# Appendix E
# Proposed Sample Design

# Sampling Design for the NPSAS:08 Full-Scale Study
# Collection of Information Employing Statistical Methods

This submission requests clearance for the 2008 National Postsecondary Student Aid Study (NPSAS:08). The sampling design for the full-scale study is presented below with contingencies upon the field test results where appropriate.

## E.1 Respondents Universe

### E.1.1 Institution Universe

The institutions eligible for NPSAS:08 are required during the 2007–2008 academic year to:

- offer an educational program designed for persons who have completed secondary education; and
- offer at least one academic, occupational, or vocational program of study lasting at least 3 months or 300 clock hours; and
- offer courses that are open to more than the employees or members of the company or group (e.g., union) that administers the institution; and
- be eligible to participate in Title IV programs; and
- be located in the 50 states, the District of Columbia, or Puerto Rico; and
- be other than a U.S. Service Academy.

Institutions providing only avocational, recreational, or remedial courses or only in-house courses for their own employees are excluded. U.S. Service Academies are excluded because of their unique funding/tuition base.

Consistency of this definition of the institution universe relative to previous NPSAS studies was discussed in Section B.1.a.

### E.1.2 Student Universe

The eligible students to be listed by the sample institutions for selection of the student sample for NPSAS:08 are those who will have attended a NPSAS-eligible institution at any time from July 1, 2007 through April 30, 2008 and who will be:

- enrolled in *either* (a) an academic program; (b) at least one course for credit that could be applied toward fulfilling the requirements for an academic degree; *or* (c) an occupational or vocational program that required at least 3 months or 300 clock hours of instruction to receive a degree, certificate, or other formal award; and
- not currently enrolled in high school; and
- not enrolled *solely* in a GED or other high school completion program.

## E.2 Statistical Methodology

### E.2.1 Institution Sample

The institution sampling frame for NPSAS:08 will be constructed from the 2005–06 Integrated Postsecondary Education Data System (IPEDS) header, Institutional Characteristics (IC) file, Fall Enrollment, and Completions files. The sample for NPSAS:08 will be selected prior to selection of the field test institutions. Then, the sample of field test institutions will be selected purposively from the complement of the full-scale sample institutions. This will ensure that no institutions are in both the field test and full-scale samples without affecting the representativeness of the full-scale sample.

Records on the IPEDS files that do not represent NPSAS-eligible institutions will be deleted, including those that represent central offices, U.S. service academies, or institutions located outside the U.S. and Puerto Rico.

The IPEDS files will then be "cleaned" to resolve the following types of problems:

- missing or zero enrollment or completions data, because these data are needed to compute measures of size for sample selection; and
- unusually large or small enrollments, especially if imputed, because if incorrect, these data would result in inappropriate probabilities of selection and sample allocation.

Table E-1 presents the proposed allocation of the NPSAS:08 institution sample to the 22 institutional sampling strata. The planned number of sample institutions is 1,667 with 1,374 institutions providing lists for sample selection.

We will select a direct, unclustered sample of institutions, like the sample selected for NPSAS:04, NPSAS:2000, and NPSAS:96, rather than a clustered sample like those used for previous NPSAS studies.

The NPSAS:08 student sampling design is based on fixed stratum sampling rates, not fixed stratum sample sizes, as discussed below. The student sampling rates are designed to produce about 120,000 sample students, distributed by institutional sector and student type as shown in table E-2: about 27,111 baccalaureate recipients; about 78,927 other undergraduate students; and about 13,962 graduate and first-professional students.

There will be seven student sampling strata:

1. potential baccalaureate recipients who are business majors;
2. potential baccalaureate recipients who are not business majors;
3. other undergraduate students;
4. masters students;
5. doctoral students;
6. other graduate students; and
7. first-professional students.

Potential baccalaureate recipients, other undergraduates, masters students, doctoral students, other graduate students, and first-professional students will be sampled at different rates

to control the sample allocation, as was done for NPSAS:2000 and NPSAS:04. Using different rates will allow us to obtain the target sample sizes and is a technique necessary in the full-scale study to meet analytic objectives for defined domain estimates.

**Table E-1. NPSAS:08 preliminary full-scale institution sample sizes and yield**

| Institutional sector | Frame count[1] | Number sampled | Number eligible | List respondents |
|---|---|---|---|---|
| Total | 6,610 | 1,667 | 1,646 | 1,374 |
| Public less-than-2-year | 245 | 21 | 342 | 292 |
| Public 2-year | 1,165 | 342 | 175 | 149 |
| Public 4-year non-doctoral | 357 | 175 | 289 | 249 |
| Public 4-year, doctoral | 289 | 289 | 20 | 18 |
| Private not-for-profit less-than-4-year | 321 | 20 | 312 | 256 |
| Private not-for-profit 4-year, non-doctoral | 1,010 | 324 | 249 | 193 |
| Private not-for-profit 4-year doctoral | 589 | 249 | 105 | 89 |
| Private for-profit less-than-2-year | 1,387 | 112 | 135 | 114 |
| Private for-profit 2-year or more | 1,247 | 135 | 1,646 | 1,374 |

NOTE: Detail may not sum to totals because of rounding. NPSAS:08 = 2008 National Postsecondary Student Aid Study.
[1] Institution counts based on IPEDS:2003–04 header file.

Based on past experience, we expect to obtain a minimum of 92 percent eligibility rates and 70 percent student interview response rates, overall and in each sector. We also plan to employ a variable-based (rather than source-based) definition of study respondent, similar to that used in NPSAS:04 with revisions as deemed necessary by the National Center for Education Statistics (NCES). We expect the study response rate to be about 90 percent. Approximately 99,051 student survey respondents, including 23,590 baccalaureate recipients; 63,287 other undergraduate students; and 12,175 graduate and first-professional students are expected.

The NPSAS sampling rates for students identified as potential baccalaureates by the sample institutions will be adjusted to yield the appropriate sample sizes after accounting for the percentage of students with a Baccalaureate and Beyond (B&B) flag of "yes" who actually receive a baccalaureate degree during the NPSAS year (about 87 percent, based on NPSAS:2000 data)[1].

[1] In NPSAS:2000, the baccalaureate recipients were identified by separate lists usually sent close to the end of the spring term, so the 87 percent estimate may need to be adjusted downwards to help determine the appropriate field test sampling rates.

**Table E-2. NPSAS:08 preliminary full-scale student sample sizes and yield**

| Institutional sector | Sample students | | | | Eligible students | | | | Study respondents | | | | Responding students per responding institution |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Total | Baccalaureates | Other undergraduate students | Graduate/first-professional students | Total | Baccalaureates | Other undergraduate students | Graduate/first-professional students | Total | Baccalaureates | Other undergraduate students | Graduate/first-professional students | |
| Total | 120,000 | 27,111 | 78,927 | 13,962 | 110,894 | 25,700 | 71,946 | 13,248 | 99,051 | 23,590 | 63,287 | 12,175 | 72.1 |
| Public less-than-2-year | 1,705 | 0 | 1,705 | 0 | 1,360 | 0 | 1,360 | 0 | 1,119 | 0 | 1,119 | 0 | 81.2 |
| Public 2-year | 24,931 | 0 | 24,931 | 0 | 21,912 | 0 | 21,912 | 0 | 17,414 | 0 | 17,414 | 0 | 59.6 |
| Public 4-year non-doctoral | 17,231 | 5,788 | 10,099 | 1,344 | 16,352 | 5,493 | 9,584 | 1,276 | 14,689 | 4,934 | 8,609 | 1,146 | 98.6 |
| Public 4-year doctoral | 35,635 | 12,231 | 16,799 | 6,605 | 33,892 | 11,633 | 15,978 | 6,281 | 31,042 | 10,655 | 14,634 | 5,753 | 124.5 |
| Private not-for-profit less-than-4-year | 1,539 | 0 | 1,539 | 0 | 1,369 | 0 | 1,369 | 0 | 1,262 | 0 | 1,262 | 0 | 70.9 |
| Private not-for-profit 4-year non-doctoral | 13,033 | 4,819 | 7,132 | 1,083 | 12,236 | 4,524 | 6,695 | 1,016 | 11,518 | 4,259 | 6,302 | 957 | 45.0 |
| Private not-for-profit 4-year doctoral | 13,661 | 4,039 | 4,948 | 4,674 | 12,971 | 3,835 | 4,699 | 4,438 | 11,982 | 3,543 | 4,340 | 4,099 | 61.9 |
| Private for-profit less-than-2-year | 7,049 | 0 | 7,049 | 0 | 6,003 | 0 | 6,003 | 0 | 5,569 | 0 | 5,569 | 0 | 62.9 |
| Private for-profit 2-year or more | 5,217 | 234 | 4,726 | 257 | 4,798 | 215 | 4,346 | 237 | 4,456 | 200 | 4,036 | 220 | 39.1 |

NOTE: NPSAS:08 = 2008 National Postsecondary Student Aid Study.

To develop the mathematical foundation for the institutional and student sampling design, we use the following notation to represent the institutional and student sampling strata:

r = 1, 2, ..., 22 indexes the institutional strata, and

s = 1, 2, ..., 7 indexes the student strata.

We further define the following notation:

j = 1, 2, ..., J(r) indexes the institutions that belong to institutional stratum "r,"

$M_{rs}(j)$ = number of students during the NPSAS year who belong to student stratum "s" at the j-th institution in stratum "r" based on the latest IPEDS data, and

$m_{rs}$ = number of students to be selected from student stratum "s" within the r-th institutional stratum, per table E-2 for students, referred to henceforth as student stratum "rs."

The overall population sampling rate for student stratum "rs" is then given by

where

$$f_{rs} = m_{rs} / M_{rs}(+,+) \ ,$$

$$M_{rs}(+,+) = \sum_{i=1}^{I} \sum_{j=1}^{J(r,i)} M_{rs}(i,j) \ .$$

The student sampling rates, $f_{rs}$, will be computed based on the final sample allocation and IPEDS data regarding the population sizes.

The composite measure of size for the j-th institution in stratum "r" will then be defined as

$$S_r(j) = \sum_{s=1}^{11} f_{rs} M_{rs}(j) \ ,$$

which is the number of students that would be selected from the j-th institution if all institutions on the frame were to be sampled.

An independent sample of institutions will be selected for each institutional stratum using Chromy's sequential pmr sampling algorithm to select institutions with probabilities proportional to their measures of size (Chromy, 1979).[2] However, rather than allow multiple selections of sample institutions, we will select with certainty those institutions with expected frequencies of selection greater than unity (1.00), and we will select the remainder of the institutional sample from the remaining institutions in each stratum. This process makes it unnecessary to select

[2] Chromy, J.R. (1979). "Sequential Sample Selection Methods." Proceedings of the *American Statistical Association Section on Survey Research Methods*, pp. 401-406.

multiple second-stage samples of students by precluding institutions with multiple selections at the first stage of sampling. Therefore, the expected frequency of selection for the j-th institution in institutional stratum "r" is given by

$$S_r(+) = \sum_{j=1}^{J(r)} S_r(j),$$

where

$$\pi_r(j) = \begin{cases} \dfrac{n_r S_r(j)}{S_r(+)}, & \text{for non-certainty selections;} \\ 1, & \text{for certainty selections;} \end{cases}$$

and $n_r$ is the number of non-certainty selections from stratum "r."

Within each of the "r" institutional strata, we will stratify implicitly by sorting the stratum "r" sampling frame in a serpentine manner (see Williams and Chromy, 1980)[3] by the following variables:

(1) Historically Black Colleges and Universities (HBCU) indicator;

(2) Hispanic Serving Institutions (HSI) indicator;

(3) Carnegie classifications of postsecondary institutions;[4]

(4) the Office of Business Economics (OBE) Region from the IPEDS header file (Bureau of Economic Analysis of the U.S. Department of Commerce Region);[5] and

(5) the institution measure of size.

Further implicit stratification within region by state and system will be done for three states with large systems: the SUNY and CUNY systems in New York; the state and technical colleges in Georgia; and the California State and University of California systems. The objective of this implicit stratification will be to approximate proportional representation of institutions on these measures. Additionally, for-profit 2-year or more institutions will be implicitly stratified by 2-year and 4-year institutions.

### E.2.2 Student Sample

Many aspects of the procedures for obtaining and sampling from student lists were described for the field test, including:

[3]Williams, R.L. and J.R. Chromy (1980). "SAS Sample Selection MACROS." Proceedings of the *Fifth Annual SAS Users Group International Conference*, pp. 392-396.

[4] We will use the new Carnegie codes and decide what, if any, collapsing is needed of the categories for the purposes of implicit stratification.

[5] For sorting purposes, Alaska and Hawaii will be combined with Puerto Rico in the Outlying Areas region rather than in the Far West region.

- obtaining as many lists via uploads to the project website or zipped, password protected e-mail attachments;
- processing lists on a flow basis as they are received;
- implementing quality assurance checks against IPEDS data; and
- compiling a master sample file on a flow basis as sample students are selected, including student selection probabilities.

The procedures proposed for the field test will be refined based on the results of the field test and then implemented for the full-scale study.

Student samples will be selected as stratified, systematic random samples for both paper and electronic lists primarily because of its ease of implementation with paper lists. The student sampling rates will be fixed for each sample institution, rather than the student sample sizes:

- to facilitate selecting the samples on a flow basis as the student lists are received from sample institutions; and
- because sampling at a fixed rate based on the overall stratum sampling rate and the institution probabilities of selection results in approximately equal overall probabilities of selection within student strata.

The overall population sampling rate for student stratum "rs" is given by

where

$$f_{rs} = m_{rs} / M_{rs}(+) \ ,$$

$$M_{rs}(+) = \sum_{j=1}^{J(r)} M_{rs}(j) \ .$$

For the unconditional probability of selection to be a constant for all eligible students in stratum "rs," the overall probability of selection should be the overall student sampling fraction, $f_{rs}$; that is to say, we must ensure that

$$\frac{m_{rs}(j)}{M_{rs}(j)} \pi_r(j) = f_{rs} \ ,$$

or equivalently,

$$m_{rs}(j) = f_{rs} \frac{M_{rs}(j)}{\pi_r(j)} \ .$$

Thus, the conditional sampling rate for stratum "rs," given selection of the j-th institution, becomes

$$f_{rs|j} = f_{rs} / \pi_r(j) \ .$$

However, in this case, the desired overall student sample size, $m_s$, is achieved only *in expectation* over all possible samples.

Achieving the desired sample sizes with equal probabilities within strata in the particular sample that has been selected and simultaneously adjusting for institutional nonresponse and ineligibility requires that

$$\sum_{j \varepsilon R} m_{rs}\ (j) = m_{rs}\ ,$$

where "R" denotes the set of eligible, responding institutions. If we let the conditional student sampling rate for stratum "rs" in the j-th institution be

$$\hat{f}_{rs|j} = \hat{f}_{rs} / \pi_r\ (j)\ ,$$

we then require

$$\sum_{i \varepsilon R} \hat{f}_{rs} \frac{M_{rs}\ (j)}{\pi_r\ (j)} = m_{rs}\ ,$$

or equivalently,

$$\hat{f}_{rs} = m_{rs} / \hat{M}_{rs}\ ,$$

where

$$\hat{M}_{rs} = \frac{\sum_{j \varepsilon R} M_{rs}\ (j)}{\pi_r\ (j)}\ .$$

Because it will be necessary to set the student sampling rates before we have complete information on eligibility and response status, $\hat{M}_{rs}$ will be calculated as follows:

$$\hat{M}_{rs} = \sum_{j \varepsilon S} \frac{M_{rs}\ (j)}{\pi_r\ (j)} * [\ E_r\ R_r\ E_{rs}\ ]\ ,$$

where "S" denotes the set of all sample institutions,

$E_r$ = the institutional eligibility factor for institutional stratum "r,"

$R_r$ = the institutional response factor for institutional stratum "r,"

$E_{rs}$ = the student eligibility factor for student stratum "rs."

NPSAS is a multivariate survey with a *p*-dimensional parameter space, $\boldsymbol{\theta} = \{\theta_j\}$, j = 1, ….., *p*, for which it is desired to estimate $\boldsymbol{\theta}$ with $\hat{\boldsymbol{\theta}}$ while minimizing cost (sample size) subject to a series of precision requirements. Consequently, optimal sampling rates can be obtained by solving the following nonlinear optimization problem:

Minimize: $$C = C_0 + \sum_{i=1}^{I} \left( C_{1i} n_{1i} + \sum_{f=1}^{F} C_{2if} n_{2if} \right)$$

$$\text{Subject to:} \quad \begin{cases} V\left(\hat{\theta}_j\right) \le v_j, \forall j \\ 2 \le n_{1i} \le N_{1i}, i \in [1,I] \\ 2 \le n_{2if} \le N_{2if}, f \in [1,F] \end{cases}$$

Where,

$C_0$ = fixed cost not affected by changes in the numbers of institutions or students selected;

$C_{1i}$ = variable cost per institution, depending on the number of participating institutions in the $i^{th}$ institutional stratum;

$n_{1i}$ = number of participating institutions in the $i^{th}$ stratum;

$C_{2if}$ = variable cost per student, depending on the number of participating students in the $f^{th}$ student stratum within the $i^{th}$ institutional stratum; and

$n_{2if}$ = number of participating students in the $f^{th}$ student stratum within the $i^{th}$ institutional stratum.

In the above, variance constraints $V(\hat{\theta}_j) \le v_j$ correspond to precision requirements that have been specified by NCES for key survey estimates. Using data from the NPSAS:04 and NPSAS:2000, all of the required variance components and their associated precision constraints have been developed. Subsequently, the resulting nonlinear optimization problem to determine the most effective sample allocation will be solved using Chromy's algorithm[6] to obtain feasible solutions to the above problem.

The large sample sizes proposed for NPSAS:08 are required to achieve reliable precision expected by users of NPSAS data. A baseline cohort of baccalaureate recipients must be selected for the B&B studies. Moreover, many NPSAS:08 statistical analyses will focus on relatively rare domains, thereby requiring large overall sample sizes and disparate sampling rates. Discussions with NCES have been used to identify the domains of interest and the study will be designed to ensure adequate sample sizes for those domains.

[6] Chromy, J.R. (1987). "Design Optimization with Multiple Objectives." *Proceedings of the American Statistical Association, Section on Survey Research Methods*.