	754	9.0	908	9.7	1,662	9.4
Father's education:						
Less than high school graduate	2,492	29.8	1,953	20.9	4,445	25.1
High school graduate	1,883	22.5	2,420	25.9	4,303	24.3
College graduate	1,770	21.1	3,306	35.3	5,076	28.6
Not reported	2,227	26.6	1,675	17.9	3,902	22.0
Total	**8,372**	**100.0**	**9,354**	**100.0**	**17,726**	**100.0**
Adjusted estimates:[1]						
Sex:						
Male	4,099	49.0	4,696	50.2	8,795	49.6
Female	4,273	51.0	4,659	49.8	8,932	50.4
Race:						
White	5,248	62.7	8,475	90.6	13,723	77.4
Black	1,986	23.7	279	3.0	2,265	12.8
Other	1,139	13.6	600	6.4	1,739	9.8
Father's education:						
Less than high school graduate	3,395	40.6	2,379	25.4	5,774	32.6
High school graduate	2,565	30.6	2,948	31.5	5,513	31.1
College graduate	2,411	28.8	4,027	43.1	6,438	36.3
Total	**8,372**	**100.0**	**9,354**	**100.0**	**17,726**	**100.0**

[1]Adjusted estimates computed by proportionally allocating "not reported" sample size to other categories.

of minority races and students with poorly educated fathers make up larger percentages of the low SES stratum than they do of the high SES stratum.

In table 5, it may be noted that the overall participation rate was 77.5 percent in the low SES stratum and 86.6 percent in the high SES since the target sample size was 10,800 students in each. The percentages of sample students who were black, other races, with poorly educated fathers, and with father's education unknown would have been higher if both SES groups had participated at the same rate.

The amount of oversampling achieved for various subpopulations in the 1972 NLS base-year survey has been estimated by Fetters [6]. What is perhaps less well-known is the amount of oversampling that should have been expected, given the sample design and the distribution of the target populations within the oversampled and undersampled portions of the universe. Prior to using the data from tables 4 and 5 to estimate this, we will introduce the following notation which is essentially that used in the recent article by Waksberg [7].

Let N_1 and N_2 be the total populations of stratum 1 and stratum 2, respectively, where $N_2 = v\ N_1$ and $v \geq 1$.

Let t_1 and t_2 be the proportions of stratum 1 and stratum 2 belonging to a specified subgroup.

Let r_1 and r_2 be the sampling rates used in stratum 1 and stratum 2, respectively, where $r_1 = k\ r_2$ and $K \geq 1$.

Now we can write the expected increase in subpopulation sample sizes, due to oversampling, as the ratio

$$\frac{r_1\ t_1\ N_1 + r_2\ t_2\ N_2}{r\ (t_1\ N_1 + t_2\ N_2)} = \frac{r_1}{r}\ \frac{t_1\ N_1}{t_1\ N_1 + t_2\ N_2} + \frac{r_2}{r}\ \frac{t_2\ N_2}{t_1\ N_1 + t_2\ N_2} \qquad (12)$$

where r = the uniform sampling rate for a proportional allocation which will give the same expected sample size; that is, $r(N_1 + N_2) = r_1\ N_1 + r_2\ N_2$. The numerator of equation 12 is the subpopulation sample size expected with oversampling and the denominator is the subpopulation sample size expected with no oversampling. The estimates in table 4 were used to calculate the first two columns of table 6, which are estimated values of

$$\frac{t_1\ N_1}{t_1\ N_1 + t_2\ N_2}$$

and

$$\frac{t_2\ N_2}{t_1\ N_1 + t_2\ N_2}$$

The sampling rates for the 1972 NLS were calculated from data in the Westat report [5] as

$r_1 = 10{,}800/983{,}240 = .010984,$

$r_2 = 10{,}800/2{,}064{,}647 = .005231,$ and

$r = 21{,}600/3{,}047{,}887 = .007087.$

Thus

$$\frac{r_1}{r} = 1.550,$$

$$\frac{r_2}{r} = .738,$$ and

$$k = \frac{r_1}{r_2} = 2.100$$

Using these figures in equation 12, the expected increase in sample sizes for various subpopulations were computed and are shown in the third column of table 6. For comparison, the actual oversampling achieved in the survey, calculated as the percent of sample cases belong to the subpopulation (table 5) divided by the estimated percent of the population (table 4) belonging to the subpopulation, is shown in the last column of table 6. The actual oversampling achieved agrees quite closely with that which would be expected, given the design. Note that to obtain much increase in the subpopulation sample size from oversampling, there must be a *large* proportion of the population in the stratum which is sampled at the higher-than-proportional rate.

Table 6.—Expected and actual effect of oversampling on subpopulation sample sizes, 1972 NLS base-year survey

Subpopulation	Estimated proportion of subpopulation members in		Effect of oversampling on sample sizes	
	Low SES stratum	High SES stratum	Expected	Actual
Sex:				
Male	0.315	0.685	0.99	1.00
Female	.323	.677	1.00	1.00
Not reported	.304	.696	.98	1.01
Race:				
White	.245	.755	.94	.94
Black	.781	.219	1.37	1.35
Other	.505	.495	1.15	1.13
Not reported	.313	.687	.99	1.01
Father's education:				
Less than high school graduate	.398	.602	1.06	1.05
High school graduate	.286	.714	.97	.97
College graduate	.218	.782	.92	.93
Not reported	.411	.589	1.07	1.08

2. Effect of Oversampling

Waksberg [7] gives a convenient formula for computing the approximate increase or decrease in variances of subpopulation means as

$$\frac{\sigma_B^2}{\sigma_A^2} \doteq \frac{(k + v)\ (u + kv)}{k\ (1 + v)\ (u + v)} \quad (13)$$

where

σ_B^2 = the variance of an estimated subpopulation mean with oversampling.

σ_A^2 = the variance of an estimated subpopulation mean with proportional sampling.

$k = r_1/r_2$,

$v = N_2/N_1$, and

$u = t_1/t_2$.

(All of the above symbols except σ_B^2, σ_A^2 were defined in the previous section.) Equation 13 assumes simple random sampling of subpopulation members within each of the two strata, a common variance within strata, and a very small sampling rate within strata. The first of these assumptions is considerably different from the NLS design, which points out again that using equation 13 permits only rough approximations to the effect of oversampling on the variances of survey estimates.

The approximate effect of oversampling on the variances of survey estimates was calculated using equation 13 with $k = 2.100$, $v = 2.148$, and the values of u for each subpopulation obtained from the t_1, t_2 estimates in table 4. Table 7 shows the estimated effect of oversampling for each subpopulation. The variances were increased by oversampling in the NLS design for most subpopulations and a moderate reduction was obtained only for blacks. Variances of estimates for the total population of students were increased by 13 percent. Proportional sampling is optimal for total population estimates. The increase in variance of estimates for the total population may also be written as

$$1 + \frac{(k - 1)^2}{k}\left(\frac{N_1}{N}\right)\left(\frac{N_2}{N}\right) \text{ for } k \geq 1. \quad (14)$$

Waksberg also shows that, with the assumptions stated earlier, the optimum rate of oversampling for estimated subpopulation means is

$$\text{opt } k = \sqrt{u}. \quad (15)$$

Table 7 shows the approximate optimum k for each subpopulation. The NLS design, with $k = 2.100$, employed more than the optimum oversampling rate for all subpopulations shown here except blacks, where a higher degree of oversampling would have been optimal. For a number of subpopulations with $u < 1$, proportional sampling was indicated.

The effect of oversampling on the variances, as estimated here, is only a part of the effect of unequal weighting. The assumption of simple random sampling within the two strata implies equal weighting within strata, whereas the NLS sample had unequal weights. The increase in variance due to unequal weighting from factors other than oversampling is discussed in the next section.

Table 7.—Estimated effect of oversampling on the variances of survey estimates and optimum oversampling rates for subpopulations

Subpopulation	$u = \frac{t_1}{t_2}$	σ_B^2/σ_A^2	Optimum k
Sex:			
Male	0.989	1.13	0.99
Female	1.024	1.12	1.01
Not reported	.928	1.14	.96
Race:			
White	.697	1.18	.83
Black	7.778	.80	2.79
Other	2.211	.99	1.49
Not reported	.978	1.13	.99
Father's education:			
Less than high school graduate	1.415	1.07	1.19
High school graduate	.859	1.15	.93
College graduate	.600	1.20	.77
Not reported	1.500	1.06	1.22
Total	1.000	1.13	1.00

3. Effect of Unequal Weighting within the Low SES and High SES Strata

The effect of unequal weighting within the low SES and high SES strata can be approximated by considering the estimated total, 'X', written as

$$X' = \sum_{h=1}^{300} \sum_{i=1}^{n_h} \sum_{j=1}^{n_{hi}} W_{hij} X_{hij} + \sum_{h=301}^{600} \sum_{i=1}^{n_h} \sum_{j=1}^{n_{hi}} W_{hij} X_{hij} , \quad (16)$$

and its variance

$$Var(X') = \sum_{h=1}^{300} \sum_{i=1}^{n_h} \sum_{j=1}^{n_{hi}} W_{hij}^2 \sigma_h^2 + \sum_{h=301}^{600} \sum_{i=1}^{n_h} \sum_{j=1}^{n_{hi}} W_{hij}^2 \sigma_h^2 , \quad (17)$$

where

W_{hij} = weight for student-hij;

X_{hij} = value of variable-X for student-hij,

n_h = number of sample schools in stratum-h, and

n_{hi} = number of sample students in in school-i of stratum-h.

If the weights within the low SES stratum (strata 1-300) were all equal to $\overline{W}_1$ and if those within the high SES stratum all equalled $\overline{W}_2$, then we could rewrite equation 17 as

$$Var_e(X') = \sum_{h=1}^{300} \sum_{i=1}^{n_h} \sum_{j=1}^{n_{hi}} (\overline{W}_1)^2 \sigma_h^2 + \sum_{h=301}^{600} \sum_{i=1}^{n_h} \sum_{j=1}^{n_{hi}} (\overline{W}_2)^2 \sigma_h^2 \quad (18)$$

Now we can approximate the increase in variance due to unequal weighting within the high and low SES strata as

$$\frac{Var(X')}{Var_e(X')} = \frac{\sum_h \sum_i \sum_j W_{hij}^2}{n_1 (\overline{W}_1)^2 + n_2 (\overline{W}_2)^2} \quad (19)$$

where

$$n_1 = \sum_{h=1}^{300} \sum_{i=1}^{n_h} n_{hi}$$

$$n_1 = \sum_{h=301}^{600} \sum_{i=1}^{n_h} n_{hi}, \text{ and}$$

$\sigma_h^2 = \sigma^2$ for all h.

Table 8 shows the average weight values, the sum of the squared weights, and the approximate increase in variance estimated using equation 19. The estimated increase is fairly sizable for all subpopulations and for the total population. This portion of the unequal weighting arises from unequal final stratum sizes, imprecise size measures, and from weight adjustments to correct for non-response. The results in this section should be regarded as rough approximations since assumptions of equal variances within strata and fixed subpopulation sizes are required.

Table 8.—Estimated effect of unequal weighting within low and high SES strata on variances of survey estimates

Subpopulation	Average weight		Sum of squares of weights	Estimated effect of unequal weighting
	Low SES ($\bar{W}_1$)	High SES ($\bar{W}_2$)	(ΣW_i^2)	
Sex:				
Male	112.76	216.17	287,468,845	1.16
Female	111.22	216.57	284,457,979	1.15
Not reported	113.30	205.51	51,472,065	1.21
Race:				
White	112.53	216.36	479,929,309	1.15
Black	109.15	219.83	40,290.221	1.20
Other	114.00	213.58	44,009.940	1.15
Not reported	113.59	207.43	59,169,419	1.21
Father's education:				
Less than high school graduate	113.03	218.45	143,103,131	1.14
High school graduate	112.73	218.60	160,955,721	1.15
College graduate	111.34	213.37	198,398,932	1.15
Not reported	111.04	211.39	120,941,105	1.18
Total sample	112.08	215.43	623,398,888	1.16

III. COMPARING THE STRATIFICATION VARIABLES

The variance modeling described in section II.B. of this report suggests that the NLS school stratification variables reduced the variances of national estimates by approximately 20 percent compared with sampling clusters of students selected from an unstratified school frame. Variances of subpopulation estimates were reduced by lesser amounts, from 6 to 20 percent, depending on the subpopulation. In this section, analyses aimed at determining which stratification variables accounted for most of the reduction in variance are described.

The analysis involved calculating several sets of variance component estimates for a linear variance model which includes terms for the five major stratification variables. By extending the linear model given in section IV of reference [1], variance components corresponding to the following stratification variables were estimated—SES (socioeconomic status), size of school, type of control (public, Catholic, non-Catholic private), geographic region, and proximity to college or university. When the sampling frame was stratified, crossing of the first four of these variables divided the population of schools into 35 strata. Then the fifth stratification variable, proximity, was used to subdivide certain of the 35 strata; this resulted in 64 strata based upon these five variables. Next, a total of 289 major strata were defined by constructing nested substrata within the 64 strata mentioned above based on percent minority (public schools) and average income level (public and Catholic schools). Final strata were defined as nested substrata within major strata, based on degree of urbanization. For the purposes of this analysis, only the five major stratification variables described above were studied.

A difficulty was encountered which relates to the order in which the stratification variables are placed in the model. Since the five major stratification variables may be regarded as crossed, the model could be specified using any one of 5! = 120 models corresponding to the 120 possible arrangements of the five variables. Also, the earlier in the model a variable is placed, the more negative estimates (set equal to zero) will be calculated since the components are estimated from right to left in the model (component for the last term of the model is estimated first). With eight components to be estimated (five stratification effects plus final stratum, school, and student components), the number of negative estimates obtained was expected to be sizable. Thus, it was not clear how to proceed and computing a set of components for each of the 120 models was not considered feasible.

As a first step toward gaining some feel for the relative utility of the five stratification variables, five models were specified and five variance components runs were completed. The models were chosen so that each of the five variables was first in one model and fifth in another model. A subset of 10 of the 21 variables used in the previous variance components study [1] was chosen for this part of the analysis. Also, only four subpopulation estimates were included (males, females, whites, and blacks). Thus 90 statistics were included in each of the 5 variance components runs, 10 national estimates, 40 domain estimates, and 40 differences of domain and national estimates. The analysis consisted of comparing the number of negative variance component estimates for the five stratification variables when the variable was first in the model and also fifth in the model. If the effect of

one of the stratification variables was zero, then we should observe about 50 percent of the estimates for that variance component to be negative. Table 9 shows the number of negative estimates obtained for each of the five variables by type of statistic. When one of the variables is written fifth in the model, estimates of the component are least biased by the large number of terms in the model. Looking at the lower part of table 9, we note that all five of the stratification variables have positive effects. (Using a simple sign test based on the numbers of positive and negative estimates, the hypothesis of zero effect would be rejected for each variable for national means, domain means, differences of domain and national means, and all statistics.) Looking at the upper half of table 9, we see the effects of position in the model on the numbers of negative variance component estimates. Using this data, we would reject the hypothesis of effect equal to zero only for control and region, based on a sign test. But since we know the number of negative estimates will be biased upward due to the large number of terms estimated, we cannot conclude anything from this type of test. We must also keep in mind that we have used only five of the 120 possible arrangements of the model and that the results here may depend on the model used.

About all we can conclude from table 9 is that region appears perhaps the strongest stratification variable, that control is perhaps the weakest, and that the other three variables are somewhere in between. There were also indications that the numbers of negative component estimates for several of the variables were sensitive to the position in the model of the control variable. This was thought to arise from the extreme large differences in the population and sample sizes for the three levels of control—students enrolled in public, Catholic, and non-Catholic private schools.

Table 9.—Number of negative variance component estimates for stratification terms in first and fifth positions in model

Variable and position in model	National means		Domain means		Difference of domain and national means		All statistics	
	Negative estimates	Total	Negative estimates	Total	Negative estimates	Total	Negative estimates	Total
First position								
SES	2	10	17	40	25	40	44	90
Size	1	10	13	40	25	40	39	90
Control	5	10	29	40	34	40	68	90
Region	0	10	5	40	13	40	18	90
Proximity	2	10	17	40	24	40	43	90
Fifth position								
SES	0	10	2	40	1	40	3	90
Size	1	10	5	40	8	40	14	90
Control	0	10	9	40	12	40	21	90
Region	0	10	0	40	4	40	4	90
Proximity	1	10	4	40	3	40	8	90

For the aforementioned reasons, it was decided to eliminate control from the model and enter region in the model at first position. Then to evaluate the relative importance of the remaining three variables, the three were permuted in all 3! = 6 possible orders and six additional variance component runs were made using the same 10 variables and the same four domains as used in the previous runs. The orderings of the stratification variables for the six variance component runs were:

Region—SES—Size—Proximity,
Region—SES—Proximity—Size,
Region—Size—SES—Proximity,
Region—Size—Proximity—SES,
Region—Proximity—SES—Size, and
Region—Proximity—Size—SES.

The number of negative component estimates was observed for each of the three stratification variables in positions two, three, and four These counts are shown in table 10. A sign test would result in rejection of the hypothesis of a zero effect for each variable in each position. Thus, we can conclude that each of these variables was effective in reducing the variances of estimates. If we use the number of negative variance component estimates as a criterion describing the magnitude of the effects, then we might conclude that the five stratification variables might be ranked from most useful to least useful as region, SES, proximity, size, and control. Thus, while we have not been able to precisely estimate how much of the stratification effect to attribute to each of the variables, we have some rough indications of the relative importance of the five major stratification variables. We also have an indication that control may not have been a very useful stratification variable, but that region, SES, size, and proximity were all useful stratification variables.

Table 10.—Number of negative variance component estimates for terms in second, third, and fourth positions in model

Variable and position in model	National means		Domain means		Difference of domain and national means		All statistics	
	Negative estimates	Total	Negative estimates	Total	Negative estimates	Total	Negative estimates	Total
Second position								
SES	0	20	14	80	25	80	39	180
Size	2	20	15	80	28	80	45	180
Proximity	0	20	19	80	28	80	47	180
Third position.								
SES	0	20	6	80	7	80	13	180
Size	0	20	6	80	17	80	23	180
Proximity	0	20	6	80	12	80	18	180
Fourth position.								
SES	0	20	3	80	2	80	5	180
Size	0	20	6	80	10	80	16	180
Proximity	2	20	6	80	4	80	12	180

REFERENCES

[1] R.P. Moore, B.V. Shah, and R.E. Folsom, "Efficiency Study of NLS Base-Year Design," Report on RTI Project 22U-884-3, Research Triangle Institute, Research Triangle Park, N.C., November 1974.

[2] Leslie Kish, *Survey Sampling*, John Wiley and Sons, New York, 1965.

[3] William B. Fetters, "Sampling Errors of NLS Base-Year Student Questionnaire Percentages and Test Battery Mean Scores," Memorandum and enclosures, dated September 26, 1974.

[4] Ralph E. Folsom, Jr., "Variance Components for NLS: Partitioning the Design Effect," Report on RTI Project 22U-884-3, Research Triangle Institute, Research Triangle Park, N.C., January 1974.

[5] "Sample Design for the Selection of a Sample of Schools with Twelfth-Graders for a Longitudinal Study," unpublished manuscript, Westat, Inc., Rockville, Md., June 1972.

[6] William B. Fetters, "Class of 1972 Sample—Degree of Oversampling of Target Groups Achieved," Memorandum and enclosures, dated October 18, 1974.

[7] Joseph Waksberg, "The Effect of Stratification with Differential Sampling Rates on Attributes of Subsets of the Population," *Proceedings of the Social Statistics Section, American Statistical Association*, 1973.

☆ U. S. GOVERNMENT PRINTING OFFICE : 1977 241-055/2010