Chapter 5
Variable Construction and File Development

5.1 Overview of the NPSAS Files

The NPSAS:2000 data files contain student-level and institution-level data collected from institution records, government databases, admission test vendors, and student interviews. The primary analysis file, from which the study Data Analysis Systems (DASs) were constructed, contains data for about 62,000 students—50,000 undergraduates, 11,000 graduate students, and 1,200 first-professional students. Among the undergraduates, about 10,400 were confirmed to have received their baccalaureate degrees between July 1, 1999, and June 30, 2000.

The primary analysis file contains over 1,000 variables, most of which were derived from multiple NPSAS:2000 data sources. The NPSAS:2000 data sources, along with the corresponding numbers of study respondents for which data were obtained, appear in table 5-1. Additional students for whom data were obtained through database matching who do not appear on the analysis file, and therefore are not represented in the table (due to incomplete data).

Table 5-1.—Record counts from NPSAS:2000 data sources, by student type

Data source	Total [1]	B&B	Total undergraduate	Graduate/first-professional
CADE (institution records)[2]	59,280	9,940	48,010	11,280
CATI (student records)	44,490	10,400	35,540	8,960
CPS 1999–2000 (Central Processing System)	31,500	5,930	27,790	3,710
CPS 2000–2001 (Central Processing System)	18,330	1,530	16,030	2,300
NSLDS Pell grants (any year)	21,430	4,010	19,750	1,680
NSLDS loans (any year)	34,090	6,830	27,360	6,730
NSLDS Pell grants (NPSAS year)	13,550	2,430	13,490	60
NSLDS loans (NPSAS year)	21,410	4,650	18,140	3,270
ACT (years 1991–92 through 1999–2000)	16,540	5,340	10,070	1,130
SAT (years 1995 through 1999)	14,680	3,880	14,330	350

[1] The numbers presented here are limited to study respondents.
[2] The CADE data file contains all study respondents, which includes some CADE nonrespondents.

NOTE: To protect confidentiality, some numbers have been rounded.

SOURCE: U.S. Department of Education, National Center for Education Statistics, National Postsecondary Student Aid Study, 1999–2000 (NPSAS:2000).

Throughout the data collection period, data were processed and examined for quality control purposes. Editing of student data began shortly after the start of CATI data collection. Anomalous values were investigated and resolved if necessary. As shown in table 5-2, numerous *interim* files were delivered to NCES for review, with each delivery including more of the study data.

Table 5-2.—Interim file deliveries

Date	Description
06/26/2000	840 completed interviews delivery – CATI, CADE, and CPS
07/31/2000	5,000 completed interviews delivery – CATI, CADE, and CPS
12/15/2000	30,000 completed interviews delivery – CATI, CADE, and CPS
01/25/2001	Preliminary Analysis file #1 – File containing CATI, CADE, CPS, preliminary weights, derived demographic variables, and derived financial aid data
02/20/2001	Preliminary Analysis file #2 – File containing CATI, CADE, CPS, institution data, near-final weights, NSLDS loan data, NSLDS Pell Grant data, derived demographic variables, and derived financial aid data

SOURCE: U.S. Department of Education, National Center for Education Statistics, National Postsecondary Student Aid Study, 1999–2000 (NPSAS:2000).

Following completion of all study data collection, separate Data Analysis System files were created for undergraduate and graduate/first-professional students. The first study DAS, for undergraduate students only, was adjudicated and approved for public release in July 2001.

Complete data obtained through the NPSAS:2000 are available on restricted CD files and documented by the electronic codebook (ECB). These files and the ECB are available to researchers who have applied for and received authorization from NCES to access restricted research files. The NPSAS:2000 ECB contains information about the following files (to protect confidentiality, some numbers have been rounded):

- NPSAS Analysis File – Contains analytic variables derived from all NPSAS data sources as well as selected direct CATI variables for the 62,000 study respondents.
- CADE Data File – Contains raw data collected from institutional records for the 59,284 students with sufficient data to be considered CADE respondents, but also includes study respondents not considered CADE respondents. This file excludes any CADE "verbatim" variables such as responses to "Other, specify" items. These variables appear on the separate Verbatim Data File.
- CATI Student Data File – Contains student-level raw data collected from 44,500 students who responded to the student interview. This file excludes any CATI "verbatim" variables, which are on the Verbatim Data File.
- CATI School Data File – Contains institution data obtained from the student interview. It is a student-level file; however, a student can have more than one record in the file. There is a separate record for each postsecondary institution students

reported in CATI as somewhere they had attended during the study year (for up to 5 institutions).

- Institution File – Contains selected institution-level variables for the nearly 1,100 sampled institutions. Of those institutions, about 1,000 participated in NPSAS:2000. This file can be linked to the CATI Student Data File and CADE Data File by the IPEDS number.
- Coding Results File – Contains the verbatim text and resulting code for student major and (for employed students) industry and occupation. In addition, it contains the occupation code and corresponding verbatim text for any parent data obtained in CATI. This file also includes the field-of-study text string collected in CADE, along with the resulting code. Linkage to other data files is through the student ID.
- Verbatim Data File – Contains item-level records (i.e., one record per variable) for text variables collected in either CADE or CATI. It is possible to have multiple records per student or no records for a student.
- CATI Preload File – Contains the data preloaded into the student interview for the 44,500 CATI respondents.
- CPS 1999–2000 Data File – Contains data received from the Central Processing System for the 31,500 study respondents who matched to the 1999–2000 financial aid application files.
- CPS 2000–2001 Data File – Contains data received from the Central Processing System for the approximately 18,300 study respondents who matched to the 2000–2001 financial aid application files.
- NSLDS Pell Data File – Contains raw grant-level data received from the National Student Loan Data System for the 21,400 study respondents who received Pell Grants during the NPSAS year or prior years. This is a history file with separate records for each transaction in the Pell system.
- NSLDS Loans Data File – Contains raw loan-level data received from the National Student Loan Data System for the 34,100 study respondents who received loans during the NPSAS year or prior years. This is a history file with separate records for each transaction in the loan files.
- SAT Data File – Contains SAT data for the 14,700 study respondents who matched to the ETS SAT database for the 1995–1999 test years.
- ACT Data File – Contains ACT data for the 16,500 study respondents who matched to the ACT database for 1991–1992 through 1999–2000.
- Weights File – Contains all the sampling and analysis weights created for NPSAS:2000. There is a separate record for each study respondent.

5.2 Data Coding and Editing

The NPSAS:2000 data were coded and edited using procedures developed and implemented for previous NCES-sponsored studies. These coding and editing procedures were implemented for the NPSAS:2000 field test, and refined during the processing of NPSAS:2000 full-scale data.

The coding and editing procedures fell into two categories:

1. Online coding and editing performed during data collection, and
2. Post-data-collection data editing.

5.2.1 Online Coding and Editing

NPSAS:2000 included two major data collection systems: CADE and CATI. Both systems included edit checks to ensure data collected were within valid ranges. To the extent feasible, both systems incorporated across-item consistency edits. While more extensive consistency checks would have been technically possible, use of such edits was limited in order to prevent excessive interview and/or respondent burden.

The CATI system included online coding systems used for the collection of industry, occupation, and major field-of-study data. Additionally, the CATI system included a coding module used to obtain IPEDS information for postsecondary institutions that the student attended (other than the NPSAS institution from which they were sampled).

Below is a description of the online range and consistency checks, and the online coding systems, incorporated into the NPSAS:2000 CADE and CATI systems.

NPSAS:2000 CADE

- All fields in CADE accepted a code of –1, for the user to indicate the information was not available in the institution records.
- All state fields were checked against a master listing of 2-character state and country codes. Nonvalid entries were prohibited by the system.
- Phone numbers left blank triggered a warning to the user requesting that the information be provided. If the phone number was again left blank, it was automatically filled with –1 (data not available).
- Student date of birth entered by a CADE user was compared to values previously obtained from the Central Processing System. If the CPS date of birth was nonblank, but different from the value entered, a warning was issued and the user was asked to either keep the date of birth as entered or accept the CPS value.

- High school graduation year was compared to CADE date of birth. If student age at the time of high school graduation was calculated as 15 or younger, a warning asked the user to verify the high school graduation date.

- Student citizenship status entered by CADE users was compared to the value previously obtained from the CPS. If the CPS citizenship was nonblank, but different from the value entered, a warning was issued and the user was asked to either keep the value as entered or accept the CPS value.

- The student's military veteran status entered by CADE users was compared to the value previously obtained from the CPS. If the CPS veteran status was nonblank, but different from the value entered, a warning was issued and the user was asked to either keep the value as entered or accept the CPS value.

- Admissions test scores were collected for SAT, ACT, GRE, GMAT, MCAT, and LSAT. Soft-edit range checks were performed on all admissions test score variables.

- Values for credit hours enrolled that were outside of the normal range (according to the student's attendance status) triggered a CADE alert to the user. The user could keep the value of credit hours entered or change it.

- If the student was sampled as an undergraduate and was identified in CADE as being enrolled in a graduate or first-professional degree program, then the user received a warning. Similarly, if the student was sampled as a graduate student and the CADE user indicated the student was enrolled in an undergraduate degree program, the user also received a warning. The user had the option to keep the entered value or modify it.

- If the user selected a graduate or first-professional degree program but the institution was coded as having no graduate or first-professional levels based on IPEDS data and information from the Institutional Coordinator, the user received a warning. The user had the option to keep the entered value or modify it.

- If the user selected an undergraduate degree program but a graduate student level, an alert was issued. Similarly, if the user selected a graduate/first-professional program and an undergraduate student level, a warning appeared. In either case, the user could choose to modify the degree program or student level, or retain the entries as keyed.

- Grade-point average (GPA) entered for the student was compared to the GPA scale for the institution (previously obtained from the Institutional Coordinator). Incompatible score/scale combinations triggered a warning to the user. The user could accept what was entered or change it.

- If tuition for a specific term of enrollment was zero or less, or $15,000 or more, a warning message was triggered asking for verification from the user.

- If total tuition for the NPSAS year was $30,000 or higher, a warning message was triggered asking for verification from the user.

- Range checks were included on all financial aid award variables, with minimum and maximum values established based on published ranges in federal, state, or institution records.

- Graduate financial aid (e.g., a graduate assistantship) entered for a student sampled as an undergraduate triggered a warning message.

- If the CADE user indicated that the student received financial aid, but the total aid amount was $0, a warning was triggered. Total financial aid in excess of $30,000 for the NPSAS year also triggered a warning.

- Total cost of attendance budget (including tuition, housing, books, and technology) in excess of $25,000 triggered a warning to the user.

NPSAS:2000 CATI

- Range checks were applied to all numerical entries, such that only valid responses could be entered.
- Major field of study was entered by telephone interviewers as a text string. The coding software then standardized and analyzed the text and attempted to match the entry to a database. The interviewer was presented with one or more choices from which to select the appropriate entry in the coding dictionary, confirming entry with the student when multiple choices were presented.

- Student's occupation (if the student was employed) was coded by concatenating text strings entered for job title and job duties. The coding software then standardized and analyzed the text and attempted to match the entry to a database. The interviewer was presented with one or more choices from which to select the appropriate entry in the coding dictionary, confirming entry with the student when multiple choices were presented.

- Student's industry (if the student was employed) was entered as a text string. The coding software then standardized and analyzed the text and attempted to match the entry to a database. The interviewer was presented with one or more choices from which to select the appropriate entry in the coding dictionary, confirming entry with the student when multiple choices were presented.

- The postsecondary institution (other than the NPSAS institution) in which the student was enrolled during the NPSAS year was selected from a list, based on the respondent's report and the interviewer's entry of the city and state in which the institution was located. Upon selection, the name of the institution, as well as selected IPEDS variables (institutional level, control, tuition) was inserted into the CATI database.

- A verification check was triggered if date of attendance and date of degree completion were in conflict.
- A verification check was triggered if the highest expected degree attainment from the NPSAS target institution was in conflict with the highest level of offering at that institution.
- A verification check was triggered if employer aid exceeded $50,000.
- A verification check was triggered if parental support (beyond tuition, fees, housing, books, etc.) exceeded $35,000.
- A verification check was triggered if hours worked per week while enrolled exceeded 60 hours.
- A verification check was triggered if earnings and income exceeded $1,000,000.
- A verification check was triggered if age at time of high school completion (as calculated based on date of birth and date entered) was 15 or younger or 24 or older.
- A verification check was triggered if age of parent was 100 or higher.

5.2.2 Post-Data-Collection Editing

Following data collection, the information collected in CADE and CATI was subjected to various checks and examinations. These checks were intended to confirm that the database reflected appropriate skip-pattern relationships, and also to insert special codes in the database to reflect the different types of missing data. There are a variety of explanations for missing data within individual data elements. For example, an item may not have been applicable to certain students, a respondent may have refused to answer a particular item, or a respondent may not have known the answer to the question. Table 5-3 lists the set of special codes used to assist analysts in understanding the nature of missing data associated with NPSAS:2000 data elements.

Table 5-3.—Description of missing data codes

Missing data code	Description
–1	Don't know (CATI variables) Data not available (CADE variables)
–2	Refused (CATI variables only)
–3	Legitimate skip (item was intentionally not collected because variable was not applicable to this student—CADE and CATI variables only)
–6	Bad data, out of range
–7	Item was not reached (abbreviated and partial CATI interviews)
–8	Item was not reached due to a CATI error

SOURCE: U.S. Department of Education, National Center for Education Statistics, National Postsecondary Student Aid Study, 1999–2000 (NPSAS:2000).

In some instances, additional across-item consistency checks were performed, although such checks were kept to a minimum since, without recontacting respondents, it was difficult to know which data item was the true source of the inconsistency.

Skip-pattern relationships in the database were examined by methodically running cross-tabulations between gate items and their associated nested items. In many instances, gate-nest relationships had multiple levels within the CADE or CATI instrument. That is, items nested within a gate question may themselves have been gate items for additional items. Therefore, validating the gate-nest relationships often required much iteration and many multiway cross-tabulations.

The data cleaning and editing process for the NPSAS data consisted of the following steps.

Step 1. Replaced blank or missing data with –9 for all variables in the CADE or CATI database. Ran one-way frequency listing of every variable in the database to confirm no missing or blank values remained. These same one-way frequencies revealed any out-of-range or outlier data values, which were investigated and checked for reasonableness against other data values. Example: hourly wages of .10, rather than 10.

Some standard variable recodes were performed during this step. All Yes/No CATI variables were recoded from 1=Yes/2=No to 1=Yes/0=No. RTI's Telephone Survey Department standard is to use 1 for Yes and 2 for No. However, 1/0 for Yes/No is more appropriate in the DAS and ECB.

Step 2. Using CADE or CATI source code as specifications, defined all gate-nest relationships in SAS code. Format of SAS statement should have been:

IF gate-variable EQUAL gate-value AND nest-variable EQUAL –9 THEN nest-variable EQUAL –3.

This code replaced –9's with –3's (the legitimate skip code) as appropriate. Two-way cross-tabulations between each gate-nest combination revealed either unusually high numbers of nonreplaced –9 codes, or unusually high numbers of "valid" responses in items that should have been skipped. Each such instance was investigated to ensure skip-pattern integrity. Typically, resolution involved reprogramming the gate-nest relationship to be consistent with the CADE or CATI instrument. Occasionally, this check revealed errors in the CADE or CATI source code.

Some logical imputations could occur during this step if nonnegative values were assigned to variables that were "missing" and whose values could have been implicitly determined (and were thereby skipped in CADE or CATI). For instance, if the student did not work while enrolled, then the amount earned should have been coded to $0 rather than –3 or –9. If a student indicated he or she was not disabled, then the "nested" disability items under the gate question were logically imputed to "no."

Step 3. Based on the section completion indicators, and/or the abbreviated interview indicator, replaced –9 and –3 with –7 (item not administered). This code, which was used for the first time in BPS:96/98, allows analysts to easily distinguish those items that were not administered to the respondent due to a partial interview or abbreviated interview versus items that were either skipped or left blank unintentionally.

Step 4. Regenerated and examined one-way frequencies on all categorical variables. Investigated high counts of –9. Checked new frequencies for out-of-range or outlier data items. Confirmed that responses in the one-way frequencies had corresponding entries in the VALCODES documentation file. Replaced any remaining –9 codes with the appropriate missing data code.

Step 5. Produced descriptive statistics for all continuous variables using SAS PROC UNIVARIATE. The SAS program first temporarily recoded all values less than zero (–1, –2, –3, –7, –8) to missing. Examined minimum, median, maximum, and mean to assess reasonableness of responses. Investigated anomalous data patterns and corrected as necessary.

5.3 Composite and Derived Variable Construction

Analytic variables were created by examining the data available for each student from the various data sources, establishing relative priorities of the data sources—on an item-by-item basis—and reconciling discrepancies within and between sources. In some cases the derived or composite variables were created by simply assigning a value from the available source of information given the highest priority. In other cases, raw interview items were recoded or otherwise summarized to create a derived variable. A listing of the set of analysis variables derived for NPSAS:2000 appears in appendix J. Specific details regarding the creation of each variable appear in the variable descriptions contained in the ECB and DAS.

5.4 Statistical Imputations

After the editing process (which included logical imputations), the remaining missing values for 23 analysis variables were imputed statistically. The imputations were performed primarily to reduce the bias of survey estimates caused by missing data. The imputed data also made the data complete and easier to analyze. Most of the variables were imputed using a weighted hot deck procedure.[1] Table 5-4 lists the variables in the order in which the missing data were imputed. The order of imputation addressed problems of multivariate association by using a series of univariate models fitted sequentially such that variables modeled earlier in the hierarchy had a chance to be included in the covariate set for subsequent models.

The weighted hot deck imputation procedure is best understood by first understanding unweighted hot deck imputation. The unweighted procedure partitions the sample into imputation classes based on auxiliary data available for both nonrespondents and respondents.

[1] Cox, B.G. (1980). "The Weighted Sequential Hot Deck Imputation Procedure." *Proceedings of the American Statistical Association Section on Survey Research Methods*, pp. 721–726.

Within these classes, it is assumed the nonrespondents answer in a manner similar to the respondents.

Also, the data records are often sorted within the classes to place individuals who shared additional characteristics closer to each other. The procedure is implemented by sequentially processing the database and replacing missing responses with the response from the previous respondent within each imputation class.

Table 5-4.—Statistically imputed variables and the amount of data imputed

Statistically imputed variable	Study respondent data used in imputations	Percent under-graduates	Percent graduates/ first-professionals	Number statistically imputed	Percent statistically imputed
Age (Age)	All	0.5	1.0	343	0.6
Gender (Gender)	All	1.3	2.5	959	1.6
Citizenship (Citizen2)	All	3.3	6.5	2,408	3.9
Hispanic ethnicity (Hispanic)	All	5.0	5.1	3,087	5.0
Race[1]	All	8.1	7.7	4,968	8.0
Student marital status (Smarital)	All	7.8	9.6	5,032	8.1
Dependents indicator (Anydep)	All	14.2	17.7	9,179	14.9
Dependency status indicator – 2 levels (Depend)	All	8.0	0.0	3,969	6.4
Dependency status indicator – 3 levels (Depend2)	All	14.7	17.7	9,447	15.3
Fall attendance status (Attend)	Students enrolled in fall 1999 (51,200)	1.4	1.2	691	1.3
High school degree indicator and type (Hsdeg)	All	7.1	18.6	5,772	9.3
Local residence (Localres)	All	17.0	18.6	10,704	17.3
Number of dependents (Ndepend)	Independents with dependents (15,600)	3.3	19.8	4,673	30.0
Parents' marital status (Pmarital)	Dependents (26,200)	13.9	†	3,582	13.7
Parent family size (Pfamnum)	Dependents (26,200)	13.9	†	3,582	13.7
Parents' income (Depinc)[2]	Dependents reporting parents' income category (14,300)	49.2	†	6,901	48.3
	Dependents not imputed in 1st stage (19,000)	19.1	†	3,602	19.0
High school graduation year (Hsgradyy)	Students with diploma/GED/cert. (61,100))	11.1	25.1	8,416	13.8
Student's income (Indepinc)	Independents (35,600)	23.9	26.1	8,761	24.6
Expected family contribution (Efc4)	All	42.8	65.0	29,086	47.1

†Not applicable.

[1]Race was an intermediary variable allowing for a full racial pattern of all possible multiple-listings of race. From this value, the variables R2WHITE, R2BLACK, R2ASIAN, R2ISLAND, and R2INDIAN were logically assigned. Appendix K provides further details.

[2]Of the approximately 26,200 dependent study respondents, 10,500 (40%) had missing values for parent income; however, parent income category was known for 6,900 of these students. Therefore, the imputation for parent income was performed in two stages. The first stage used a cross-classification of parent income category and parent marital status as the imputation classes among students who reported their parents' income category. The second stage imputed the remaining missing values among students who did not report their parents' income category. Appendix K provides details of the imputation for parents' income.

NOTE: To protect confidentiality, some numbers have been rounded.

SOURCE: U.S. Department of Education, National Center for Education Statistics, National Postsecondary Student Aid Study, 1999–2000 (NPSAS:2000).

The unweighted hot deck procedure reduces nonresponse bias if the response distributions differed *across* the imputation classes. However, a potential consequence of not using the sample weights is that bias may remain in the survey estimates due to the weighted distribution of the imputed data *within* the classes being different from the weighted distribution of the respondent data.

The weighted hot deck procedure is an extension of the hot deck procedure that considers the weighted distribution. The procedure takes into account the unequal probabilities of selection by using the student weights to specify the expected number of times that a particular respondent's answer will be used to replace missing data. Use of these expected selection frequencies allows the weighted distribution of the affected data to replicate the weighted distribution of the respondent data. Hence, the weighted hot deck imputation was designed so that, within each imputation class, the weighted survey estimates based on the imputed data are equal in expectation to the weighted survey estimates based on the respondent data.

To implement the weighted hot deck procedure, imputation classes and sorting variables that were relevant for each item being imputed were defined. If more than one sorting variable was chosen, a serpentine sort was performed where the direction of the sort (ascending or descending) changed each time the value of a variable changed. The serpentine sort minimized the change in the student characteristics every time one of the variables changed its value.

The respondent data for five of the items being imputed was modeled using a Chi-squared automatic interaction detector (CHAID) analysis to determine the imputation classes. These items were

- parent income (imputed for dependent students only),
- student income (imputed for independent students only),
- student marital status,
- local residence, and
- dependents indicator.

A CHAID analysis was performed on these variables because of their importance to the study and the large number of candidate variables available to form imputation classes. Also, for the income variables, trying to define the best possible imputation classes was important due to the large amount of missing data.

The CHAID analysis divided the respondent data (of each of these six items) into segments that differed with respect to the item being imputed. The segmentation process first divided the data into groups based on categories of the most significant predictor of the item being imputed. It then split each of these groups into smaller subgroups based on other predictor variables. It also merged categories of a variable that were found insignificant. This splitting and merging process continued until no more statistically significant predictors were found (or until some other stopping rule was met). The imputation classes were then defined from the final CHAID segments.

The federal methodology Expected Family Contribution (EFC) was available for 53 percent of the students in the NPSAS:2000 sample. The major sources for the EFC were the

1999-2000 Pell grant records(21 percent) and the student financial aid application records reported in the federal central processing system (CPS) for the 1999-2000 academic year (28 percent). In 5 percent of the cases neither of these was available, but an EFC was reported in CADE by the institution. For Pell Grant recipients, the EFC from the Pell record was always used.

The EFC was imputed for 47 percent of the 61,767 students on the file. Imputation regression equations were developed separately for the three categories of student dependency that have separate EFC formula types, using the EFC's recorded in the 1999-2000 CPS student records. EFC's were imputed for 40 percent of the dependent students, 55 percent of the independent students without dependents, and 50 percent of the independent students with dependents. More details on the EFC imputation are provided in Appendix K.

Appendix K presents the imputation classes and sorting variables used for all of the variables imputed by the hot deck approach, as well as other imputation procedures that were used. This appendix also includes a table showing the distribution of variables before and after imputation. When characteristics of nonrespondents significantly differed from characteristics of respondents and the imputation procedure successfully accounted for these differences, the distribution after imputation will be different from the distribution before imputation. . Following data imputations, variables were reviewed and revised [if necessary] to adjust for inconsistencies with other known data. Therefore, the distribution after imputation may differ from the final distribution.