Appendix E
Sample Design for the NPSAS:04 Full-scale Study

Appendix E
Sample Design for the NPSAS:04 Full-scale Study

E.1. Respondents Universe

E.1.1 Institution Universe

The institutions eligible for NPSAS:04 were required during the 2003–04 academic year to

- offer an educational program designed for persons who have completed secondary education;
- offer at least one academic, occupational, or vocational program of study lasting at least 3 months or 300 clock hours;
- offer courses that are open to more than the employees or members of the company or group (e.g., union) that administers the institution;
- be eligible to participate in Title IV programs;
- 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 were excluded. U.S. Service Academies were excluded because of their unique funding/tuition base.

Consistency of this definition of the institution universe relative to previous NPSAS studies is 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:04 are those who attended a NPSAS-eligible institution at any time from July 1, 2003 through April 30, 2004 and who were:

- 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.

Students concurrently enrolled in high school or who were enrolled only in a GED or other high school completion program were not eligible. Students taking only courses for remedial or avocational purposes and not receiving credit, those only auditing courses, and those taking courses only for leisure, rather than as part of an academic, occupational, or vocational program or course of instruction, were not eligible.

E.2. Statistical Methodology

E.2.1 Institution Sample

The institutional sampling frame for NPSAS:04 was constructed from the 2001 Integrated Postsecondary Education Data System (IPEDS) Institutional Characteristics (IC) file, the 2001 IPEDS Completions file, and the 2001 IPEDS Fall Enrollment file. The sample for NPSAS:04 was selected prior to selection of the field test institutions. Then, the sample of field test institutions was selected purposively from the complement of the full-scale sample institutions. This ensured that no institutions were in both the field test and full-scale samples without affecting the representativeness of the full-scale sample.

Records on the IPEDS IC file that did not represent NPSAS-eligible institutions were deleted. Hence, records that represented central offices, U.S. service academies, or institutions located outside the United States and Puerto Rico were deleted. The IPEDS files were then "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 enrollment, especially if imputed, because, if incorrect, these data would result in inappropriate probabilities of selection and sample allocation.

Table E-1 presents the allocation of the NPSAS:04 institutional sample to the nine institutional sampling strata. The number of sample institutions is 1,500, accounting for historical rates of participation in CADE, institution eligibility rates, and rates with which sample institutions provide student lists for sample selection. Table E-1 shows the resulting institutional sample sizes, which was 1,370 institutions providing lists for sample selection and 1,285 institutions providing CADE data.

We selected a direct, unclustered sample of institutions, like the sample selected for NPSAS:2000 and NPSAS:96, rather than a clustered sample like those used for previous NPSAS studies. A subset of approximately 1,000 institutions selected for NPSAS was also in the 2004 National Study of Postsecondary Faculty (NSOPF:04) sample. In addition, to allow analysis of the effects of state tuition and student aid policies in individual states, representative samples of institutions were selected from three strata—public 2-year institutions; public 4-year institutions; and private not-for-profit 4-year institutions—in each of the following 12 states: CA, CT, DE, GA, IL, IN, MN, NE, NY, OR, TN, and TX.

The NPSAS:04 student sampling design was based on fixed stratum sampling rates, not fixed stratum sample sizes, as discussed below. The student sampling rates were designed to produce about 80,925 student web/CATI respondents, distributed by institutional and student sampling strata as shown in table E-2: about 22,091 first-time beginner (FTB) students; about 45,401 other undergraduate students; and about 13,433 graduate and first-professional students.

There were two student sampling strata for undergraduates (FTB and other undergraduates), three student sampling strata for graduate students (master's, doctoral, and other graduate students), and one stratum for first-professional students. Differential sampling rates were used for the three types of graduate students to get adequate representation of students

pursuing doctoral degrees and to limit the sample size for "other" graduate students, who are of limited inferential interest.

Table E-1. NPSAS:04 institution sample sizes and yield

	Institutions				
Institutional sector	Frame	Sample	Eligible	List respondents	CADE respondents
Total	6,674	1,500	1,483	1,370	1,285
Public less-than-2-year	321	50	48	41	37
Public 2-year	1,225	322	319	303	288
Public 4-year nondoctorate granting	358	150	150	143	136
Public 4-year doctorate granting	276	251	251	238	226
Private not-for-profit 2-year or less	379	60	55	52	48
Private not-for-profit, 4-year nondoctorate granting	1,076	252	249	212	195
Private not-for-profit 4-year doctorate granting	537	165	165	155	147
Private for-profit less-than-2-year	1,390	150	146	131	118
Private for-profit 2-year or more	1,112	100	100	95	90

NOTE: Institution counts based on the Fall 2000 IPEDS data collection. Institution eligibility rate: 98.9 percent. Institution list response rate: 92.4 percent. 1,000 of the 1,500 institutions also are in the NSOPF:2004 sample.
SOURCE: U.S. Department of Education, National Center for Education Statistics, 2004 National Postsecondary Student Aid Study, "Field Test Methodology Report" (NPSAS:04).

The NPSAS:04 web and CATI data collection procedures were expected to produce about a 70 percent student response rate based on historical experience. Given prior NPSAS experience regarding institutional CADE response rates and sample student eligibility rates, the student sample sizes planned to support the desired student web/CATI yield are shown in table E-2. We selected approximately 121,684 sample students for NPSAS:04, including 36,228 FTBs; 67,596 other undergraduate students; and 17,860 graduate and first-professional students.

The numbers of FTB students shown in table E-2 include both "true" FTBs who began their postsecondary education for the first time during the NPSAS field test year and effective FTBs who had not completed a postsecondary class prior to the NPSAS field test year. Unfortunately, postsecondary institutions cannot readily identify their FTB students. Therefore, the NPSAS sampling rates for students identified as FTBs and other undergraduate students by the sample institutions were adjusted to yield the sample sizes shown in table E-2 after accounting for expected false positive and false negative rates. The false-positive and false-negative FTB rates experienced in NPSAS:96 were used to set appropriate sampling rates for the NPSAS:04 field test.[1]

[1] The NPSAS:96 false-positive rate was 27.6 percent for students identified as potential FTBs by the sample institutions, and the false-negative rate was 9.1 percent for those identified as other undergraduate students.

Table E-2. NPSAS:04 student sample sizes and yield

	Web/CATI respondents				Eligible students				Sample student			
Institutional sector	Total	BPS	Other UG	G1P	Total	BPS	Other UG	G1P	Total	BPS	Other UG	G1P
Total	80,925	22,091	45,401	13,433	114,738	33,033	63,845	17,860	121,684	36,228	67,596	17,860
Public less-than-2-year	1,442	650	792	#	2,218	1,000	1,218	#	2,773	1,250	1,523	#
Public 2-year	14,410	7,096	7,314	#	22,169	10,917	11,252	#	24,632	12,130	12,502	#
Public 4-year nondoctorate granting	11,152	2,157	7,645	1,350	15,022	2,915	10,331	1,776	15,719	3,068	10,875	1,776
Public 4-year doctorate granting	23,545	2,882	14,730	5,933	31,607	3,895	19,905	7,807	32,092	3,974	20,311	7,807
Private not-for-profit 2-year or less	2,147	1,265	882	#	3,303	1,946	1,357	#	3,476	2,048	1,428	#
Private not-for-profit 4-year nondoctorate granting	8,898	1,646	6,206	1,046	12,005	2,224	8,386	1,395	12,563	2,341	8,827	1,395
Private not-for-profit 4-year doctorate granting	9,945	1,042	4,601	4,302	13,362	1,408	6,218	5,736	13,518	1,437	6,345	5,736
Private for-profit less-than-2-year	5,459	3,840	1,619	#	9,098	6,400	2,698	#	10,703	7,529	3,174	#
Private for-profit 2-year or more	3,927	1,513	1,612	802	5,954	2,328	2,480	1,146	6,208	2,451	2,611	1,146

Rounds to zero.

NOTE: Student eligibility rate: 94.3 percent. Student response rate: 70.5 percent. BPS = Confirmed first-time beginners (design will account for false positive and false negative FTB rates to yield these sample sizes)

SOURCE: U.S. Department of Education, National Center for Education Statistics, 2004 National Postsecondary Student Aid Study, “Field Test Methodology Report” (NPSAS:04).

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

r = 1, 2,..., 58 indexes the institutional strata, and

s = 1, 2,..., 11 indexes the student/faculty strata.

Note that the NSOPF sample of institutions was a subset of the NPSAS institutions, so the institution strata were expanded to accommodate the selection of certain types of institutions for NSOPF. The strata also accounted for selection of institutions in the 12 states where there were representative samples. The institution measure of size (described below) accounted for student as well as for faculty counts and sampling rates.

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 and faculty during the NPSAS year who belong to person stratum "s" at the j-th institution in stratum "r" based on the latest IPEDS data, and

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

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

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

where

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

The person sampling rates, f_{rs}, were 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 persons that would be selected from the j-th institution if all institutions on the frame were to be sampled.

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

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

remaining institutions in each stratum. This process made it unnecessary to select multiple second-stage samples of persons 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 stratified implicitly by sorting the stratum "r" sampling frame in a serpentine manner (see Williams and Chromy, 1980[3]) by the following variables:

- HBCU (historically black colleges and universities);
- OBE Region (from the IPEDS IC file) with Alaska and Hawaii moved to Region 9 with Puerto Rico;
- state; and
- the institution measure of size.

The objectives of this additional, implicit stratification are to ensure some HBCUs, to ensure proportionate representation of all geographic regions and states, and to ensure representation of both large and small 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

- obtaining as many lists as possible in machine-readable form, including e-mails, uploads to the project website, and diskettes or CD-ROMs;
- processing lists on a flow basis as they are received;
- unduplicating samples selected when an institution provides only a hard-copy list for each term of enrollment;
- ensuring that each sample institution receives a sufficient sample allocation that 30 respondents can be expected;
- implementing quality assurance checks against the latest IPEDS data; and

[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.

- compiling a master sample file on a flow basis as sample students are selected, including student and institution sampling weight factors.

The procedures proposed for the field test were refined based on the results of the field test and implemented for the full-scale survey.

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

- to facilitate selecting the samples on a flow basis as the student lists were received from sample institutions;
- to facilitate unduplicating the samples selected when an institution provided only hard-copy lists by term; 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.

Recall that the overall population sampling rate for student stratum "rs" is given by

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

where

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

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}; i.e., 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 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

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

or equivalently,

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

where

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

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

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

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, $\boldsymbol{\theta} = \{\theta_j\}$, $j = 1, \ldots\ldots, p$, 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: $C = C_0 + \sum_{i=1}^{I} \left(C_{1i} n_{1i} + \sum_{f=1}^{F} C_{2if} n_{2if} \right)$

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

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 $V(\hat{\theta}_j) \leq v_j$ correspond to precision requirements that have been specified by NCES for key survey estimates. Using data from the NPSAS:2000 and NPSAS:96 (and NSOPF:99 for faculty constraints), 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 was solved using Chromy's algorithm[4] to obtain feasible solutions to the above problem.

The large sample sizes proposed for NPSAS:04 were required to achieve the many objectives of the study, including estimates for three domains—public 2-year, public 4-year, and private not-for-profit 4-year institutions—in each of 12 states. A baseline cohort of FTBs must be selected for the BPS studies. Moreover, many NPSAS:04 statistical analyses 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.

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