

UNDERCOVERAGE IN A LARGE NATIONAL SCREENING SURVEY FOR YOUTHS

Mike Horrigan, BLS, Whitney Moore, Steven Pedlow, Kirk Wolter, NORC
Steven Pedlow, National Opinion Research Center, 55 E. Monroe, Chicago, IL 60603

Key Words: NLSY97, Coverage Ratios.

1. Introduction. The National Longitudinal Survey of Youth (NLSY97) is the latest in a series of surveys sponsored by the U.S. Department of Labor (DoL) to examine issues surrounding youth entry into the work force and subsequent transitions in and out of the work force. Immediately following the data collection for the first round of the survey, the NLSY97 respondents and an additional sample of persons aged 18 to 23 were administered the Armed Services Vocational Aptitude Battery (ASVAB). This additional sample of 18- to 23-year-olds participated in a separate study called the Profile of American Youth (PAY97).

PAY97 was sponsored by the U.S. Department of Defense (DoD). Originally done in 1980, its purposes were to assess the vocational aptitudes of contemporary American youths and to establish current national norms for the ASVAB. This second round of the project was done not only to renorm the ASVAB, but also to assess the vocational interests of American youths through the administration of the Interest Finder (IF). PAY97 actually encompassed two different surveys: the Enlistment Testing Program (ETP) assessed roughly 6,000 youths aged 18-23, and the Student Testing Program (STP) assessed around 4,700 youths enrolled in grades 10, 11, or 12 as of Fall, 1997.

The NLSY97 will follow a cohort of approximately 9,000 youths who completed an interview in 1997 (the base year). These youths were between 12 and 16 years of age as of December 31, 1996.

The overall study design for NLSY97 incorporated the PAY97 sample and resulted in one large screening sample of over 90,000 housing units to generate youth participants for both NLSY97 and PAY97.

This paper details the large screening sample for NLSY97/PAY97, and briefly discusses the design of the sample. We then provide information about the undercoverage of youths in the eligible age range and possible reasons for the shortfalls.

2. Overview of the Sample. The NLSY97, ETP, and STP samples were selected in two broad phases. First, we specified a large sample of more than 90,000 housing units. Through fieldwork, we determined which housing units were occupied and which were vacant, and for the occupied housing units, we conducted brief screening interviews. In the second phase, we selected subsamples of the eligible youths identified during screening for participation in the main NLSY interview and in the ASVAB and IF tests. Both the NLSY and the ETP targeted sizeable oversamples of Hispanic and non-Hispanic, black youths. The STP targeted an approximately proportionate representation of six grade (10, 11, and 12) by sex (male and female) domains.

We designed the sampling methods and procedures to yield a database of eligible youths that could be projected (with known confidence levels) to represent the entire eligible population of American youths. For the large screening sample, our approach involved the selection of two independent area-probability samples: 1.) a cross-sectional (CX) sample designed to represent the various segments of the eligible population in their proper population proportions, and 2.) a supplemental (SU) sample designed to produce, in the most statistically efficient way, the required oversamples of Hispanic and non-Hispanic, black youths.

Both the cross-sectional and supplemental samples were selected by standard area-probability sampling methods. Sampling was in three essential stages: primary sampling units (PSUs) consisting mainly of Census metropolitan statistical areas (MSAs) or single counties, segments consisting of single census blocks or clusters of neighboring blocks, and housing units (HUs). All eligible youths in each household were then selected for interviewing and testing.

We rigorously tested our sampling procedures by using sampling weights to “weight-up” counts at each stage of sampling. These weighted-up counts verified representativeness of the selected sample.

However, it should be noted that the sample was selected using 1990 Decennial U.S. Census counts, while the survey was performed in 1997.

3. Results of Screening. We screened in a total of 90,957 households in order to meet our targets for the NLSY and PAY samples. 80,204 of these households were eligible to participate (88% eligibility rate), and 75,409 completed a screener (94% response rate).

Target sample sizes were 10,000 completed interviews for the NLSY sample, 6,000 completed tests for the ETP, and 6,164 completed tests for the STP. We located 9,817 youths eligible for the NLSY, 8,846 ETP-eligible youths, and 6,364 STP-eligible youths. Of the 9,817 total NLSY97-eligible youths found, 9,022 interviews were completed (92% response rate), which is almost 1,000 short of the goal.

4. Coverage ratios. Various analyses were conducted by NORC to determine both the coverage and representativeness of the NLSY97 and PAY97 samples. The March 1997 Current Population Survey (CPS) was used as a comparison (i.e. the denominator for the coverage ratios). The analyses illustrated the shortfalls found for the age ranges critical to the NLSY97/PAY97. Estimated coverage ratios are given below in Table 1:

Table 1. Estimated NLSY97/PAY97 Coverage Ratios

Age Range	Both CX and SU	CX	SU
0-11	0.90	0.89	0.94
12-16	0.74	0.72	0.83
18-23	0.68	0.69	0.64
24-35	1.01	1.01	1.00

While screening located 90 percent of the expected persons below age 12, and a surplus (101 percent of the expected number) of persons between ages 24 and 35, there is a large underrepresentation of youths aged 12 to 23. In fact, less than 75 percent of the expected number of youths aged 12 to 23 were located in the screening operation. Figure 1 on the following page shows our estimates of these coverage ratios graphically for each year of age between 0 and 35. Using special household

weights, the NLSY line on Figure 1 shows an estimate of the number of persons of each age year nationally. The CPS-Expected line shows an estimate given by the special household weights multiplied by CPS actual percentages. The difference between the two lines is the undercoverage NLSY97 shows for that age year.

However, all surveys suffer from some level of general undercoverage. The presumed true population is the Decennial U.S. Census population adjusted for the census undercount. This undercount adjustment is the undercoverage in the Census. Estimated coverage rates for the 1990 Decennial Census are shown in Table 2 below:

Table 2. Estimated 1990 Census Coverage Ratios (Mar. 1998)¹

Race	Sex	Age	Coverage Ratios
Black	Male	10-19	.9805
		20-29	.9091
	Female	10-19	.9787
		20-29	.9653
Non-Black	Male	10-19	1.0089
		20-29	.9830
	Female	10-19	1.0053
		20-29	.9937

Coverage ratios are not given for Hispanics, but Census experts often assume that they are similar to coverage ratios for blacks. Among the age groups of NLSY97/PAY97, this undercoverage is greatest for black males aged 20-29 – over 9% were not counted. For most other groups, the undercoverage is small, and non-black 10-19 year-olds actually show overcoverage (i.e. some counted more than once).

The Current Population Survey has greater undercoverage than the Decennial Census. Overall, the CPS covers about 92 percent of the adjusted decennial census population. However, the CPS has less coverage for the age groups of NLSY97/PAY97, as shown by Table 3 below:

¹ These coverage ratios come from Robinson et al. (1993) "Estimation of Population Coverage in the 1990 United States Census Based on Demographic Analysis," *Journal of the American Statistical Association*, Vol. 88, pp. 1061-1071.

Figure 1: Weighted Counts of Persons, Total Sample

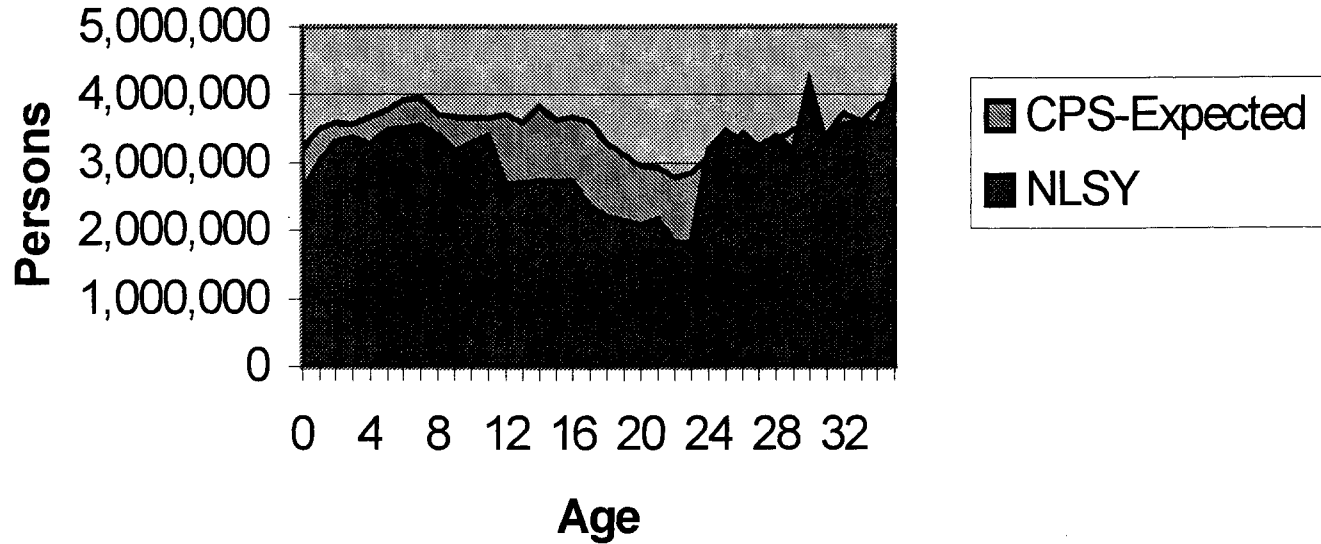


Table 3. Estimated CPS Coverage Ratios (March, 1998)

Race	Sex	Age	Coverage Ratios
Black	Male	0-15	.844
		16-19	.711
		20-29	.660
	Female	0-15	.838
		16-19	.802
		20-29	.811
Non-Black	Male	0-15	.929
		16-19	.881
		20-29	.847
	Female	0-14	.964
		15-19	.891
		20-29	.897

Again, coverage ratios are not given separately for Hispanics. The coverage ratios tend to be lowest among 20-29 year-olds and 16-19 year-olds. Undercoverage is most severe for black males aged 20-29 (.660), which indicates that one in three is not represented. Clearly, the CPS has trouble with some of the age groups studied by NLSY97/PAY97.

Reviewing Table 1, it is clear that NLSY97/PAY97 suffers from more undercoverage than the CPS. An overall summary, then, is that even the overwhelming and costly Decennial Census suffers from some undercoverage. Large and ongoing samples such as the CPS logically suffer from more undercoverage than the Decennial Census. One-time samples such as NLSY97/PAY97 logically suffer from more undercoverage than the CPS.

5. Examination of the Shortfall. The sample of housing units has also been verified to be representative. Therefore, the shortfall and undercoverage must have occurred during the screening operation. As we have seen above, the NLSY97/PAY97 undercoverage is greatest for the eligible age range (12-23 year-olds), and is much smaller outside of this age range. This undercoverage has been widely examined, and seems to be uniform across the sample. Undercoverage appears to be equal in the cross-sectional and supplemental samples, as well as metropolitan and non-metropolitan areas, etc. While we do not yet have coverage ratios for demographic and

socio-economic status (SES) groups, preliminary indications are that the coverage ratios for NLSY97/PAY97 will follow well-established patterns (e.g. lower among minorities).

One difference that has been identified in NLSY97/PAY97 is that smaller household sizes were observed than for the CPS, as shown in Table 4:

Table 4. Average Household Size

Sample	Number of Persons
CPS, March 1993 – Weighted	2.62
CPS, March 1997 – Weighted	2.89
NLSY97/PAY97 – Weighted	2.34
NLSY97/PAY97 – Weighted (CX PSUs)	2.34
NLSY97/PAY97 – Weighted (SU PSUs)	2.41

There are two possibilities for this difference. Either larger households (those with kids?) were less likely to be screened into the survey, or, individuals (kids?) were missed by the screening effort.

Table 5 below shows distributions of households by number of youths in the relevant NLSY97 age range. It shows that the NLSY97/PAY97 sample had more households with zero NLSY-eligible 12-16 year-old youths (89% vs. 85%) in the pertinent age range and fewer households with one or two youths than the CPS did. It should be noted that the fair comparison here is between the CPS and the CX sample because the SU sample oversamples minorities, while the CPS and CX are representative national samples.

Table 5. Distribution of Households by Reported Number of Youths Aged 12 to 16

Number of Youths	Proportion of Households			
	CPS Weighted	NLSY97-Weighted		
		Both	CX	SU
0	.848	.892	.894	.876
1	.114	.079	.077	.091
2	.034	.026	.026	.029
3	.004	.004	.004	.003
4	.000	.000	.000	.000
5+	.000	.000	.000	.000

So NLSY97 had fewer households with eligible 12-16 year-olds than the CPS would lead us to expect. Of the households with at least one NLSY97 eligible, Table 6 shows that NLSY97 had a higher number of eligibles per household:

Table 6. Conditional Mean Youths, Age 12-16, Per Household

Race/ Ethnicity	Mean Number of Youths, Age 12-16, Per Household With ≥ 1 Such Youth			
	CPS	NLSY97-Weighted		
		Total	CX	SU
Hispanic	1.303	1.377	1.376	1.386
Non-Hisp., Black	1.300	1.328	1.324	1.337
All Other	1.275	1.303	1.308	1.288
Total	1.282	1.316	1.320	1.302

The conditional household distribution of the number of NLSY97 eligibles, given that there was at least one, also shows that NLSY97 households with eligibles tend to have more eligibles than the CPS would lead us to expect. After controlling for the existence of at least one youth, the overall NLSY97 sample has fewer households with only one such youth, and more with two or more such youths.

It seems clear that the NLSY97 shortfall is concentrated among households with only one eligible youth. It also seems clear that NLSY97 did not fail to record some youths in households with more than one eligible youth. If youths were screened in for a household, all youths in these households seemed to be screened in. The evidence shown here leads us to believe that children are being hidden from NLSY97, and that it is easier to hide one eligible child than more than one. As a survey organization, we would much prefer that the children be acknowledged, and then refuse to participate in our survey. But we believe that people like avoiding this confrontation, and like the quickness with which they can dismiss the interviewer by claiming that no eligibles are present.

6. Possible Reasons for “Hidden” Refusals. Presumably, in order to hide eligible youths from NLSY97, they would have to know who is eligible. Survey information, such as the eligible age range, could have been discovered prior to screening in several ways. First of all, the screening materials sent to the household prior to interviewing, which included an advance letter and information about the study, disclosed the approximate age range and may have discouraged participation. The first letter names the survey, “The National Longitudinal Survey of Youth, 1997,” and says that its purpose is “to construct a statistical profile of young persons...” Both of these statements indicate that the survey involves young people, though it doesn’t disclose exact ages. Later in the letter, the length of the interview is given: “...we will ask for his or her permission to spend approximately one hour with the interviewer...” This may have discouraged participation, leading to possible underreporting of youths. A follow-up letter discloses exact ages: “We are contacting households in your area to identify individuals who will be 12 through 16 years old as of January 1, 1997.” It also requests that the individual “take a 2-hour test on a computer...” Any of these statements may have resulted in purposeful underreporting of youths or misreporting of resident youths’ ages to avoid participation in the study.

Household informants may also have gathered information from neighboring households about the nature of the survey. Respondents who had already been screened by a field interviewer may have disclosed information about the survey, such as the eligible age range, to neighboring households. This, too, may have caused respondents to underreport youths in that age range.

Finally, the field interviewers themselves may have given this information to respondents during the screening process. In fact, of the 96 interviewers participating in the debriefing study, 38 (nearly 40 percent) reported telling informants the ages of youths eligible to participate in NLSY97/PAY97. Twenty-two of the interviewers, about 23 percent, followed strict protocol during the screening interviews.

7. Conclusions and Recommendations. It is well known among researchers that the response rate for surveys has been dropping, and that the decline in the willingness to participate is among the greatest challenges facing survey research today. With the saturation of the American public with market research polling, and an increase in their privacy concerns, it is a bad idea to disclose the screening criteria in advance of the actual screening. Rather than disclose the presence of eligible youths and then decline to participate, it may seem easier to avoid the disclosure of age-eligible children. This appears to be the only possible explanation for the underrepresentation of youths aged 12 to 23 in the age distribution.

We recommend that similar studies that target subgroups of the population refrain from literature that identifies the subgroup of interest. It is our belief that

general statements such as “youths” are much preferable to specific statements such as “12-23 year-olds.” We also recommend that interviewers are strongly coached to follow the proper protocol, and not to tell respondents about the eligible subgroup until eligible persons in the household are identified.

It is quite possible that NLSY97/PAY97 gave out too much information on the intended population, making it easier for household informants to avoid participation by “hiding” eligibles. From a data quality perspective, it is much better to identify a nonrespondent eligible than to miss an eligible altogether. Response rates will drop, of course, but coverage rates will increase. Adjustments and estimates can be made for nonrespondents, but not much can be done for missed eligibles.