OVERSAMPLING MINORITY SCHOOL CHILDREN

Ralph DiGaetano, David Judkins, and Joseph Waksberg, Westat Inc.
Westat Inc., 1650 Research Boulevard, Rockville, Maryland 20850

Keywords: Rare populations, sampling frame, design effects, effective sample sizes, sample allocation.

1. Introduction

This paper examines sampling strategies for a survey of school children in which minority school children are to be oversampled. We recently faced this problem in planning a survey utilizing a two-stage design, where the first stage was to be a sample of schools and the second stage a sample of children. Constraints included a fixed overall sample size and selection of one class (roughly 20 to 30 students) per grade in a sample of grades per school. For at least part of the study, it did not appear practical to use variable sampling rates within schools, and the oversampling needed to be restricted to the first stage units. There is a fair amount of literature on the efficiency of such designs, mostly referred to as sampling for rare populations. The empirical studies, as in the case of the previous paper in this session, are mostly concerned with area samples, but obviously the theory applies equally to other definitions of first stage units. A frequently used data source for sampling schools contains information that can be used for oversampling minorities. However, there are inadequacies in the data file from this source, and we describe modifications in the sample design that are needed to compensate for the frame problems.

The sampling frame analyzed was the national list of school districts and schools maintained by QED (the National Education Database of Quality Education Data, Denver, Colorado). The QED data base provides school counts for several variables, including total enrollment, grade ranges, and the number of Black and Hispanic students. Data on Black and Hispanic students are missing for most private schools, but a relatively small percentage of minority students are enrolled in private schools.

For most purposes, QED is considered a very good data source. It uses the most recent information available in the Common Core of Data (CCD) listing for public schools put together by the National Center for Education Statistics (NCES), and attempts to update the CCD material on an annual basis. The QED data base appeared to be the best source of school data for sampling purposes. The QED data base used in analyses for this paper reflected schools in existence during the 1993-1994 school year.

2. Sample Design if QED Data are considered Accurate for Minorities

2.1 Distribution of Minority students Among Schools

In this section we discuss the implications of using the QED frame for purposes of oversampling minority students. For the present, we will assume the QED figures on minority enrollment are correct. We change this assumption later on. The context for our evaluation is an attempt to achieve approximately equal reliability for estimates of Black, Hispanic, and all other students. Schools are assigned to sample strata based on the racial/ethnic distribution of their enrollment as obtained from the QED frame. For this analysis, we assume that operational factors require that all students in a school be selected at the same rate, and, as a result, students in the same stratum have approximately the same probability of selection. We wish to examine the extent to which the effective sample sizes for the demographic groups of interest under several sampling plans for oversampling minorities can be made approximately equal, thus providing estimates of roughly equal precision. The effective sample size is the ratio of the actual sample size for a group of interest divided by the expected design effect associated with the specified sample design. The expected design effect reflects both differential weighting associated with the use of different sampling rates for the designated strata and the effect of clustering the sample of students (within sample school district and within sample school).

Our research concentrated on two issues: (1) what are efficient sampling rates within strata, and (2) is it possible to achieve approximately equal precision for the three race/ethnic groups with the two-stage sample design?

Five strata were established for the analyses, taking into account coverage and student distribution. Because we were attempting to achieve maximum precision for two separate domains of interest, the choice of strata could not be based on a straightforward optimization strategy. Table 1 shows the distribution of the three race/ethnicity groups (Hispanic, Black, and other) within each of the five strata separately and the percentage of each group which falls into a given stratum. For example, the students in the Hispanic-high stratum (elementary, middle, and secondary schools combined) are about 77 percent Hispanic, 7 percent Black, and 16 percent other, and these students represent 50, 3.4, and 1.6 percent of all Hispanic, Black, and other students, respectively.

It should be noted that the Hispanic-mid and Black-mid strata are predominantly "other", although they do contain roughly one-third of the targeted demographic group. About three-fourths of the students in the Hispanic-high and Black-high strata are in the targeted group, and each stratum contains about half of the targeted population. As will be seen, a major limitation in oversampling minority students stems from the fact that large proportions of the Black and Hispanic student populations are found in schools that are heavily "other."

2.2 Effective Sample Sizes

For the survey being considered, estimates were desired for each grade (age) separately. We have assumed a design effect associated with clustering alone of roughly 2 for the three individual race/ethnicity groups of interest and 2.5 for all three groups combined, arising from the roughly 20 to 30 students per grade selected per school. The estimated relvariances of the weights resulting from specified differential sampling rates were added to the design effects associated with clustering to obtain the overall estimated design effects used in the evaluations. Other surveys will, of course, have different clustering patterns, but unless the design effects are very different from 2, the implications for sample design will be similar to the results of our analysis.

In order to achieve equal reliability (e.g., equal coefficients of variation) for the three race/ethnicity categories, the effective sample sizes should be roughly the same. (The planned survey required a total of 31,000 responding students or roughly 2,400 per grade but the same oversampling rates and relationships of effective sample sizes would apply with any other total.) The sample sizes and design effects were examined for a number of alternate sampling rates among the five strata. The alternatives explored the effect of different oversampling rates for the high and mid-minority strata, although all were constrained to have a total of 31,000 students.

Increasing the rates in the high minority strata obviously produces higher numbers of minorities in the sample, but it also increases the design effects. Beyond a certain point, increases in rates are counter-productive because they result in reductions in effective sample size for the minorities. This is illustrated in Table 2.

Section (2a) of Table 2 shows the actual and effective sample sizes with an equal probability sample. The distribution of the sample among race/ethnic groups reflects their proportions in the population, both for the actual and effective sample sizes. The non-black, non-Hispanics amount to about three-fourths of the sample and are about five times as large as blacks or Hispanics. Section (2b) shows comparable data when the higher minority schools are oversampled at a fairly high rate. The disparity in effective sample sizes has been sharply reduced, but the effective size for nonminorities is still about 2.5 times that for blacks or Hispanics. Increasing the sampling rate in the high minority strata further does not improve the distribution, as is illustrated in Section (2c). The only way to make the variances equal for the three groups is to reduce the effective sample sizes for the "other" group substantially without improving the minority statistics. For example, in Section (2b) when the relative sampling rate is 4.75 for the Hispanic-high stratum, 3 for the Hispanics-mid stratum, 3.5 for the Black-high stratum, and 2 for the Black-mid stratum, the expected sample yield for Hispanics is around 6,700, but the effective sample size is close to 2,900 due to an estimated overall design effect of 2.34. For corresponding relative sampling rates of 6, 2, 4, and 2 (Section (2c)), the actual sample yield for Hispanics is around 7,200 (an increase of 600) but the effective sample size is only around 2,800 because the estimated overall design effect is substantially increased (2.58). For similar reasons, the effective sample sizes for Blacks are about the same.

3. Quality of QED Data on Race/Ethnicity

The distributions of school children by race/ethnicity in Table 1 and the implications of these distributions for an efficient sample design are based on the assumption that the QED data are accurate and up-to-date. The QED figures are reasonably current; the data are revised annually as new information becomes available. However, a comparison of the number of Black and Hispanic students in the QED data file with other sources of similar information raises serious questions about the accuracy of the QED data. (We should point out that this concern relates only to the accuracy of the classification of students as Black, Hispanic, or other. As far as we are aware, the estimates of total students are very good. For example, the total number of school children in public schools in the 1993-1994 QED file differed by only 1 percent from the CPS estimate of public school children.)

The comparison of QED data with similar counts from the National Assessment of Educational Progress (NAEP) for 12th grade students shown in Table 3 illustrates the reasons for concern. Table 3 uses the same type of stratification of schools as the earlier tables, that is, the strata comprise schools with designated percentages of Blacks or of Hispanics. The two columns for Blacks use stratification based on percentage of Blacks; similarly, the two columns for Hispanics use the percentage of Hispanics as the stratifying variables. The QED columns show what is reported by QED for the strata. The NAEP columns contain weighted totals of the school children in the NAEP sample. The classification of schools by strata are based on the QED data base for both sets of data.

There are several reasons to be skeptical of the QED numbers. It is obvious that the distributions by strata are very different. For Blacks, the QED and

NAEP estimates are reasonably close for all strata except the one containing the 0-9 percent Blacks. Here, QED shows only a trivial number of Black students while NAEP reports that over 20 percent of Black 12th graders are in schools in the stratum. A somewhat similar pattern applies to Hispanics, although the shortage in the 0-9 percent stratum is much greater, and there is a fairly sizable discrepancy in the other direction in the 30-59 percent stratum.

One could consider that a possible explanation for the discrepancy is poor reporting of race/ethnicity in NAEP. We think this is very unlikely for 12th graders. The NAEP data come from self-identification of race/ethnicity of students in the NAEP sample. It would be surprising to find that 12th grade white students who are mostly 17 and 18 years old are not keenly aware of their race/ethnicity status. It would be even stranger to discover that almost all of the ones who erroneously report themselves as minorities go to schools with small percentages of minorities.

Another possible explanation is that QED numbers are just slightly off, but enough to affect the classification of schools by percent minority. Table 4 compares NAEP estimates of Black and Hispanic 12th graders when schools are classified according to the QED and when the principals' reports of the number of minority students in their schools during NAEP are used for the classification. It can be seen that the principals' classifications correspond closely to that of QED. If there is any problem in QED, it comes from the counts of minorities, not from errors in classification. It seems likely that the source of the understatement of minority student counts in low density minority schools are reports from principals of these numbers.

Comparison of the 1993-1994 QED total number of 12th graders who were black (330,669) and Hispanic (184,686) with figures from the 1993 Current Population Survey (CPS) sheds further light on the problem. The corresponding CPS figures are 532,000 black and 333,000 Hispanic 12th grade students. Of these, 420,000 blacks and 268,000 Hispanics were 18 years old or younger. Comparing QED with CPS shows a shortage that is close to 40 percent for Blacks and 45 percent for Hispanics. The shortage may be somewhat overstated since some of the older students in CPS may be studying for high school equivalency or not formally enrolled for other reasons. However, even if all of the older (more than 18) students are eliminated from the comparisons, shortages of 20 to 30 percent are indicated. The actual shortages are probably closer to 30 to 40 percent.

Our analysis of QED coverage has concentrated on 12th grade students. Comparisons with NAEP have also been carried out for fourth and eighth graders, the other two groups covered in NAEP. The general patterns were similar, although the shortages were not as great. Furthermore, the total number of minority school children in QED is about 15 percent below the comparable CPS data for both Blacks and Hispanics.

Basically, it looks as if the QED contains reasonable data for schools with a relatively large proportion of minority students but seriously underestimates minorities in schools that are predominantly white. A sampling strategy that is based only on QED, and thus assumes that only a trivial proportion of the minorities are in one of the strata would be highly inefficient because it would seriously undersample that stratum. It seems clear that even though QED is needed as the sampling frame and for the school stratification, the NAEP distributions should be used to determine sampling rates. We do this in the next section.

4. Revising the Projected Impact of Differential Sampling

Acceptance of the NAEP estimates of the distribution of minority student populations has dramatic effects upon optimal sample design. As was discussed earlier, even if the QED distributions were accurate, the potential for simultaneously increasing (through differential school-level sampling only) the effective sample sizes for Black and Hispanic students is limited unless one is willing to suffer serious degradation in precision for statistics about miscellaneous minorities such as Asian students, about white students, and about all students together. Once we factor in the deficiencies in the stratification information, the effectiveness of differential school-level sampling to improve the efficiency of black and Hispanic student statistics is even more limited.

The ideal methodology for revising the impact of differential sampling would be to tabulate the NAEP sample according to the strata defined in Table 1. Unfortunately, timing and resource issues prevented such a tabulation, but the tabulations shown in Table 5 can be used to shed light on the likely actual impact of taking QED data to guide the allocation. Instead of studying simultaneous improvement for blacks and Hispanics, we studied the effects of taking QED data to guide an "optimal" allocation for either blacks or Hispanics, but not both. Since the results were fairly similar for 4th, 8th, and 12th graders, the results presented here concern only 12th graders. Table 5 contrasts how the optimal sample allocation varies depending upon the information source that is used to guide the allocation process. A proportional allocation (using QED total enrollment counts by stratum) are also indicated. The optimal allocation indicated by NAEP lies between those indicated by the QED data and those suggested by simple proportional allocation.

Table 6 shows the impact of these different allocations. If the QED data were accurate, then a sample allocation that sought to maximize precision for black student statistics would yield a 139 percent increase in the effective black student sample size relative to a multi-stage design with proportional allocation. Such an allocation would reduce the effective sample for everyone else by about two thirds. However, the 1992 NAEP data indicate that the improvement in black student statistics would be illusory. No significant improvement relative to proportional allocation would be achieved for black student statistics if the "optimal" sampling rates driven by the QED data were employed. Interestingly, the predicted degradation for statistics for other domains is quite close to what was predicted based solely on QED data. Real improvement relative to proportional allocation would be achieved if the optimal sampling rates (for black students) suggested by the NAEP data are employed. In this case, the best improvement that can be achieved for black statistics is about 40 percent. There is still a penalty to be paid in the precision for other domains (a roughly 25 percent loss for the other domains), but the penalty is much smaller than if the "optimal" sampling rates driven by the QED data were employed (a roughly 65% loss for the other domains). Similar results hold true when the sample is allocated to maximize the precision of Hispanic student statistics.

5. Conclusions

The main conclusion that we draw for student statistics is that oversampling schools can lead to modest improvements in minority statistics, but that this technique is not powerful enough to yield the same precision for minority statistics as for statistics about students who are neither black nor Hispanic. Black and Hispanic students are not sufficiently concentrated and where they do concentrate, they tend to concentrate separately, making the simultaneous improvement of statistics for both domains difficult. Furthermore, measurement error in the sampling frame data can have a major impact on the precision achieved. Additional improvements in the precision for minority statistics require either an overall increase in sample size or some sort of screening and subsampling operation within schools. Subsampling procedures could be somewhat problematical for young students, as they are generally associated with a single class and thus the sampling of entire classes is most efficient and least disruptive for them. Consequently, there is probably very little that can be done to improve the precision for statistics about black and Hispanic primary students other than to increase the total sample size. However, it appears that lists of middle and high school students indicating race/ethnicity may often be available, and these students are generally not associated with a single class throughout a school day. Thus, subsampling of students beyond the primary school level is likely to be much more feasible, improving the precision of estimates for these minority students.

A broader conclusion that we draw from this work and from work on area sampling (Judkins, Massey and Waksberg, "Patterns of Residential Concentrations By Race and Hispanic Origins", ASA 1992 Proceedings of the Section on Survey Research Methods) concerns the importance of allowing for error in the information on the concentrations of rare domains across various strata when attempting to improve the precision for rare-domain statistics through differential allocation across those strata. In the case of area sampling, the movement of people from one type of neighborhood to other types of neighborhoods over the course of a decade implies that a compromise allocation between proportional allocation and the apparent optimal allocation driven by data from the last decennial census will often yield better results than optimal allocation that is fully conditioned on the auxiliary data. In the case of student statistics, QED does have reasonably current information, but apparent inaccuracies in racial statistics provided by school principals to the individual state departments of education mean that a compromise allocation should also be used in this setting.

Table 1. Distribution and coverage of race/ethnicity groups by strata for elementary, middle and high schools combined*

School stratum	No. of students	% of students within stratum	Distribution within stratum				Race/ethnic coverage by stratum		
			Hispanic	Black	Other	Total	Hispanic	Black	Other
Hispanic-high	2,955,378	7.3	76.5	7.2	16.2	100.0	50.0	3.4	1.6
Hispanic-mid	4,070,733	10.0	33.3	15.9	50.8	100.0	30.0	10.3	6.9
Black-high	4,250,778	10.5	2.3	74.1	23.7	100.0	2.1	50.0	3.4
Black-mid	3,818,819	9.4	2.9	32.9	64.2	100.0	2.4	20.0	8.2
Other	25,511,490	62.8	2.7	4.0	93.2	100.0	15.4	16.3	79.9
Total	40,607,198	100.0	11.1	15.5	73.4	100.0	100.0	100.0	100.0

*Students in private, non-Catholic schools or schools in nonstandard grade ranges not included

Table 2. Expected effective sample sizes for the three race/ethnicity groups for specified sampling rates within strata

School strata	Relative within strata sampling rates of schools	Student group of analytic interest	Expected sample yield of students for race/ethnicity group of interest	Estimated design effect for both clustering and differential sampling rates	Expected effective sample size of students
(2a)					
Hispanic-high	1.00	Hispanic	3,452.8	2.00	1,726.4
Hispanic-mid	1.00	Black	4,805.7	2.00	2,402.8
Black-high	1.00	Other	22,741.5	2.00	11,370.6
Black-mid	1.00	Total	31,000.0	2.50	12,399.9
Other	1.00				
(2b)					
Hispanic-high	4.75	Hispanic	6,706.2	2.34	2,862.9
Hispanic-mid	3.00	Black	7,313.1	2.25	3,257.0
Black-high	3.50	Other	16,980.8	2.20	7,731.9
Black-mid	2.00	Total	31,000.0	2.88	10,767.4
Other	1.00				
(2c)					
Hispanic-high	6.00	Hispanic	7,171.1	2.58	2,784.4
Hispanic-mid	2.00	Black	7,631.8	2.32	3,284.4
Black-high	4.00	Other	16,197.2	2.18	7,427.4
Black-mid	2.00	Total	31,000.0	2.93	10,580.8
Other	1.00				

Table 3. Black and Hispanic 12th graders reported in NAEP and QED when schools are classified by QED data

Schools classified by % designated minority (QED)	Blacks total (NAEP)	Blacks total (QED)	Hispanics total (NAEP)	Hispanics total (QED)
0-9	83,548	6,562	99,697	8,044
10-29	80,275	75,909	39,832	43,880
30-59	129,694	119,380	35,472	60,467
60+	108,192	128,848	71,285	72,295
Total	401,708	330,699	246,286	184,686
	(Weighted estimates)	(Actual counts)	(Weighted estimates)	(Actual counts)
Percent				
0-9	20.8	2.0	40.5	4.4
10-29	20.0	23.0	16.2	23.8
30-59	32.3	36.1	14.4	32.7
60+	26.9	39.0	28.9	39.1
Total	100.0	100.0	100.0	100.0

Table 4. Black and Hispanic 12th graders reported in NAEP when schools are alternatively classified by QED data and reports from school principals

Schools classified by percent designated minority	NAEP sample distribution	Percent of total	NAEP sample distribution	Percent of total
By QED data				
0-9	83,548	20.8	99,697	40.5
10-29	80,274	20.0	39,832	16.2
30-59	129,694	32.3	35,472	14.4
60+	108,192	26.9	71,285	28.9
Total	401,708	100.0	246,286	100.0
By principals' report				
0-9	82,679	20.6	99,697	40.5
10-29	82,479	20.5	42,576	17.3
30-59	122,611	30.5	38,559	15.7
60+	113,939	28.4	65,454	26.6
Total	401,708	100.0	246,286	100.0

Table 5. Comparing optimal and proportional allocation schemes for 12th graders based on different information sources

% of schools who belong to targeted minority	1993-1994 QED		1992 NAEP		Optimal allocation of total sample				Proportional allocation	
	National percentage of targeted population who attends such schools				Given 1993-1994 QED data		Given 1992 NAEP data		Given 1993-1994 QED data	
	Black	Hisp	Black	Hisp.	Black	Hisp.	Black	Hisp.	Black	Hisp.
0-9	2.0	4.4	20.8	40.5	179	323	458	689	718	867
10-29	23.0	23.8	20.0	16.2	302	267	194	126	134	72
30-59	36.1	32.7	32.3	14.4	295	221	213	78	100	31
60+	39.0	39.1	26.9	28.9	223	189	135	108	48	30
Total	100.0	100.0	100.0	100.0	1,000	1,000	1,000	1,000	1,000	1,000

Table 6. Change in effective sample size relative to an equi-probability sample with the same extent of clustering

	When the sample is designed to maximize precision for black 12th graders			When the sample is designed to maximize precision for Hispanic 12th graders		
Population of interest	Expectation s using QED data	"True" impact of using rates suggested by QED data	Result of using rates suggested by NAEP data	Expectations using QED data	"True" impact of using rates suggested by QED data	Result of using rates suggested by NAEP data
For targeted minority	138.7%	0.8%	40.5%	193.9%	1.3%	35.2%
For other minority (Black or Hispanic)	-63.3%	-63.5%	-24.6%	-53.2%	-47.6%	-14.4%
For majority population and miscellaneous other minorities	-66.7%	-66.0%	-28.1%	-55.4%	-49.9%	-16.8%
For total population	-63.1%	-61.9%	-22.0%	-52.8%	-47.1%	-13.3%