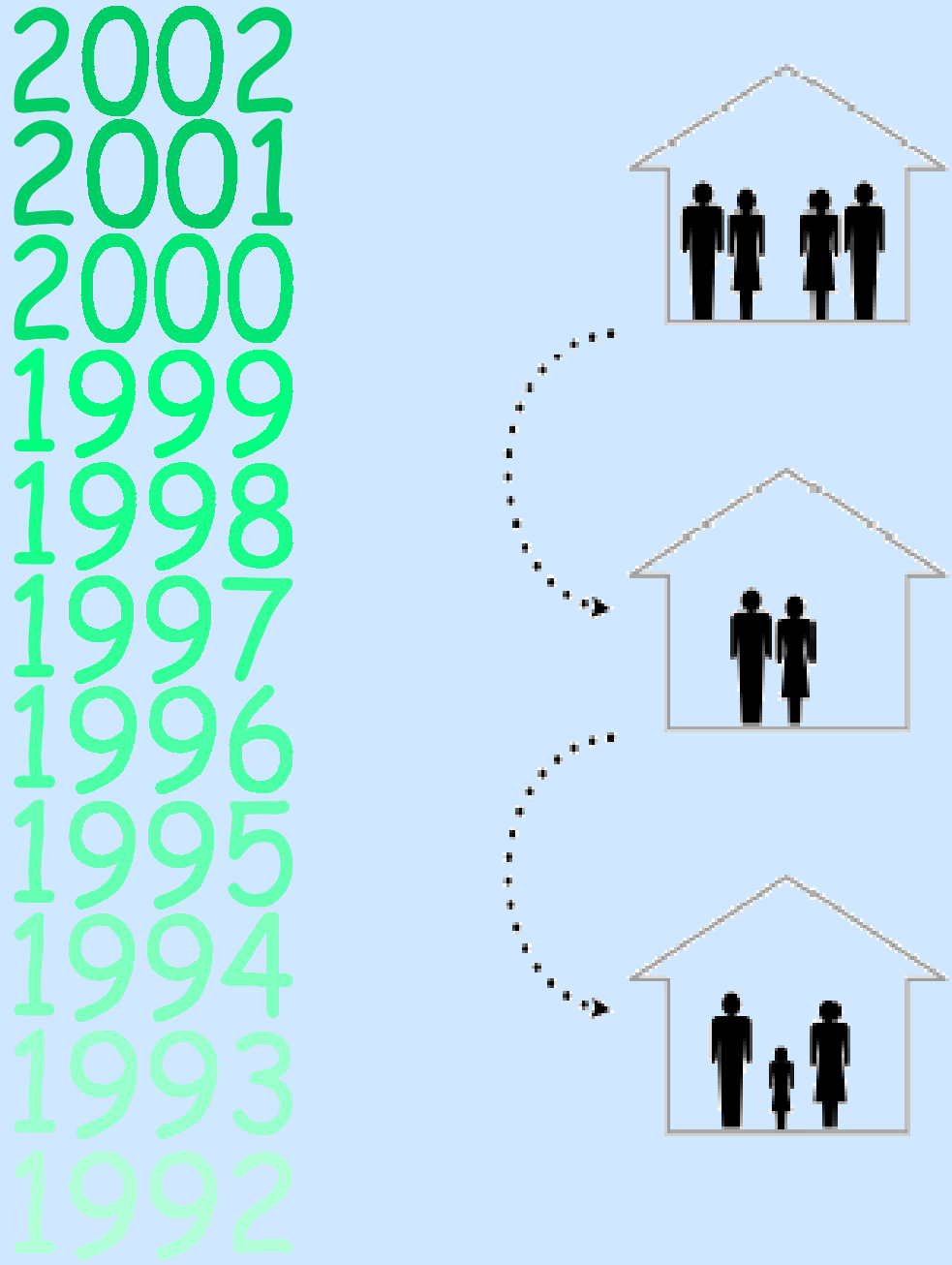
SPD Survey of Program Dynamics

2002

Users' Guide



Demographic Programs

USCENSUSBUREAU
Helping You Make Informed Decisions • 1902-2002

U.S. Department of Commerce
Economics and Statistics Administration
U.S. CENSUS BUREAU

Table of Contents

Abstract i

Executive Summary iii

Section I: Overview of the Survey 1
Chapter 1. Introduction 3
Chapter 2. The SPD Design and Implementation 11
Chapter 3. The SPD Survey Content 17

Section II: Accuracy of the Data 29
Chapter 4. Editing and Imputation 31
Chapter 5. Weighting 35
Chapter 6. Error Estimation 43

Section III: Working With the Public Use Microdata Files 51
Chapter 7. Using The 1997 SPD Experimental File 53
Chapter 8. Using The Unedited 1998 Calendar Year File 59
Chapter 9. Using the Longitudinal Files 65
Chapter 10. Analytic Uses of the Data 73

References 83

Appendixes 85
Acronyms and Abbreviations 87
Glossary 89

Index 91

Abstract

The *Survey of Program Dynamics Users' Guide* is primarily intended as a reference for analysts using data files produced and distributed by the U.S. Census Bureau. This document provides an overview of the Survey of Program Dynamics and its goals, a history of its development and implementation, an explanation of the survey's design (and the implications of that design), a description of available data products, and specific methods for accessing and analyzing the data.

The *Users' Guide* is divided into three sections, plus an appendix. The first section contains introductory material, including background information on the survey. The second section contains more technical information on how to properly use the data and interpret the results. The third section contains directions for working with the Survey of Program Dynamics data files. The appendix contains a list of abbreviations and a glossary of terms used in this document.

Some of the information contained in the first section of this guide may be useful to a broader audience, especially people who are interested in measuring the effects of welfare reform or in the methodology of sample surveys. An invaluable companion volume to this guide would be the *Survey of Income and Program Participation Users' Guide*, also available from the Census Bureau.

Executive Summary

The Survey of Program Dynamics (SPD) is a longitudinal, demographic survey designed to collect data on the economic, household, and social characteristics of a nationally representative sample of the U.S. population over time. The survey was created in response to the "Personal Responsibility and Work Opportunity Reconciliation Act of 1996" (Public Law 104-193), which required the U.S. Census Bureau to continue collecting data on the 1992 and 1993 panels of the Survey of Income and Program Participation (SIPP). The goal of the SPD is to provide policy makers the data necessary to assess the effects of national welfare reforms—how these reforms interact with each other, and with employment, income, and family circumstances.

This guide is intended as a reference for analysts using data from the SPD. The document provides an overview of the SPD and its goals, a history of its development and implementation, an explanation of the survey's design (and the implications of that design), a description of available data products, and specific methods for accessing and analyzing the data. Additional material includes a glossary of survey terms and a bibliography of relevant resources. An invaluable companion volume to this guide would be the *Survey of Income and Program Participation Users' Guide*, also available from the Census Bureau.

The Census Bureau has designed a series of the SPD data products for public use: one interim calendar year file (for 1998); six fully edited cross-sectional files (for 1997 through 2002); and three longitudinal files, containing fully edited, consistently-formatted, and longitudinally-processed core variables derived from the information collected over time (for 1992-1997, 1992-1999, and 1992-2001). Only the 1992 SIPP panel produced information for 1992; because neither of the SIPP panels produced information for the entire calendar year in 1995, none of the longitudinal files will contain data for that year.

Since the SPD estimates are based on a sample of households, the estimates may differ from those that would be obtained through a complete census. This document describes the use of sampling weights in analyzing the SPD data. It also describes methods for estimating the magnitude of errors resulting from sampling.

Additional information is available on the SPD Internet site: <http://www.sipp.census.gov/spd/>.

Section I: Overview of the Survey

Chapter 1. Introduction

This guide is intended primarily as a reference for analysts using data from the Survey of Program Dynamics (SPD). The document provides an overview of the SPD and its goals, a history of its development and implementation, an explanation of the survey's design (and the implications of that design), a description of available data products, and specific methods for accessing and analyzing the data. For analysts, an invaluable companion volume to this guide would be the *Survey of Income and Program Participation Users' Guide*, also available from the Census Bureau.

This chapter and the ones that follow come under three main sections: Section I contains introductory material, including background information on the survey; Section II contains more technical information on how to properly use the data and interpret the results; Section III contains directions for working with the SPD public use microdata files to answer specific research questions. This introduction offers a brief overview of each of those topics and an annotated outline of the chapters that follow.

Precursors of the SPD

During the Great Depression, the Enumerative Check Census (taken as a part of the 1937 unemployment registration) was the first attempt to estimate unemployment on a nationwide basis using probability sampling. There had been earlier attempts to estimate the number of unemployed, ranging from guesses to enumerative counts. Experience with the Enumerative Check Census, and research performed by the Work Projects Administration (WPA), led to the creation in 1940 of the Sample Survey of Unemployment. Responsibility for that survey was transferred from WPA to the Bureau of the Census in 1942, and the name of the survey was changed to the Current Population Survey (CPS). Since 1948, the CPS has included supplemental questions (at first, in April; later, in March) on income received in the previous calendar year.

In April 1973, the Office of Management and Budget's Statistical Policy Division asked the Interagency Committee on Income Distribution and the Interagency Committee on Poverty Statistics to conduct a thorough review of federal income and poverty statistics (Fisher 1992). Subcommittees were formed to study the following topics: updating the poverty threshold, improving the measurement of cash income, and measuring noncash income. One of the recommendations made by the Subcommittee on Measurement of Cash Income was for a separate income survey that would encompass items not covered by the March supplement of the CPS—to collect better money (and nonmoney) income data.

To address inadequacies in available survey data, the U.S. Department of Health, Education, and Welfare established the Income Survey Development Program (ISDP). The goal of the ISDP was to plan a recurring survey of income, assets, program eligibility, and participation. The ISDP researched and resolved a series of technical and operational issues before adopting a final design framework for a new survey, which became fully operational in 1983. That survey became the Survey of Income and Program Participation (SIPP).

The original design of the SIPP called for a nationally representative sample of individuals (15 years of age and older), to be selected in households in the civilian noninstitutionalized population. Those individuals, along with others who subsequently lived with them, were to be interviewed once every four months over a 32-month period. The first sample, the 1984 Panel, began interviews in October 1983 and finished in July 1986. The second sample, the 1985 Panel, began interviews in February 1985 and finished in August 1987. Subsequent panels (through 1993) began interviews in February of the calendar year. The 1993 panel finished interviewing in January 1996. There were no panels in 1994 and 1995, and the program was redesigned for 1996.

In early 1993, a group of Census Bureau scientists began discussions about developing an extended SIPP panel, to follow respondents for a period longer than four years. By the fall of 1994, rough goals for the survey were set: to provide information on actual and potential program participants over a ten-year period and to examine the

consequences of program participation on the well-being of recipients, their families, and their children. To deal with the likelihood of major welfare reform legislation (and with funding from the Departments of Agriculture and Health and Human Services), by the beginning of 1995 a Census Bureau workgroup had assembled content material for the survey—mostly from the content of the SIPP, with additional material submitted by various experts on children's research issues. Although a pretest of the instrument was planned for the spring of 1996 (with implementation planned for the spring of 1997), a lack of funding for the program resulted in its being sidelined.

In August 1996, the U.S. Congress enacted legislation to reform the national welfare system. That legislation, the "Personal Responsibility and Work Opportunity Reconciliation Act of 1996" (Public Law 104-193), specified (in Section 414) that the Census Bureau continue to collect data on the 1992 and 1993 panels of the SIPP. The legislation directs the Census Bureau to pay particular attention to the issues of out-of-wedlock births, welfare dependency, the beginning and end of welfare spells, the causes of repeat welfare spells, and the status of children in the surveyed households. In response to that legislation, the Census Bureau created the Survey of Program Dynamics (SPD).

Overview of the SPD

The Survey of Program Dynamics is a longitudinal, demographic survey designed to collect data on the economic, household, and social characteristics of a nationally representative sample of the U.S. population over time. The primary goals of the SPD are to provide information on spells of actual and potential program participation (over a ten-year period), to examine the causes of program participation and its long-term consequences (on recipients and their families), and to monitor the possible long-term changes (for individuals) that result from implementing welfare reform.

To provide policy makers the data necessary to assess the effects of national welfare reforms (how these reforms interact with each other, and with employment, income, and family circumstances), the SPD was designed to create a longitudinal database spanning a ten-year period and consisting of three components: information collected in the 1992 and 1993 panels of the SIPP; information collected in 1997 using a modified version of the March CPS; and information collected from 1998 to 2002 using the SPD instrument.

All SIPP people interviewed in the first wave of the 1992 and 1993 panels, and still being interviewed at the end of their panel, were eligible for the SPD sample. The 1997 SPD was a "bridge" between the earlier SIPP interviews and the new SPD survey, and used a modified version of the March CPS questionnaire (which includes the annual income supplement). The CPS annual income supplement obtains data for the previous calendar year on topics such as work experience, earnings, program participation, income, and health insurance.

The SPD questionnaire developed for 1998 through 2002 covers a wider variety of topics, to measure the impact of welfare reform legislation on previous program participants, and to compare their situations with those of the rest of the country.

SPD Uses

Analysts will be able to use the SPD longitudinal database to address all of the following research objectives: determine the types of jobs that previous welfare recipients are getting (and the types of employers hiring them); determine if their new employers are providing benefits (and how these benefits compared to those they received while on welfare); determine whether previous recipients used any type of training to obtain a job, whether they stay at the first job obtained after leaving the welfare system, or if they move on to a new job; measure the economic impact of welfare reform directly (by comparing information that shows whether a family's economic situation is better or worse after welfare reform, and whether those who have several jobs over a period of time make more money than those who stay at one job); measure how long people are unemployed between jobs, and how children are affected by parent's employment; estimate how long individuals go without health insurance and examine such lapses in coverage; illustrate

the relationship between work training, education, employment and earnings; show the effect of the welfare reform measures on people with disabilities, making it possible to relate disability status to income, employment, health insurance coverage and receipt or discontinuance of program benefits; and, monitor the effects of welfare reform on the nation.

The SPD Universe

The SPD universe consists of people who resided in the United States (except those living in institutions, such as prisons and nursing homes or entire military households) in March 1992 or March 1993. The universe is represented by original sample members from the 1992 and 1993 Survey of Income and Program Participation (SIPP) panels, except those who were subsampled out because of cost constraints or who left the survey universe before the 1998 interview. This population includes people (including children) living in group quarters, such as dormitories, rooming houses, and religious group dwellings. It does not include crew members of merchant vessels, Armed Forces personnel living in military barracks, and institutionalized people, such as correctional facility inmates and nursing home residents. In addition, United States citizens residing abroad were not eligible to be in the survey. Foreign visitors who work or attend school in this country and their families were eligible. All others were not eligible to be in the survey. With the exceptions noted above, people who were at least 15 years of age at the time of the interview were eligible to be asked income and job experience.

The SPD Sample

Based on their inclusion in the first and last waves of the 1992 and 1993 SIPP panels, there were 34,609 households eligible for the SPD. For 1998, the SPD sample was reduced (for budgetary reasons) to 19,129 households. The table below summarizes the sample sizes by year, along with information on the numbers of eligible and interviewed households.

SPD Households			
Sample	Eligible Households	Sample Households	Interviewed Households
1992/1993 SIPP	47,273	54,600	35,291
1997 SPD	48,633	34,609	30,125
1998 SPD	32,800	19,129	16,395
1999 SPD	33,200	19,303	16,659
2000 SPD[1]	33,600	23,258	18,716
2001 SPD[2]	34,000	29,341	22,340
2002 SPD	TBD[3]	TBD[3]	TBD[3]

[1]The 2000 SPD sample includes 19,802 households (base sample), plus 3,456 households that were selected for the 1997 SPD but were not interviewed.
[2]The 2001 SPD sample includes 20,185 households (base sample), plus 3,616 households that were selected for the 1997 SPD but were not interviewed, plus 5,540 households selected for the 1992/1993 SIPP that were not interviewed.
[3]TBD = To Be Determined

The SPD Content

The SPD longitudinal database will contain data collected using three different survey instruments: the 1992/1993 SIPP paper instruments, used to collect data for calendar years 1992, 1993, and 1994; a modified March CPS computer-assisted personal interviewing (CAPI) instrument, used to collect data for calendar year 1996; and the 1998 SPD CAPI instrument, used to collect data for calendar years 1997 through 2001.

SIPP Content. Information collected in the SIPP falls into two categories: core and topical module. The core content includes questions asked at every interview and covers demographic characteristics, labor force participation, program participation, amounts and types of earned and unearned income received (including transfer payments), noncash

benefits from various programs, asset ownership, and private health insurance. Most core data are measured on a monthly basis, although a few core items are measured only as of the interview date (once every four months). Topical module questions, asked less frequently to produce in-depth information on specific subjects, ask about particular social and economic characteristics, as well as personal histories. Topics include assets and liabilities, school enrollment, marital history, fertility, migration, disability, and work history.

1997 SPD "Bridge" Content. From April through June of 1997, the 1997 SPD used a modified version of the March annual income supplement to the Current Population Survey (CPS), to collect information about the previous calendar year. The instrument consists of demographic questions (questions about age, race, sex, ethnic group, marital status, and other personal characteristics) and questions on a wide variety of income sources.

1998-2002 SPD Content. Data collection for the 1998-2002 SPD occurs once each year, in May through July, gathering information about the previous calendar year. The information collected includes economic, demographic, and social characteristics of the people interviewed. Questions about demographic and social characteristics include educational enrollment and work training, functional limitations and disability, and health care use and health insurance. Questions about economic characteristics include employment and earnings, income sources and amounts, assets, liabilities, program eligibility information, and food security. Information about children is also collected, including their school enrollment and enrichment activities, disability, health care, child care arrangements, contact with an absent parent, and payment of child support on their behalf. A separate, self-administered section of the CAPI questionnaire collects information from adults on marital relationship, marital conflict, and parental depression.

In 1998 and 2001, a separate, self-administered paper questionnaire collected information from adolescents – on family conflict, vocational goals, educational aspirations, crime-related violence, substance abuse, and sexual activity. In 2000, the SPD included a Children's Residential History Calendar (RHC), designed to collect complete childhood histories of all children in SPD respondents' households. The RHC measures the number and timing of moves that children make.

For 1999 and 2002, the SPD included additional questions on children's extended measures of well-being, positive behavior/social competence, and conflict between parents.

The SPD Data Products

The Census Bureau has designed a series of the SPD data products for public use: one interim calendar year file, for 1998 (to support preliminary analysis of income and program participation among the original cohort); six fully edited cross-sectional files, for 1997 through 2002; and three longitudinal files, containing fully edited, consistently-formatted, and longitudinally-processed core variables derived from the information collected over time. The three longitudinal files will contain data for the following years: (1) for 1992 - 1994 and 1996 - 1997; (2) for 1992 - 1994 and 1996 - 1999; and (3) for 1992 - 1994 and 1996 - 2001. Only the 1992 SIPP panel provided information for 1992.

The SIPP 1992 and 1993 Longitudinal files, the 1997 SPD Bridge file, the 1998 SPD file, and the SPD First Longitudinal file are available from Marketing Services Office, Customer Services Center, U. S. Census Bureau, Washington, D.C. 20233. An extract file of the SIPP 1992 and 1993 Longitudinal files are available for downloading from the SIPP Internet site at <http://www.sipp.census.gov/sipp> under "Data Access" using one of the following extraction systems: the Federal Electronic Research and Review Extraction Tool (FERRET) or Data Extraction System (DES). Extract files of the 1997 SPD Bridge file are available for downloading from the SPD Internet site at <http://www.sipp.census.gov/spd> under Data Access using FERRET. Files are also available on CD-ROM (compact disc-readable) in ASCII format (call 301-457-4100 for price information).

Comparison to Other Surveys

The Census Bureau's Survey of Income and Program Participation (SIPP) and the University of Michigan's Panel Study of Income Dynamics (PSID) are two longitudinal surveys that can also be used to study the effect of welfare reform. Analysts can use the SIPP data to address many of the same questions they can address with the SPD data—except for the differences experienced by families and individuals, before and after national welfare reform. Because the PSID has interviewed individuals from the families in its core sample every year since 1968, the PSID data can be used to measure differences experienced by families and individuals, before and after national welfare reform. Additional information on the SIPP is available on the Internet at this address: www.sipp.census.gov/sipp. Additional information on the PSID is available on the Internet at this address: www.isr.umich.edu/src/psid.

The Census Bureau's Current Population Survey (CPS) and the Urban Institute's National Survey of American Families (NSAF) are two cross-sectional surveys that can also be used to study the effect of welfare reform. The CPS has already been used to study other non-experimental welfare changes, such as those made in 1981 to the Aid to Families with Dependent Children (AFDC) program. The NSAF data are being collected specifically to evaluate the 1996 changes. Additional information on the CPS is available on the Internet at this address: www.bls.census.gov/cps. Additional information on the NSAF is available on the Internet at this address: http://newfederalism.urban.org/nsaf/.

Researchers may also examine the effects of welfare reform by looking at pre-existing continuing experimental studies, such as welfare waiver demonstration projects. Other useful approaches include ethnographic studies, such as the Manpower Demonstration Research Corporations' Urban Change Study and the General Accounting Office's (GAO's) studies of welfare reform in selected states. Each of these surveys and studies will provide insights into some aspects of welfare reform and should be considered part of the portfolio needed to understand that major program change. Additional information on the Urban Change Study is available on the Internet at this address: www.mdrc.org/welfare_reform.htm. Additional information on the GAO's studies of welfare reform is available on the Internet by going to the GAO website at www.gao.gov/ and then searching for the phrase "welfare reform."

The SPD is a unique tool for evaluating reform because of its welfare reform-specific content, and because it offers the ability to analyze the economic and social well-being of families at two points in time as well as longitudinally over a 10-year period.

Guide to This Document

The balance of this *Users' Guide* is organized as follows:

The next two chapters are also introductory, designed mainly for beginning SPD users:

- Chapter 2 discusses how the SPD survey is designed and implemented. The chapter describes the structure of the survey, sample size and selection, and field procedures.
- Chapter 3 examines the general nature of questions in the SPD. The discussion focuses on the core and topical module content, including brief descriptions of individual topical modules.

Chapters 4 through 6 provide more technical information on how to properly use the data and interpret the results:

- Chapter 4 describes what happens after data collection. This chapter covers all aspects of post data collection processing, including consistency checks, data editing, and procedures for imputing missing data.
- Chapter 5 discusses the topic of weights in the SPD, with a focus on how to choose weights.
- Chapter 6 discusses the types and sources of error in the SPD, and discusses how to calculate sampling errors for the SPD estimates.

Chapters 7 through 11 provide specific instructions for the use of the SPD public use microdata files:

- Chapters 7 describes how to use the 1997 calendar year file.

- Chapter 8 describes how to use the interim, minimally edited, 1998 calendar year file.
- Chapter 9 describes how to use the edited, cross-sectional files. This chapter also describes the structure of those files and how to use the accompanying technical documentation.
- Chapter 10 describes how to use the longitudinal files. This chapter also describes the structure of those files and how to use the accompanying technical documentation.
- Chapter 11 describes some analytic applications using the SPD longitudinal data.

The SPD Users' Guide includes the following additional information:

- A list of references cited
- A guide to the acronyms and abbreviations used in this manual.
- A glossary of terms that may be unfamiliar to some users.
- An index of important topics and concepts.

Where to Go for More Information

The SPD Internet site <http://www.sipp.census.gov/spd/> provides links to the SPD files and documentation.

Chapter 2. The SPD Design and Implementation

Because the SPD is based on the 1992 and 1993 SIPP panels, this chapter begins with a brief examination of the SIPP sample design. Additional information on that design is available in the *Survey of Income and Program Participation Users' Guide* and in the *SIPP Quality Profile*. Following the discussion of the SIPP sample design, the topic turns to the design of the SPD sample.

The 1992/1993 SIPP Sample Design

The SIPP sample is a multistage, stratified sample of the U.S. civilian noninstitutionalized population. That population includes people living in group quarters, such as dormitories, rooming houses, and religious group dwellings. Foreign visitors who work or attend school in this country and their families were eligible. Crew members of merchant vessels, Armed Forces personnel living in military barracks, and institutionalized people, such as correctional facility inmates and nursing home residents, were not eligible to be in the survey. Also, U.S. citizens residing abroad were not eligible to be in the survey.

Sample selection for SIPP has three stages: the selection of primary sampling units (PSUs); the selection of address units in sample PSUs; and the determination of people and households to be included in the sample for the initial and subsequent waves of each panel. The first two stages are common to all household surveys, whether cross-sectional or longitudinal, that use multistage sample designs. The third stage is an additional requirement for longitudinal surveys.

The samples are located in 284 PSUs, each consisting of a county or group of contiguous counties. Within these PSUs, expected clusters of two living quarters (LQs) were systematically selected from lists of addresses prepared for the 1980 Decennial Census of Population and Housing to form the bulk of the sample. To account for LQs built within each of the sample areas after the 1980 census, a sample containing clusters of four LQs was drawn from permits issued for construction of residential LQs up until shortly before the beginning of the panel. In jurisdictions that do not issue building permits or have incomplete addresses, small land areas were sampled, expected clusters of four LQs within were listed by field personnel, and then these LQs were subsampled. In addition, sample LQs were selected from a supplemental frame that included LQs identified as missed in the 1980 census.

The SIPP is administered in panels and conducted in waves and rotation groups. The original design of SIPP called for an annual selection of a nationally representative sample of households (a panel), with all adults in those households being interviewed once every four months (a wave). Interviews were also conducted with any other adults living with original sample members at subsequent waves. Each panel was divided into four subsamples of roughly equal size (rotation groups), with one rotation group getting interviewed each month, for information about the previous 4-month period. Since the first panel in 1984, the number of waves per panel has varied from 3 to 13.

People interviewed in the first wave of an SIPP panel are called original sample members. The original sample members are the "units" that are followed longitudinally. If an original sample member (age 15 or older) leaves a household, he or she is followed and interviewed in the new household. If a household was not interviewed for a wave of SIPP, the household was recontacted during the next wave to be brought back into sample. Households that were non-interviews for two waves in SIPP were dropped from further attempts. The exception being households that were not interviewed because they could not be located: for them, a third attempt for contact was permitted.

In preparation of the 1996 redesign of the SIPP, the Census Bureau canceled the 1994 and 1995 panels and extended the 1992 panel an additional wave. The last interview for the 1992 panel took place in April 1995; the last interview for the 1993 panel took place in January 1996.

The 1997-2002 SPD Survey Design

The 1997 SPD bridged the gap in data between the close of the SIPP panels and the start of the SPD by recontacting the interviewed sample people from the 1992 and 1993 SIPP panels. The sample size for the SPD Bridge Survey was 34,609 households. Census field representatives interviewed 30,125 households during the SPD Bridge Survey.

At any given point in time, a household is eligible to be interviewed if it contains an original sample member (age 15 or older). The number of eligible households fluctuates from round to round of interviewing because of household formation and dissolution—and because original sample members move from one (previously eligible) household to another (previously ineligible) household.

The sample for the 1998 SPD was 19,129 households, subsampled from households interviewed in the 1997 SPD Bridge Survey. The 19,129 households selected for the SPD 1998 (and beyond) met one of the following criteria:

- One hundred percent of the households where the primary family or the primary individual has a total family income below 150 percent of the poverty threshold. The number of cases is 6,182.

- One hundred percent of the households where the primary family or the primary individual has a total family income between 150 percent and 200 percent of the poverty threshold, and there are children under 18. The number of cases is 1,075.

- Ninety percent of the households where the primary family or the primary individual has a total family income above 200 percent of the poverty threshold, and there are children under 18. The number of cases is 6,623.

- Eighty percent of the households where the primary family or the primary individual has a total family income between 150 percent and 200 percent of the poverty threshold, and there are no children under 18. The number of cases is 1,461.

- Twenty-seven percent of the households in the balance. The number of cases is 3,707.

- Twenty-seven percent of the SIPP cases that were institutionalized during the SPD Bridge. The number of cases is 81.

Census Bureau field representatives interviewed 16,395 of the eligible households during the 1998 interview period.

The 1999 SPD sample consisted of all eligible households from the 1998 SPD, including households that were interviews, refusals, temporarily absent, and unable to locate. Census Bureau field representatives interviewed 16,659 of the eligible households during the 1999 interview period.

The 2000 SPD sample consisted of all eligible households from the 1999 SPD, supplemented with 3,456 households subsampled from those that were noninterviews for the 1997 SPD Bridge. For 2000, Census Bureau field representatives interviewed 18,716 of the eligible households.

The 2001 SPD sample consists of 20,185 eligible basic SPD cases, 3,616 eligible 1997 SPD Bridge non interviews added in 2000, further supplemented with 5,540 eligible 1992 and 1993 SIPP non interviews (from Waves 2 through 10). For 2001, Census Bureau field representatives interviewed 22,340 of the eligible households.

Methods to Maximize Response

The SIPP respondents provided 9 or 10 waves of detailed data over a three-year period. The SIPP data collection had a burden of 30 minutes per adult respondent per wave. So the average SIPP household (2.1 adults per household) had provided more than 10 hours of their time. At the end of the last wave of the SIPP interviews, respondents were thanked for their time and told that there would be no more interviews. Then, 1 to 2 years later, the respondents were

contacted and told they were still in a panel survey. Therefore, it was not surprising that the SPD would have nonresponse problems.

The reduction of sample through attrition is a concern. The SPD inherited a 26.6 percent sample loss rate from the SIPP sample. However, after two years of the SPD, the sample loss rate was 50 percent. Procedures used during 1999 and 2000 helped to slow the sample loss rate.

Sample Attrition Rates for the 1992 and 1993 SIPP Panels and the SPD				
Survey	Eligible Households	Interviewed Households	Average Sample Loss Rate (%)	Interview Rate (%)
1992/1993 SIPP	47,273	35,291	26.6	73.4
1997 SPD	48,633	30,125	41.3	58.7
1998 SPD	32,800	16,395	50.0	50.0
1999 SPD	33,200	16,659	49.8	50.2
2000 SPD Basic	33,600	16,845	49.9	50.1
2000 SPD Basic + 97NI	33,600	18,716	44.3	55.7
2001 SPD Basic + 97NI + 1992/93 SIPP NI	34,000	22,340	37.0	63.0

Previous studies on the SIPP sample loss have shown that the sample loss is not uniform (Mack and Petroni 1994; Lamas et al. 1994; Zabel 1993). Households in and near poverty are lost at a higher rate than other households. Since poverty households are a key target population in the study of welfare reform, there is some concern about nonresponse bias.

The 1998-2002 SPD uses several techniques to maximize response rates and ensure the accuracy of the information collected. One of the techniques the Census Bureau uses is the special training given to field personnel. This training emphasizes the conversion of nonresponse households or refusals to complete interviews. Field employees are taught to stress the importance of the survey to the respondent, the positive results that can be obtained from the survey if everyone participates fully, and the satisfaction of knowing that the respondents' answers helped their government or other individuals. Another technique is nonresponse follow-up. Senior field representatives attempt to interview any household that refuses to do the study.

Households receive an interim mailout letter approximately two months before the start of interviewing. This letter thanks the respondent for their past participation in the SPD and explains how their continued participation in the SPD helps make decisions that affect all citizens. In addition, the households also receive a fact sheet containing information from previous data collections of the SIPP and the SPD and a change of address card to let regional office (RO) staff know when the household has a new address.

The recontacts and attempts to interview nonrespondents from earlier interviewing cycles have helped to maximize response. The introduction of 3,456 noninterviews has increased the longitudinal response rate of 50 percent in the 1999 SPD to 55.7 percent in the 2000 SPD.

To determine the effectiveness of monetary incentives on improving response rates, an experiment was included in the 1997 SPD. Low income households in a subset of sample clusters received $20 vouchers. Compared to a group of low income households in a similar subset of sample clusters, the response rate for the voucher households was slightly (but

not significantly) higher (Creighton et al. 2000).

No incentives were used in the 1998 SPD. For the 1999 SPD, eligible but not interviewed households from the 1998 SPD received $40 debit cards in an advance letter, by priority mail, prior to the interview cycle. Each receiving household was allowed to cash the incentive regardless of the 1999 interview outcome. In addition, other households that were reluctant to continue the survey in 1999 were given a $40 debit card as part of the conversion procedures.

For the 2000 SPD, they distributed a $40 debit card to households that received (or were eligible for) an incentive in 1999—and to potential refusals. An incentive of $100 was offered to a sample of households that had been noninterviews for the 1997 SPD and had not been contacted in 1998 or 1999. The incentive was given whether an interview was obtained or not.

For the 2001 SPD, a $40 debit card was distributed to households that received incentives in 1999 or 2000 (and gave interviews); to households that were eligible but not interviewed in 1998, 1999, or 2000; to households that refused to participate in the 2001 SPD (but had not refused in the past); and to eligible but not interviewed households from the 1997 SPD.

Households that were part of the 5,540 eligible but not interviewed cases from the 1992 and 1993 SIPP panel received an advance letter containing a $100 debit card incentive prior to the SPD field representatives visit in 2001 (assuming a valid address was available). The advance letter and incentive were sent via priority mail. Households receiving a $100 incentive were allowed to cash the incentive regardless of the interview outcome.

In 1998, when the Adolescent SAQ was conducted with the SPD, the response rate was 58 percent. To increase this response, households with adolescents in the Basic and 1997 SPD noninterview sample receive an additional conditional $40 incentive in 2001. The incentive was provided to the household respondent/parent if all children 12 to 17 completed their Adolescent SAQ. However, households receiving the $100 incentive did not receive an additional incentive for completion of the Adolescent SAQ.

Following Movers

The SPD rules call for following original sample members (15 years old or older) who move—provided they are not institutionalized, do not live in military barracks, or do not move abroad. People added to a household roster after the initial SIPP interview are called additional people or non-sample people. Non-sample people are not followed to new addresses if they move unless they move with a sample person.

If an entire household moves, an interviewer tries to find the original sample members and interview them at their new address(es). If only some original sample members move, an interviewer completes interviews with all eligible household members at both the original address and the address(es) of those who have moved.

Chapter 3. The SPD Survey Content

This chapter provides an overview of the SPD content. Tables at the end of the chapter summarize the differences in content among the three components of the longitudinal data collection: the 1992/1993 SIPP, the 1997 Bridge Survey, and the 1998-2002 SPD.

The 1992/1993 SIPP

For the 1984 to 1993 Panels, SIPP data were collected by means of paper and pencil instruments that consisted of a control card and a questionnaire. Basic demographic characteristics and other classification variables associated with a household and its members were recorded on the control card in the initial interviews for a panel and updated in each subsequent wave. The survey questionnaire consisted of core questions, which were repeated at each wave, and topical modules, which included questions on selected topics. The topical modules varied from wave to wave. The main topics covered by the core questions were labor force participation and sources and amounts of income. Information for most items in these categories was obtained at every interview for each of the four months included in the interview reference period.

SIPP distinguishes between two kinds of topical modules: fixed and variable. Fixed topical modules are modules that are included in one or more waves during the life of each panel to augment the core data. They include, for example, modules on annual income, retirement accounts, income taxes, educational financing and enrollment, personal history, and wealth. Variable topical modules, which are designed to satisfy the special programmatic needs of other federal agencies, are not necessarily repeated from one panel to the next. Some topics that have been covered are child care arrangements, child support agreements, support for nonhousehold members, long-term care, pension plan coverage, housing costs, and energy usage. Variable modules were usually included in Waves 3 and 6 while fixed modules appeared in other waves.

More detailed information on the SIPP content is available in the *SIPP Users' Guide*.

The 1997 SPD "Bridge" Survey

The 1997 SPD used a slightly modified version of the March 1997 Current Population Survey (CPS), which asks questions about employment and income in the past year. The 1997 SPD Bridge Survey also included a few questions not collected in 1995 from the 1992 SIPP panel, questions about the receipt of public assistance.

The 1998-2002 SPD

The 1998-2002 SPD uses the core SPD questionnaire (described below) and two self-administered modules: one set of questions for adults, focusing on marital relationship, marital conflict, and parental depression; the other, a completely separate questionnaire for adolescents (administered only with parental consent), focusing on family conflict, vocational goals, educational aspirations, crime-related violence, substance abuse, and sexual activity.

The SPD core instrument included retrospective questions for all people aged 15 years and over, focusing on such topics as jobs, income, and program participation. Additional questions focusing on children in the household gathered information on school status, activities at home, child care, health care, and child support.

For the 1999 SPD, the core questions were expanded to include the following topics: new questions asked about independence, assets, vehicle operating expenses, substance abuse, health care utilization while uninsured, and food expenditures. In addition to the core questions, the 1999 SPD asked questions on Extended Measures of Children's Well-Being: new questions asked about positive behavior and social competence, family routines, and conflict between parents.

In addition to the core questions, the 2000 SPD employed the Children's Residential History Calendar topical module, which asked about with whom children have lived and the reasons for any changes in living arrangements.

In addition to the core questions, the 2001 SPD used the same Adolescent Self-Administered Questionnaire employed in 1998.

The 2002 SPD will use the core SPD instrument, plus additional questions on Extended Measures of Children's Well-Being.

Core Questions for Adults

Household Roster and Coverage

The SPD tracks movements into and out of family groups. The household roster and coverage questions establish the household composition and the relationships of those who live with the original sample members. They obtain important information about the household members for future reference in the interview and for future tabulations.

Employment and Earnings

For each person age 15 or over in the household, the SPD collects a detailed account of work-related activities in the past calendar year, including weeks worked, weeks on layoff, and weeks spent looking for work, as well as whether or not they are currently working. In addition, the SPD collects detailed employment data, for up to four jobs in the previous calendar year including annual earnings from each job.

Income Sources

These questions are similar to those from SIPP. An inventory consists of all the types of income received during the previous calendar year for all household members age 15 and older. Household-level screening questions determine if anyone in the household received income from specific sources. If so, the FR asks who received that type. This section also contains questions about cash assistance for low-income households. Cash assistance questions comparable to these have also been added to the Current Population Survey instrument.

Dependent Interviewing

Questions in the "Independent/Dependent Comparison" section of the questionnaire are asked about each person 15 years of age and older. If a household member reported in a prior interview the receipt of a particular type of income, this section seeks to confirm if the household member received the same type of income in the previous year. This series of questions also provides an option for replacing incorrect data reported in the prior interview.

Amounts

This section of the questionnaire is designed to obtain the amount of income received during the reference period from each income source reported in the previous section, and the number of months it was received for selected income. This section also contains cash assistance questions for low-income households. Cash assistance questions comparable to these are also included in the Current Population Survey instrument.

Eligibility and Assets

Selected questions about assets and debts are included because they are critical to measuring program eligibility. These assets include the value of homes, cars, stocks, bonds, and mutual funds. Other payments critical to eligibility include medical expenses, child support, and energy costs. Some items, such as stocks and bonds are covered in previous sections on income sources and amounts. These questions are asked of everyone—to measure changes that occur among previous program participants, and to obtain a picture of the rest of the population with which to compare their answers.

Vehicle Operating Expenses

The purpose of these questions is to find out what types of transportation are available to respondents, which type is used, how much is spent on work-related travel, and whether transportation issues are limiting respondents' employment or training opportunities.

Educational Enrollment

The educational enrollment part of the SPD instrument collects information on the enrollment of people age 18 and older in regular school, including post-secondary vocational, technical, or business school. People 15 to 17 will be included in the children's school enrollment questions since we believe that the children's series of questions is more appropriate for that age group.
The focus of this section is on basic education and general skills development, and will track the progress of adults toward receiving high school or high school equivalency degrees as well as college and graduate degrees.

Work Training

The work training part of the SPD instrument is intended to collect information on the training that people age 15 and older have received either to help them find a job or to get a better job. This training may focus on the following:

1. Basic academic preparation—e.g., reading skills, math skills or preparation for high school equivalency (GED).
2. Training to learn a specific job skill—e.g., word processing, auto mechanics.
3. Other training to improve job skills or learn a new job.
4. Job search assistance—placement service.
5. Job readiness training—e.g., resume writing, interviewing.
6. Unpaid work experience or community service work.

The first question, on basic academic preparation, is asked only of people whose current educational attainment is below the associate degree level. Questions 2, 5, and 6 are asked only of those who have received or applied for public assistance in the past year. The two remaining questions (3 and 4) are asked of all respondents.

Work training is not basic education of the sort one would receive in a high school or college. Nor is it the general skill development that one would expect to receive in a post-secondary vocational, technical, or business school. These are covered in the adult Educational Enrollment section of the instrument.

The main differentiating factor between training and education is the nature of the credential awarded. Training is strictly vocational in nature. Any award or certificate for completion of the program is purely incidental to the purpose of training for employment. Only in rare instances would training count in a program in regular school leading to a degree.

In some instances the training programs focus on the job search process itself. These programs may focus on résumé preparation, interviewing skills, or organizing one's schedule or life's circumstances to allow work.

Substance Abuse

Substance abuse can prevent people from getting and keeping jobs. States can deny benefits to people to who use drugs. Therefore, it is necessary to ask about substance usage when talking about welfare. These questions are only asked of adults 18 or older. All answers are respondent-defined.

Functional Limitation and Disability

The ability to see, hear, carry items, walk short distances, and perform other activities may affect employment status and the ability to live independently. These questions are condensed versions of a similar series included as topical modules in the Census Bureau's Survey of Income and Program Participation (SIPP) and National Health Interview Survey (NHIS) surveys.

Health Care Utilization

These questions are included to measure changes in the U.S. health care system and how the changes affect accessibility to health services. As the health care system of the U.S. changes, an important goal of the SPD will be to chart how these changes affect coverage, health utilization, and outcomes. To that extent, we need to know how individuals are accessing the health care system.

Health Insurance

Questions on health insurance are condensed versions of a similar series included in the SIPP core. These questions are included to measure changes in the U.S. health care system and how the changes affect accessibility to government health insurance such as Medicaid and Medicare as well as private or employer-provided insurance. The first series of questions are about health insurance coverage for the previous year, with a follow-up about current health coverage.

Health Care Utilization While Uninsured

As the health care system in the United States evolves, an important goal of the SPD will be to chart how these changes affect coverage, health care utilization, and outcomes. It is therefore important to know to what extent individuals without health care coverage are able to access health care services.

Food Expenditures and Food Security

This series of questions is taken from the USDA-sponsored Food Security Supplement to the CPS. It is intended to measure the subjective experience of hunger. The questions are used as a scale to measure the severity of hunger in a household. Food expenditure questions ask how people spend their money on food. The introductory food security question serves as a screening question—those with higher incomes who have "enough and the kinds of food" skip to the next section of the instrument.

The subsequent scale incorporates:

- increasing food insecurity
- anxiety
- perceptions
- incidents of reduced food intake in adults
- incidents of reduced food intake in children.

Core Questions for Children

Children's School Enrollment

These questions track children's progress through and out of school over time. A critical element of the well-being of children is their enrollment, at an appropriate age, in school and their normal progress through the educational system. School enrollment includes both preschool and regular school, kindergarten through twelfth grade. The former includes both Federally-funded Head Start and other pre-kindergarten programs with a substantial educational or school readiness component.

Children's Enrichment Activities

This section of the SPD instrument is intended to collect information on activities, in addition to schooling, which promote the development of children. Some of these activities are school-related functions such as sports and clubs. Others are home or community activities that the child might do independently or jointly with parents or other household members.

Children's Disability

Parents of children with disabilities often have special financial burdens, and there is concern about access to educational services. This series of questions is asked for children 14 and under. The FR interviews the designated parent or guardian of the child.

Children's Health Care Utilization

This series of questions is asked for children 14 and under, to record how children are accessing the health care system. The FR interviews the designated parent or guardian of the child.

Mother's Work Schedule

The SPD asks about activities associated with work, school, training, and looking for work for the designated parent to determine the demands for child care on the family for each child.

Child Care

The SPD collects information on child care arrangements for working and non-working parents: what child care arrangements parents make, especially while they are working, looking for work, going to school, or attending work training; how much parents pay for child care and whether these costs are paid in part or in full by the government, an employer, or someone else; and how often parents miss work or leave children to care for the children themselves because regular child care arrangements are not available.

Child Support Agreement

These questions are asked of households containing children no older than 20 years of age. One aspect of welfare reform is improved compliance with child support agreements. The amount of child support received by a parent or legal guardian is an important factor in determining the economic well-being of children. Also, the child support questions asked in this section will allow users of SPD data to examine the evolving system of child support awards and enforcement in the U.S.

Contact with Absent Parent

An objective of welfare reform is to encourage closer family ties and greater responsibility of parents for their children. Absent parents may participate in and contribute to their children's well-being by providing economic resources or by spending time with them, or both. These questions measure the amount of time absent parents spend with their children.

Adult Self-Administered Questions

Marital Relationship and Conflict

Marital relationships may be affected by changes in welfare reform policies, (e.g., a spouse's finding a job may improve the relationship if household income rises, or it may cause the relationship to decline if child care problems are exacerbated). It is also evident from prior research that the frequency and level of inter-parental conflict are related to children's adjustment.

Depression Scale

This section is about feelings the respondent may have experienced over the past 30 days. These questions explore the respondents' feelings about themselves and how they perceive their lives.

Self-Administered Adolescent Questionnaire

Adolescents between the ages of 12-17 are interviewed directly, because their knowledge of their own behavior often differs radically from parents' or guardians' knowledge of their children's behavior and attitudes. These questions are administered by audio-cassette with the adolescent filling in an answer booklet. The questionnaire takes about 20-30 minutes. The answer booklet contains only the answers to the questions and not the questions themselves in order to protect the adolescent's privacy.

Children's Residential History Calendar

The Residential History module collects information about the childhood residential histories of people who were recorded as children (18 years of age and younger) in Wave 1 of the 1992/1993 SIPP or on a subsequent SPD roster. The module is designed to measure variability in the living arrangements of children. To gauge the disruptions in children's lives, this module measures all instances of more than three months in which children lived away from their biological mothers or biological fathers, and all instances of more than three months when children shared a residence with adults other than their biological parents, regardless of whether the biological parent(s) was also present.

Extended Measures of Child Well-Being

In 1999, the SPD asked a series of questions devoted to measuring child well-being. These questions will be asked again in the 2002 SPD.

Comparability of Content Across SPD Components

The components comprising the SPD (the 1992/1993 SIPP, the 1997 SPD Bridge, and the 1998-2002 SPD) have different recall periods and different levels of aggregation. Respondents in the SIPP panels were interviewed three times per year and faced a recall period of four months. Respondents in the 1997-2002 SPD, interviewed only once a year, faced recall periods up to fifteen months. Also, some topics that appear on more than one component (for example, receipt of welfare) are covered by questions that have different categories of response. The following table summarizes the major differences in the content and comparability of the SPD components.

Content and Comparability of the Three Components of the Survey of Program Dynamics

(applies to age 15+ unless otherwise indicated)

Topic	Instrument		
	1992/1993 SIPP Panels	**1997 SPD Bridge Survey**	**1998-2002 SPD**
Basic Demographic Characteristics	at time of survey	at time of survey	at time of survey
Adolescent and Child Questions*			
Family Routines			at time of survey
Interaction with Parents			at time of survey and last 12 months
School Routines and Behaviors			last school year
Parental Rules			at time of survey
Delinquent Behaviors			last 12 months
Substance Use			ever, first, last 30 days
Dating and Sexual Behavior			first, last, and at time of survey
Armed Forces Status	at time of survey and ever	at time of survey and ever	at time of survey and ever
Child Care			
Child Care Arrangements	at time of survey**, last month, and changes during last 12 months		Jan. of previous year to May of current year and at time of survey
Child Care Hours and Amounts	at time of survey** and typical week		this April and last calendar year
Mother's Work Schedule			this April
Work and Child Care Conflicts	last month** and typical week**		last calendar year
Child Enrichment Activities			
Sports, Clubs, and Lessons	at time of survey (6-17)** past month**		last September to this April
TV, Reading, Outings			at time of survey
Gang Activity			at time of survey and ever
Job			at time of survey
Education			
Attainment	at time of survey (6-17** and 15+)	at time of survey	at time of survey

Topic	1992/1993 SIPP Panels	1997 SPD Bridge Survey	1998-2002 SPD
Education (continued)			
School Enrollment	at time of survey (children)** and monthly	last week and 1996	monthly (3+) and last (children)
Financial Aid	last 12 months** and last 4 months	1996	last 12 months and last school year
Post-Secondary Educational Expenses	last 12 months**		
Expelled or Repeated Grade	past (children)**		ever, when?
Child's School Progress*			overall and last school year
English Ability and Other Language			at time of survey
Family Context			
Marital Relationship and Conflict			at time of survey, past few months, last year
Parental Depression Scale			last 30 days
Child Activities*			at time of survey and last week/month/year
Problem and Positive Child Behaviors*			last 3 months
Family Structure			
Marital Status	past** and at time of survey	at time of survey and change since last month	at time of survey and month/year of change
Contact with Absent Parent			last and at time of survey
Residential History*			birth to age 18
Food Security	last 4 months**		last 12 months
Food Sufficiency			last 12 months
Immigration Status			
Nativity, Citizenship	at time of survey**	at time of survey	at time of survey
Date of Entry	at time of survey**	when	
Migration	past**	where living 1 year ago, 3/1/96	
Work Training	past**		last 12 months

Topic	1992/1993 SIPP Panels	1997 SPD Bridge Survey	1998-2002 SPD
Employment & Earnings	monthly	previous calendar year	
Work/Employment Status			previous calendar year, which weeks?
Layoffs/Looking for Work			previous calendar year, which weeks?
Reasons NOT Working			current
Earnings			previous calendar year
Income Sources (excluding earnings)	monthly	previous calendar year	previous calendar year
Unemployment			which months? which weeks?
Worker's Compensation			which months? which weeks?
Social Security			which months?
Supplemental Security Income (SSI)			which months?
Food Stamps			which months?
AFDC/TANF			which months?
WIC			which months?
Child Care			which months?
General Assistance			which months?
Other Assistance			which months?
Veteran's/Disability Payments			which months?
Assets			previous calendar year
Child Support			which months?, which weeks?
Income Amounts	monthly	total of previous calendar year	total of previous calendar year
(For Each Previously Listed Source)			(may be reported weekly, bi-weekly, monthly or annually)
Eligibility & Assets	topical modules	N/A	
Housing & Real Estate	topical modules		current monthly mortgage
Automobile Information	topical modules		current
Assets	topical modules		current
Debts	topical modules		current
Eligibility & Assets (continued)			
Child Support Paid	monthly		previous calendar year
Other Support Paid	monthly		previous calendar year
Disability			

Topic	1992/1993 SIPP Panels	1997 SPD Bridge Survey	1998-2002 SPD
Functional Limitations & Disabilities	once per wave, current	N/A	current
Health			
Health Care Utilization	N/A	N/A	previous calendar year
Medical Expenses	N/A	N/A	previous month
Health Insurance	monthly	previous calendar year, current	previous calendar year, which months? current
Uninsured Utilization***	N/A	N/A	previous calendar year
Food Expenditures* **	N/A	N/A	previous calendar year
Public Housing	once per wave, current	current	current

*from SPD topical module
**From SIPP topical module
* **Added in 1999 SPD

Section II: Accuracy of the Data

Chapter 4. Editing and Imputation

This chapter describes the data editing and imputation procedures applied to data from the 1997-2002 Survey of Program Dynamics, after completion of interviews. For information on data editing and imputation for the 1992/1993 Survey of Income and Program Participation, see the *Survey of Income and Program Participation Users' Guide*.

Three different approaches are used for dealing with missing data in the 1997-2002 SPD:

- Weighting adjustments (discussed in Chapter 5) are used for some types of noninterviews.
- Data editing (also referred to as logical imputation) is used for some types of item nonresponse.
- Statistical (or stochastic) imputation is used for some types of unit nonresponse and some types of item nonresponse.

This chapter begins with a brief discussion of the types of missing data and the goals of imputation in the SPD. It then presents an overview of the editing and imputation procedures used to deal with missing and inconsistent data. Next, the chapter provides a detailed description of each of the major steps used by the Census Bureau when creating its internal files and the files that are released for public use.

Types of Missing Data

As in most surveys, there are three types of missing data in the SPD: household nonresponse, person nonresponse, and item nonresponse. Household nonresponse (also called whole unit nonresponse) occurs when an interviewer finds an eligible household's address but obtains no interview. This can happen as a result of people not being at home or being unwilling or unable to participate in the survey. Household nonresponse also occurs when a sample household has moved to an unknown or unavailable address. Household nonresponse is dealt with through weighting adjustments (see Chapter 5).

Person nonresponse (also called Type Z nonresponse) occurs when an interview is obtained from at least one household member but an interview is not obtained from one or more other sample people in that household. Like household nonresponse, this can happen as a result a person being unwilling, unable, or unavailable to answer questions. Person nonresponse is dealt with through editing and imputation.

Item nonresponse occurs when a respondent completes part of the questionnaire but does not answer one or more individual questions. Item nonresponse can occur under any of the following circumstances: a respondent refuses or is unable to provide requested information; a response is inconsistent with related responses or is incompatible with the response categories; an interviewer fails to ask a question or to record an answer; an interviewer makes an error when recording or keying in the response. For item nonresponse, data are generally imputed for core items.

Goals of Imputation

Missing data cause a number of problems: analyses of data sets with missing data are more problematic than analyses of complete data sets; there is a lack of consistency among analyses because analysts compensate for missing data in different ways and their analyses may be based on different subsets of data; and, in the presence of nonresponse that is unlikely to be completely random, estimates of population parameters are biased.

Because missing data are always present to some degree, analyses of survey data must be based on assumptions about patterns of missing data. When missing data are not imputed or otherwise accounted for in the model being estimated, the implicit assumption is that data are missing at random after controlling for other variables in the model. The imputation procedures used for the SPD are based on the assumption that data are missing at random within subgroups of the population (as defined by the cells of the imputation matrices, described later in this chapter).

The statistical goal of imputation is to reduce the bias of survey estimates. This goal is achieved to the extent that systematic patterns of item nonresponse are correctly identified and modeled. In the SPD, the statistical goals of imputation are general, rather than specific. Instead of addressing the estimation of specific parameters, the SPD procedures are designed to provide reasonable estimates for a variety of analytical purposes.

Data editing is generally preferred over statistical imputation, and it is used whenever a missing item can be logically inferred from other data that have been provided. When information exists on the same record from which missing information can logically be inferred, that information is used to replace the missing information. The advantage of data editing is that it avoids the increase in variance that occurs when missing items on one record are imputed with nonmissing responses from other records.

Assessing the Influence of Imputed Data on Analysis

Users of the SPD data interested in assessing the influence of imputed data on their analyses should consider whether the SPD imputation procedures have properties that affect their specific analytical requirements. An evaluation of the effects of imputed data should include a review of rates of unit nonresponse and an assessment of the extent of item nonresponse. Unit nonresponse tends to increase over the life of a panel, as does the likelihood that nonresponse is not a random effect. As the percentage of eligible sample members re-interviewed decreases, the pool from which donors are selected shrinks accordingly. This smaller pool of donors leads to an increased likelihood that individual donors will be used more than once, which in turn increases the variance of an estimate. The effects of imputation will likely be small for items with low rates of missing data, as long as rates of item nonresponse are not high among important subclasses not controlled for in the imputation process.

Overview of the Editing and Imputation Process

The editing process effectively blanks all inappropriate entries and ensures that all appropriate questions have valid entries. For some variables, editing ensures consistency over time and

agreement within a household. The main purpose of editing and imputation is to assign values to questions where the response was "Don't know" or "Refused." This is accomplished by using one of the imputation techniques described below. Edits are run in a deliberate and logical sequence. That is, demographic variables are edited first because several of those variables are used in allocating missing values for other types of variables. Similarly, labor force participation variables are edited before income variables. In all, there are twelve different categories of variables in the editing sequence.

The SPD uses the following imputation methods:

- *Logical imputation* infers the missing value from other characteristics on a person's record or within the household. For instance, if race is missing, it is assigned based on the race of another household member or, failing that, taken from the previous record on the file. Similarly, if relationship is missing, it is assigned by looking at age and sex of the person in conjunction with the known relationship of other household members. Missing occupation codes are sometimes assigned by viewing the industry codes and vice versa.
- *"Hot deck" imputation* assigns a missing value from a record with similar characteristics. Hot decks are always defined by age, race, and sex. Other characteristics used in hot decks depend on the nature of the question being referenced. For instance, most labor force questions use only age, race, sex, and occasionally another labor force item (such as full- or part-time status).
- *"Cold deck" imputation* procedures use group estimates (such as means) for the sample as a whole or for subgroups within it as the source of information for the values to assign to those cases for which data are missing.
- *Longitudinal edits* are used for the longitudinal files. If a question is blank, the edit looks at the previous year's data to determine whether there was a non-allocated entry for that item. If so, the previous year's entry is used to assign a value to the missing item; otherwise, the item is assigned a value using the appropriate hot deck.

For the 1997-2002 SPD files, every variable that is subject to editing and imputation will have an associated flag to designate the source of its value. For example, the imputation flag for "occupation of longest job" will have one of the following values:

0 Not imputed
1 Statistical imputation (hot deck)
2 Cold deck imputation
3 Logical imputation (derivation)

All of the editing and imputation procedures described above are part of the process of preparing the data for internal Census Bureau use. Before the files are released for public use, they undergo additional editing to protect the confidentiality of respondents. Three procedures are used: topcoding selected variables (income, assets, and age), suppression of geographic information, and recoding by collapsing categories of responses into broader categories. As a result of these procedures, estimates based on data from the public use files will differ slightly from the Census Bureau's published estimates.

On the SPD longitudinal data files, there will be no imputation flags for the data from the 1992-1994 SIPP. To check for edits or imputations on those records, analysts will need to link them to the SIPP files from which they came.

Chapter 5. Weighting

This chapter describes the use of sampling weights in analyzing data from the Survey of Program Dynamics (SPD). Each SPD file contains either just one set or a number of alternative sets of weights for use in data analysis. The several different sets of weights are needed to allow optimal use of the sample data and analysis with different time periods for which survey estimates may be required.

A common mistake in the analysis of a survey like the SPD is to ignore the weights entirely, that is, to perform an unweighted analysis. This chapter explains why an unweighted analysis is likely to produce biased estimates. It also describes the different sets of weights on the SPD files and identifies the set that is appropriate for particular analyses.

What Weights Are

The weight for a responding unit in a survey data set is an estimate of the number of units (people, families, or households) in the target population that the unit represents. In general, since population units may be sampled with different selection probabilities and since response rates and coverage rates may vary across subpopulations, different responding units represent different numbers of units in the population. The use of weights in survey analysis compensates for this differential representation.

A number of data products produced from the SPD are cross-sectional (calendar year) data files. However, the weights included in those files are not cross-sectional weights, as exist on a cross-sectional file like the March CPS. Instead, because the survey is principally designed to be longitudinal, the weights are longitudinal. The survey sample was subsampled from the SIPP Panel 1992 and 1993 samples. Therefore, it is important to remember that the SPD universe consists of people who resided in the United States (except those living in institutions, such as prisons and nursing homes or entire military households) in March 1992 or March 1993. That universe is not fully representative of the U.S. population as of the time of the SPD interview.

SIPP Final Panel Weight

Several stages of weight adjustments were involved to produce the SIPP longitudinal panel weight. Each person received a base weight equal to the inverse of his or her probability of selection. Two noninterview adjustment factors were applied. One adjusted the weights of interviewed people in interviewed households to account for people who were eligible for the sample but could not be interviewed at the first interview. The second was applied to compensate for people who were not interviewed in subsequent interviews.

An additional stage of adjustment to longitudinal person weights was performed to reduce the mean square error of the survey estimates. This was accomplished by bringing the sample estimates into agreement with the monthly Current Population Survey (CPS) estimates of the civilian (and some military) noninstitutional population of the United States by age, gender, race, Hispanic origin (Note: Hispanics can be of any race), and householder/not householder status as

of the specified control date. The control months for the 1992 and 1993 SIPP panels were March 1992 and March 1993, respectively. The CPS estimates were adjusted with estimates from the 1990 decennial census for undercount and to reflect births, deaths, immigration, emigration, and changes in the Armed Forces since 1990. For the weighting of the SPD calendar year and longitudinal panel files, the control month for the SPD panel universe was nominally chosen as March 1993.

Weighting the 1997 SPD File

The longitudinal panel weight covering the time period between 1992 and 1996 on the 1997 SPD file is LGTPERWT. Each person was assigned one crude longitudinal weight. The weight assigned depended on the individual's longitudinal interview status during the SIPP panels and the SPD Bridge. Each weight is the product of three components: the SIPP Longitudinal Panel Weights, Combined Panel Factor, and the Bridge Nonresponse Factor. The product of these three components produces the SPD longitudinal weight.

The SIPP final panel weights were adjusted by a factor of one-half, due to combining two nationally representative samples together of approximately equal size. Then an additional adjustment factor was applied to each interviewed case by age, race/ethnicity, and sex that simultaneously adjusted for the SPD Bridge nonresponse and under-coverage to form the 1997 SPD Bridge final weights.

Interviewed, noninterviewed, and excluded people for the SPD Bridge are defined below. Both person and household interview status codes were used to define these groups. Only people residing in a sample household at the first interview of SIPP and considered longitudinally interviewed for the SIPP are eligible for an SPD longitudinal weight.

1. Interviewed People

   This group is comprised of eligible SPD Bridge sample people (including children) who were successfully linked to a SIPP panel, considered an interview longitudinally for the SIPP, and interviewed (or had died or moved to an ineligible address) in the SPD Bridge survey.

2. Noninterviewed People

   This group is comprised of all eligible people, (including children), who were successfully linked to a SIPP panel, considered an interview longitudinally for the SIPP, but were not interviewed in the SPD Bridge survey (excluding imputed people and people who died or moved to an ineligible address). This includes noninterviewed people in an interviewed sample household.

3. Excluded People

   Everyone else who does not meet the criteria for interviewed or noninterviewed people.

All sample people classified as interviewed for the entire longitudinal period, (that is, the SIPP, and the SPD Bridge) were assigned positive longitudinal weights for the 1998 SPD (based on the

weighting calculation procedure described earlier). People classified as noninterviewed or excluded were assigned zero weights.

Application of the Weights on the SPD 1997 File

The longitudinal panel weights on this file are only applicable for crude estimates of the longitudinal characteristics (e.g., unemployment spell length) of people and families in the SPD universe for the time period within 1992 and 1996. The crude weights were provided because the refined weights on the SPD first longitudinal file were not available at the time. They served as means to perform preliminary estimates and research in the early stage of the SPD. Since the data from 1992 to 1995 are not available on this file, they must be obtained by matching the sample people back to either the SIPP Panel 1992 and 1993 longitudinal files or the SPD first longitudinal file. The data on the SIPP Panel 1992 and 1993 files are monthly but those on the SPD first longitudinal file are yearly. Since the 1992 data are available only for the sample units from the SIPP Panel 1992, which is approximately half the SPD sample size, the weights used any 1992 estimates must be twice the longitudinal weights on the file (i.e., 2 × LGTPERWT.) The variances of the estimates for this year will need to be inflated by two as well.

Weighting the 1998 SPD File

Each person was assigned one crude longitudinal panel weight covering the time period between 1992 and 1997. The longitudinal panel weight on the file is LGTPERW8. The weight assigned depended on the individual's longitudinal interview status during the SIPP panels and the SPD (1997) Bridge, and the 1998 SPD.

In the calculation of the SPD 1998 longitudinal final weight, the SPD Bridge longitudinal final weight acted as the initial weight and then was adjusted for the additional nonresponse that occurred during the 1998 interviewing cycle. The SPD Bridge longitudinal final weight was calculated from the SIPP longitudinal final panel weight and similarly adjusted for additional nonresponse since the end of SIPP. Details of the weighting components are given below.

The SPD Bridge final weights were adjusted by the sample cut factor. Then an additional adjustment factor (similar to the one for the SPD Bridge) was applied to each interviewed case by age, race/ethnicity, and sex that simultaneously adjusted for the 1998 SPD nonresponse and under-coverage to form the 1998 SPD final weights.

Interviewed, noninterviewed, and excluded people for 1998 SPD are defined below. Codes for both person and household interview status were used to define these groups. People who met all of the following conditions are eligible for a 1998 SPD longitudinal weight: residing in a sample household at the first interview of the SIPP, considered longitudinally interviewed for the SPD Bridge, and not subjected to the 1998 SPD sample cut.

1. Interviewed People

   This group consists of the eligible 1998 SPD sample people (including children) who were considered a longitudinally interviewed person for the SPD Bridge, and were interviewed (self or proxy or imputed) or died or moved to an ineligible address in the 1998 SPD Survey.

2. Noninterviewed People

   This group consists of all eligible people (including children) who were considered a longitudinally interviewed person for the SPD Bridge, but were not interviewed (self or proxy or imputed) in the 1998 SPD survey (excluding people who died or moved to an ineligible address).

3. Excluded People

   Everyone else who did not meet the criteria for interviewed or noninterviewed people.

All sample people classified as interviewed for the entire longitudinal period, (that is, the SIPP, the SPD Bridge, and the 1998 SPD) were assigned positive longitudinal weights for the 1998 SPD (based on the weighting calculation procedure described earlier). People classified as noninterviewed or excluded were assigned zero weights.

Application of the Weights on the SPD 1998 File

The longitudinal weights on this file are only applicable for crude estimates of the longitudinal characteristics (e.g., unemployment spell length) of people and families in the SPD universe for the time period within 1992 and 1997. The crude weights were provided because the refined weights on the SPD first longitudinal file were not available at the time. They served as a means to perform preliminary estimates and research in the early stage of the SPD. Since the data from 1992 to 1995 are not available on this file, they must be obtained by matching the sample people back to either the SIPP Panel 1992 and 1993 longitudinal files or the SPD first longitudinal file. The data on the SIPP Panel 1992 and 1993 files are monthly but those on the SPD first longitudinal file are yearly. Since the 1992 data are available only for the sample units from the SIPP Panel 1992, which is approximately half the SPD sample size, the weights used any 1992 estimates must be twice the longitudinal weights on the file (i.e., 2 × LGTPERW8.) The variances of the estimates for this year will need to be inflated by two as well.

Weighting the First Longitudinal File

For the SPD longitudinal data, the sample people who meet the following definition have a positive final weight:

- Lived in a 1992/1993 SIPP panel household during Wave 1 interviews.
- Were interviewed (self, proxy or imputed) for each reference month in SIPP.
- Were interviewed (self, proxy or imputed) in the 1997 SPD Bridge and the 1998 SPD.

Not all persons with imputed waves will have positive weights, only those whose missing waves are bounded by self or proxy interviews. Unlike the crude weighting, those who continued to be interviewed until they died or moved to an ineligible address during the SPD interviews were classified non-interviewed and were assigned a zero weight. This group of original sample people jointly represents the SPD universe. Other people included in the data file have zero weights. Their presence on the data file is to facilitate development of household and family

characteristics of the people in the longitudinal sample. This will permit the user to construct contextual information on the cohort sample member's household and economic circumstances. They refine the longitudinal weights on the first longitudinal file but those on the 1997 and 1998 SPD files are crude. Since their availability on the first longitudinal file, they supersede the crude weights in the SPD 1997 and 1998 files for any analyses.

For the first longitudinal file, there are two longitudinal weights: SPDLNWGT and ANNUALWT. The first, SPDLNWGT, which is the longitudinal panel weight and should be used for calculating estimates covering multiple calendar years. The second, ANNUALWT, which is a longitudinal annual weight derived to account for the children born after the first SIPP interview. The ANNUALWT should be used for annual or calendar year estimates. The SPDLNWGT and ANNUALWT are identical except for the non-original sample children born after the first SIPP interview and under the parental care or guardianship of original sample people. For these children, SPDLNWGT are zero and ANNUALWT are identical to their designated (biological/adopted or guardian) parents.

The sample people who meet the following definition have a positive final weight:

- Lived in a 1992/1993 SIPP panel household during Wave 1 interview.
- Were interviewed (self, proxy or imputed) for each reference month in SIPP.
- Were interviewed (self, proxy or imputed) in 1997 SPD Bridge and 1998 SPD.

For SPDLNWGT, all the other sample people included on the file have zero final weights. For ANNUALWT, sample children aged 6 or less (if spawned from the SIPP Panel 1992) and aged 5 or less (if spawned from the SIPP Panel 1993) are assigned the same weight as their designated parents, if the parent is an original sample member. If the parent is not an original sample member, the child’s weight was assigned as zero. (A designated parent of a child can be a biological parent, an adopted parent, a blood-related guardian, or a not-blood-related guardian.)

An original sample member is a person who at the time of the Wave 1 interview resided in an interviewed sample household (or group quarters). An initial weight was assigned to each original sample member (including children) based on their probability of selection. The inverse of this initial weight represents the probability of an original sample member residing in an interviewed Wave 1 sample household in either the SIPP Panel 1992 or 1993 (depending on which SIPP panel he or she originally belonged). The initial weight was the base weight adjusted to account for eligible households that were selected for interview in Wave 1 but not interviewed.

Since each of the SIPP Panels (1992 and 1993) was a nationally representative sample by itself, combining them into one sample reduce the weight of each panel sample person proportionately to their sample sizes. Since the sample sizes of the SIPP Panels 1992 and 1993 are approximately the same, a combined panel factor of one-half was assigned to each of the original sample members.

Because not all of the original sample members were interviewed in each reference month in SIPP and in the 1997 SPD, weights of members who were interviewed in all periods were adjusted to compensate for members who were not. Similarly, not all of the original sample members who made it through all the SIPP interviews and the 1997 SPD interview were

interviewed in 1998. Two adjustments to weights compensated for that: one to account for the sample reduction (due to budget constraints); the other to account for those in the sample who were not interviewed.

A final adjustment to weights involved "raking" to match a set of SPD population estimates with a corresponding set of control (benchmark) population estimates for March 1993. The control population estimates were based on the following demographic variables: age, sex, race, ethnicity, householder living with or not living with a relative, not-householder related to or not related to householder. This adjustment serves as a means to improve the population coverage of the SPD sample and also serves as a post-sampling stratification to reduce the mean square error of the estimates.

Children from the 1992 SIPP aged 6 or less and from the 1993 SIPP aged 5 or less received SPDLNWGT values of zero. If the designated parent was an original sample member, the child received an ANNUALWT equal to that of the parent. Otherwise, the child received an ANNUALWT of zero.

Application of the Weights on the First Longitudinal File

On the longitudinal file, the longitudinal panel weight, SPDLNWGT, should be used for any estimates covering multiple years within 1992 to 1997, and the longitudinal annual weight, ANNUALWT, should be just for any annual or calendar year estimates. However, the SPDLNWGT is also recommended to be used for any annual or calendar year estimates if the estimates do not concern the characteristics of the children born after the first interview of the 1992/1993 SIPP panels. Some caution should be taken when using the ANNUALWT for estimating the characteristics of children aged six and less. Because of the approach used to assign the weights to the sample children born after the first SIPP interview, the estimates for the children in this age group are generally 2.2 percent higher than the corresponding 1998 benchmark estimates. By race for the children in this age group, the estimates are 3.6 percent higher than the benchmark estimates for non-Black, and 5.4 percent lower the benchmark estimates for Black children. Since the 1992 data are available only for the sample units from the SIPP Panel 1992, which is approximately half the SPD sample size, the weights used any 1992 estimates must be twice the longitudinal weights on the file (i.e., 2 × SPDLNWGT or 2 × ANNUALWT.) The variances of the estimates for this year will need to be inflated by two as well.

Summary of the Weights on the SPD Files

A summary of the weights on the SPD calendar year and longitudinal files is provided in the table below. The weight description and applications of these weights are also included in the summary.

File	Weight Variable	Weight Description and Application
1997 SPD	LGTPERWT	The LGTPERWT is a crude longitudinal panel weight for estimates covering 1992 to 1996. It is produced for use in the preliminary estimates or research on the SPD prior to the availability of the SPD first longitudinal file and the longitudinal panel weight, SPDLNWGT. The LGTPERWT is generally superseded by the SPDLNWGT since its availability.
1998 SPD	LGTPERW8	The LGTPERW8 is a crude longitudinal panel weight for estimates covering 1992 to 1997. It is produced for use in the preliminary estimates or research on the SPD prior to the availability of the SPD first longitudinal file and the longitudinal panel weight, SPDLNWGT. The LGTPERW8 is generally superseded by the SPDLNWGT since its availability.

File	Weight Variable	Weight Description and Application
SPD First Longitudinal	SPDLNWGT	The SPDLNWGT should be used for any estimates covering multiple years within 1992 to 1997. However, the SPDLNWGT is also recommended to be used for any annual or calendar year estimates if the estimates do not concern the characteristics of children born after the first interview of the 1992/1993 SIPP panels.
	ANNUALWT	The ANNUALWT should be used for any annual or calendar estimates. This weight was derived to account for children born after the first interview of the 1992/1993 SIPP panels.

Chapter 6. Error Estimation

Because the Survey of Program Dynamics (SPD) estimates are based on a sample, they may differ somewhat from the figures that would have been obtained if a complete census had been taken (using the same questionnaire, instructions, and enumerators). There are two types of errors possible in an estimate based on a sample survey: nonsampling and sampling. Although it is possible to provide estimates of the magnitude of the SPD sampling error, this is not true of non-sampling error. This chapter begins by describing sources of non-sampling error in the SPD, then discusses sampling error—its estimation and its use in data analysis.

Nonsampling Errors

Nonsampling errors can be attributed to many sources: for example, inability to obtain information about all cases in the sample, difficulties in precisely stating some definitions, differences in the interpretation of questions, inability or unwillingness on the part of the respondents to provide correct information. Other types of errors may take place in recording, coding, or processing the data; in estimating values for missing data; in biases resulting from the differing recall periods caused by the rotation pattern used; or because of undercoverage.

Undercoverage in the SPD results from missed living quarters and missed people within sample households. It is known that undercoverage varies with age, race, and gender (Martin and de la Puente 1993). Generally, undercoverage is larger for males than for females and larger for Blacks than for non-Blacks. Ratio estimation to independent age-race-gender population controls (benchmark estimates) partially corrects for the bias due to survey undercoverage. However, biases exist in the estimates to the extent that people in missed households or missed people in interviewed households have characteristics different from those of interviewed people in the same age-race-gender group. In addition, the independent population controls used have not been adjusted for undercoverage in the decennial census.

The Census Bureau has used complex techniques to adjust the weights for nonresponse. For an explanation of the techniques used, see the “Non-response Adjustment Methods for Demographic Surveys at the U.S. Bureau of the Census,” November 1988, Working Paper 8823, by R. Singh and R. Petroni. An example of successfully avoiding bias can be found in "Current Non-response Research for the Survey of Income and Program Participation" (paper by Petroni, presented at the Second International Workshop on Household Survey Non-response, October 1991). The procedure for calculating the longitudinal person weights on the first SPD longitudinal file was derived based on such complex techniques.

Sampling Errors

The sample selected for each SPD panel is a stratified multistage probability sample. This complex sample design needs to be taken into account when estimating the variances of the SPD estimates. The SPD data files contain variables, related to the sample design, that are created for the purpose of variance estimation. Several software packages are now available for computing variance estimates for a wide range of statistics based on complex sample designs. Using the variables that specify the design, these programs can calculate appropriate variances of survey estimates. The Census Bureau also provides generalized variance functions (GVFs) that can be used to obtain approximate estimates of sampling variance for the SPD estimates. Information on these functions may generally be found in the technical documentation associated with the data files.

A common mistake in the estimation of sampling errors for survey estimates is to ignore the complex survey design and treat the sample as a simple random sample of the population. This mistake occurs because most standard software packages for data analyses assume simple random sampling for variance estimation. When applied to the SPD estimates, SRS formulas for variances typically underestimate the true variances.

Direct Variance Estimation

The primary sampling unit (PSU) plays a key role in variance estimation with a multistage sample design. The SIPP PSUs are mostly counties, groups of counties, or independent cities, which are sampled with probability proportional to size within strata. The PSUs were sampled without replacement so that no PSU was selected more than once for the sample. Some PSUs are so large that they are included in the sample with certainty. Because no sampling is involved, those PUSs are, in fact, not PSUs but strata. The actual PSUs for those certainty selections are the enumeration districts and other units selected within them.

Although the SIPP PSUs were selected without replacement (as is the case with most multistage designs), for the purpose of variance estimation they are treated as if they were sampled with replacement. The with-replacement assumption greatly facilitates variance estimation, since it means that variance estimates can be computed by taking into account only the PSUs and strata, without the need to consider the complexities of the subsequent stages of sample selection. This widely used simplifying assumption leads to an overestimation of variances, but the overestimation is not great.

Several software packages are available for computing variances of a wide range of survey estimates from complex designs. For example, for means and proportions for the entire sample and for subclasses, for differences in means and proportions between subclasses, and for regression and logistic regression coefficients. These packages use a variety of methods for variance estimation. Some use an approach based on a Taylor series approximation, or linearization, method. Others use a replication method, such as jackknife repeated replications or balanced repeated replications. Although some methods have advantages in some situations, there is generally little to recommend one method over another. The variance estimates they produce are not identical, but the differences are usually small.

Using GVFs to Approximate the Standard Error of an Estimated Number

The GVFs for the SPD were derived by modeling the standard error behavior of groups of estimates with similar standard errors. The mathematical form of the function adopted is

$$s = \sqrt{(ax^2 + bx)}$$

where s represents the standard error and x the value of an estimate. The parameters a and b are derived on the basis of a selected group of estimates. They are updated annually and are included in the source and accuracy statement that accompanies each SPD data file. It is essential to use the parameter estimates for a specific panel and to follow the instructions to apply necessary adjustments to obtain the correct estimates for subgroups.

Using GVFs to Approximate the Standard Error of an Estimated Mean

A mean is defined here to be the average quantity of some characteristic (other than the number of people or households) per person or household. For example, a mean could be the average monthly household income of females 25 to 54 years of age. The formula used to estimate the standard error of a mean is

$$s_{\bar{x}} = \sqrt{\frac{b}{y} s^2}$$

where y is the size on which the estimate is based, s^2 is the estimated population variance of the characteristic, and b is the parameter associated with the particular type of characteristic. With the use of standard software for weighted data, the estimated population variance of the characteristic can be computed as

$$s^2 = \frac{\sum w_i (x_i - \bar{x})^2}{\sum w_i} \text{ or } \frac{\sum w_i (x_i - \bar{x})^2}{\sum w_i - 1}, \text{ where } \bar{x} = \frac{\sum w_i x_i}{\sum w_i}$$

Because of the approximations used in developing this formula, an estimate of the standard error of the mean obtained from this formula will generally underestimate the true standard error.

Using GVFs to Approximate the Standard Error of an Estimated Aggregate

An aggregate is defined to be the total quantity of a characteristic summed over all units in a sub-population. The formula used to estimate the standard error of an aggregate is

$$s_{x_a} = \sqrt{bys^2}$$

As with the estimate of the standard error of a mean, the estimate of the standard error of an

aggregate will generally underestimate the true standard error.

Using GVFs to Approximate the Standard Error of an Estimated Percentage

The reliability of an estimated percentage, computed using sample data for both numerator and denominator, depends upon both the size of the percentage and the size of the total upon which the percentage is based. Estimated percentages are relatively more reliable than the corresponding estimates of the numerators of the percentages, particularly if the percentages are more than 50 percent. When the numerator and denominator of the percentage have different parameters, use the parameter of the numerator. If proportions are presented instead of percentages, note that the standard error of a proportion is equal to the standard error of the corresponding percentage divided by 100.

There are two types of percentages commonly estimated. The first type is the percentage of people sharing a particular characteristic such as the percentage of people owning their own home or the percentage of 1996 food stamp recipients who were also receiving food stamps in 1997. The second type is the percentage of money or some similar concept held by a particular group of people or held in a particular form. Examples are the percentage of wealth held by people with high income and the percentage of annual income received by females.

For the percentage of people, the formula used to estimate the standard error is

$$s_{x,p} = \sqrt{\frac{b}{x} p(100 - p)}$$

Here, x is the base of the percentage, p is the percentage (0<p<100), and b is a parameter for the numerator of the percentage calculation.

To calculate the percentages of money, the formula is more complicated. A percentage of money will usually be estimated in one of two ways. It may be the ratio of two aggregates

$$p_M = 100\left(\frac{X_A}{X_N}\right)$$

where X_A and X_N are aggregate money figures, or it may be the ratio of two means with an adjustment for different bases

$$p_M = 100\left(\frac{\overline{X}_A}{\overline{X}_N}\right)\hat{p}_A$$

where $\overline{X}_A$ and $\overline{X}_N$ are mean money figures, and $\hat{p}_A$ is the estimated number in Group A divided by the estimated number in Group N. For either method of estimating the percentage, the formula used to estimate the standard error is

$$s_{p_M} = \sqrt{\left(\frac{\hat{p}_A \overline{X}_A}{\overline{X}_N}\right)^2 \left[\left(\frac{s_{\hat{p}_A}}{\hat{p}_A}\right)^2 + \left(\frac{s_{\overline{X}_A}}{\overline{X}_A}\right)^2 + \left(\frac{s_{\overline{X}_N}}{\overline{X}_N}\right)^2\right]}$$

where $s_{\hat{p}_A}$ is the standard error of $\hat{p}_A$, $s_{\overline{X}_A}$ is the standard error of $\overline{X}_A$ and $s_{\overline{X}_N}$ is the standard error of $\overline{X}_N$. To calculate $s_{\hat{p}_A}$, use the formula for estimating the standard error of an estimated percentage of people. To calculate $s_{\overline{X}_A}$ and $s_{\overline{X}_N}$, use the formula for estimating the standard error of an estimated mean. Note that there is frequently some correlation among the characteristics estimated by $\hat{p}_A$, $\overline{X}_A$, and $\overline{X}_N$. These correlations, if present, will cause a tendency toward overestimates or underestimates, depending on the relative sizes of the correlations and whether they are positive or negative.

Using GVFs to Approximate the Standard Error of an Estimated Difference

The standard error of a difference between two sample estimates, *x* and *y*, is equal to

$$s_{x-y} = \sqrt{s_x{}^2 + s_y{}^2 - 2rs_x s_y}$$

where s_x and s_y are the standard errors of the estimates *x* and *y*. The estimates can be numbers, averages, percents, ratios, etc. The correlation between *x* and *y* is represented by *r* (*0 # r # 1*). If *r* is assumed to be zero and the true correlation is really positive (negative), then this assumption will result in a tendency toward overestimates (underestimates) of the true standard error.

Using GVFs to Approximate the Standard Error of an Estimated Median

The median quantity of some item is that quantity such that at least half the group has as much or more and at least half the group has as much or less. The sampling variability of an estimated median depends upon the form of the distribution of the item as well as the size of the group. The median, like the mean, can be estimated using either data that has been grouped into intervals or ungrouped data.

If grouped data are used, the median is estimated using one of the formulas given in Step 4 below. If ungrouped data are used, the data records are ordered based on the value of the item (e.g., income level), then the estimated median is the value of the item such that the weighted estimate of 50 percent of the sub-population falls at or below that value and 50 percent is at or above that value. The method of standard error computation presented here requires the use of grouped data.

An approximate method for measuring the reliability of an estimated median is to determine a confidence interval about it. The following procedure may be used to estimate the 68-percent confidence limits (approximately ± one standard error from the median) and so the standard error of a median based on sample data.

Step 1 - Estimate the standard error of an estimate of 50 percent of the group.

Step 2 - Subtract from and add to 50 percent the standard error determined in Step 1 to obtain the percentages associated with the lower and upper limits of the 68-percent confidence interval of the item. Namely, the smaller percentage is $50 - s_{x,p=50}$ percent, and the larger percentage is $50 + s_{x,p=50}$ percent.

Step 3 - Using the distribution of the item within the group, calculate the quantity, X_{UCL} of the item such that the percent of the group owning more of the item is equal to the smaller percentage $(50 - s_{x,p=50})$ found in Step 2. This quantity (X_{UCL}) will be the upper limit for the 68-percent confidence interval (assuming that the interval with higher item value is ranked at lower percentile). In a similar fashion, calculate the quantity, X_{LCL} of the item such that the percent of the group owning more of the item is equal to the larger percentage $(50 + s_{x,p=50})$ found in Step 2. This quantity (X_{LCL}) will be the lower limit for the 68-percent confidence interval. (Note that a median computed from ungrouped data may or may not fall in this confidence interval).

Step 4 - Divide the difference between the two quantities (X_{UCL} and X_{LCL}) determined in Step 3 by two to obtain the standard error estimate ($s_{\hat{X}_{med}}$) of the median estimate ($\hat{X}_{med}$). Namely,

$$\hat{s}_{\hat{X}_{med}} = \frac{X_{UCL} - X_{LCL}}{2}$$

To perform Step 3, it will be necessary to interpolate, which may be done using different methods. The most common are simple linear interpolation and Pareto interpolation. The appropriateness of the method depends on the form of the distribution around the median. We recommend Pareto interpolation in most instances. Interpolation is used as follows. The quantity of the item, X_{pN} such that p percent own more of the item is

$$X_{pN} = A_1 \exp\left[\frac{\ln\left(\frac{pN}{N_1}\right)}{\ln\left(\frac{N_2}{N_1}\right)} \ln\left(\frac{A_2}{A_1}\right)\right]$$

if Pareto Interpolation is indicated and

$$X_{pN} = \left[\frac{pN - N_1}{N_2 - N_1}\left(A_2 - A_1\right) + A_1\right]$$

if linear interpolation is indicated, where N is the size of the group; A_1 and A_2 are the lower and upper bounds, respectively, of the interval in which X_{pN} falls; N_1 and N_2 are the estimated numbers of group members owning more than A_1 and A_2, respectively; *exp* refers to the exponential function; and *Ln* refers to the natural logarithm function. One should note that a mathematically equivalent result is obtained by using common logarithms (base 10) and antilogarithms.

Using GVFs to Approximate the Standard Error of an Estimated Ratio of Means

The standard error for a ratio of means or medians is approximated by

$$s_{\frac{X}{Y}} = \sqrt{\left(\frac{X}{Y}\right)^2\left[\left(\frac{s_X}{X}\right)^2 + \left(\frac{s_Y}{Y}\right)^2\right]}$$

where X and Y are the means or medians, and s_X and s_Y are their associated standard errors. Formula 6.C-16 assumes that the means or medians are not correlated. If the correlation between the population means or medians estimated by X and Y are actually positive (negative), then this procedure will tend to produce overestimates (underestimates) of the true standard error for the ratio of means or medians.

Variance Estimation with Imputed Data

Imputation methods are used to fill in several types of missing data in the SPD. These methods are used to complete some item nonresponse and person-level nonresponse within households (Type Z nonresponse). Imputation fills in gaps in the data set and makes data analysis easier. It also allows more people to be retained as panel members for longitudinal analysis. The concern, however, is that imputation fabricates data to some degree. Treating the imputed values as actual values in estimating the variance of survey estimates leads to an overstatement of the precision of the estimates. It is important to recognize this fact when sizable proportions of values are imputed.

Section III: Working With the Public Use Microdata Files

Chapter 7. Using The 1997 SPD Experimental File

The 1997 SPD Experimental File is intended to be used as a research file. This is a calendar year file for 1996 and allows for longitudinal analyses through the ability to match to the Survey of Income and Program Participation (SIPP) 1992 and 1993 Panel Longitudinal files, using the following match keys: SIPP panel number, sample unit identification number, person number, and entry address ID. The data were collected using a modified Current Population Survey March questionnaire.

Subject matter

The file contains basic demographic, social, and economic characteristics data for calendar year 1996 for each member of the households selected from the 1992 and 1993 SIPP Panels. These include age, sex, race (White; Black; American Indian, Eskimo, and Aleut; Asian or Pacific Islander; and Other), ethnic origin (23 categories including 7 Hispanic origin categories), marital status, household relationship, education, and veteran status. Limited data are provided on housing unit characteristics such as units in structure and tenure.

This file, also known as the SPD Research File, provides monthly labor force data, but in addition, provides supplemental data on work experience, income and noncash benefits. Comprehensive work experience information is given on the employment status, occupation, and industry of people 15 years old and over. Additional data for people 15 years old and older are available concerning weeks worked and hours per week worked, reason not working full time, total income and income components. Data on employment and income refer to the preceding year, although demographic data refer to the time of the survey.

This file also contains data covering nine noncash income sources: food stamps, school lunch program, employer-provided group health insurance plan, employer-provided pension plan, personal health insurance, Medicaid, Medicare, CHAMPUS or military health care, and energy assistance.

Geographic Coverage

State codes are shown except for nine states that are identified in three groups. The sample was not designed to produce state or MSA/CMSA level estimates. State codes are primarily useful in relating a respondent's recipiency of benefits to welfare reform thresholds that may vary from state to state.

Identification Number System and Match Key Variables

The 1997 SPD identification scheme uses match key variables designed to uniquely identify individuals, provide a means of linking data for the same individuals over time, and grouping individuals into households and families over time. The various components of the identification scheme are listed below:

Variable	1997 SPD	1992/1993 SIPP		
		Core	Topical	Longitudinal
SIPP Panel	SIPP_PNL	Panel		
Sample Unit Identification Number	PP_ID	SUID	ID	PP_ID
Entry Address ID	PP_ENTRY	ENTRY	ENTRY	PP_ENTRY
Person Number	PP_PNUM	PNUM	PNUM	PP_NUM

The SIPP panel number identifies the panel in which the respondent participated. For the 1997 SPD, the sample person should either have an entry of 1992 or 1993 for their panel number. For the 1992/1993 SIPP, the panel number will be 92 or 93.

The sample unit identification number was created by scrambling together the PSU, segment, and serial numbers used for Census Bureau administrative purposes. This identifier is constructed in the same manner as the 1992 and 1993 SIPP panel files, to enable matching to these files.

The entry address ID represents the address of the person at the time he or she was first interviewed and does not change even if the person moves. It is used in conjunction with the person number to uniquely identify people within the sample unit. This variable is the number 011 for all original sample members. For additional sample people, this can be 011 or greater, depending on the address ID of the unit that the new sample person joined.

The first two digits of the person number signify the SIPP wave during which the individual first entered the sample or when an interview took place at a particular address (usually 01 for Wave 1, but not always). Person numbers 0101, 0102, etc., are assigned in Wave 1; 0201, 0202, etc., are assigned to people added to the roster in Wave 2, and so forth. This number is not changed or updated, regardless of moves in subsequent waves.

The sample unit identification number and entry address ID code uniquely identifies each household in any given wave. The sample unit identification number can link all households in subsequent waves back to the original Wave 1 household.

Caution Statement

The SPD Bridge file was processed using programs for the March Current Population Survey Demographic Supplement. These data have undergone limited editing and review and should be used carefully. It will appear to be a cross-section snapshot of the population, like the March CPS. It is important to emphasize that this is not the case. First of all, the underlying set of individuals that receive positive weights on this file was initially representative of the population in 1992 or 1993, when the sample was originally selected. It omits children born later, as well as

persons who entered the SIPP/SPD universe subsequent to the first wave of their respective SIPP panels. Secondly, there has been differential nonresponse over the life of the SIPP and SPD Bridge surveys.

Weighting compensates for some of the bias, but may not totally eliminate its effect on estimates. It is simply one segment in the array of longitudinal data being produced for SPD. The weights are longitudinal and only valid to estimate 1996 characteristics of the 1992/1993 SIPP cohorts. Children born since the beginning of the SIPP panel do not receive longitudinal weights in this first file so the distribution of population by age will be skewed toward the older population. However, users can construct an approximate weight for these children by assigning the weights of the mother to each child.

There are several relationship variables on the person record. Only the A_EXPRRP variable has been reviewed and edited. A_EXPRRP is the "edited relationship to reference person." We recommend using A_EXPRRP in your analysis and not using, or using with caution, the following relationship variables: A_FAMREL, A_PFREL, HHDREL, FAMREL, and HHDFMX. When matching to the 1992 or 1993 SIPP panel, core, or topical module files, note the field lengths differences in the following variables:

Field Length Differences							
	SIPP 1992 Panel Files			SIPP 1993 Panel Files			1997 SPD File
Short Variable Description	Panel	Wave	Topical Modules	Panel	Wave	Topical Modules	
Person Number	4	3	3	3	3	3	4
Entry Address ID	3	2	2	2	2	2	3
Address ID	3	2	2	2	2	2	3

File Structure

There is a household record for each household or group quarters. The household record is followed by one of three possible structures:

A. If the household contains related people and is not a group quarters household:

1. The family record appears next followed by person records for members of the family who are not also members of a related subfamily. The person records would be ordered: family householder, spouse of family householder, children in the family, and other relatives of the family householder.

2. The above records may be followed by one or more related subfamily records, each related subfamily record being followed immediately by person records for members of that related subfamily. The person records would be ordered: reference person of the related subfamily, spouse of subfamily reference person, and children of subfamily reference person.

3. The above records may be followed by one or more unrelated subfamily records, each unrelated subfamily record being followed immediately by person records for members of that unrelated subfamily. The person records would be ordered: unrelated subfamily reference person, spouse of subfamily reference person, and children of subfamily reference person.

4. The above records may be followed by one or more people living with nonrelatives "family" records, each to be followed by the person record for the unrelated individual it represents.

B. If the household contains a householder with no relatives and is not a group quarters household:

1. The family record for the nonfamily householder is followed immediately by the person record for that nonfamily householder.

2. These records may be followed by one or more unrelated subfamily records, each unrelated subfamily record being followed immediately by the person records for members of that unrelated subfamily. The person records would be ordered as described in A-3 above.

3. These records may be followed by one or more family records for people living with nonrelatives, each person living with nonrelatives family record being followed immediately by the person record for that person living with nonrelatives.

Topcoding of Income Variables

To protect against the possibility that a user might recognize the identity of an SPD respondent with a very high income, income from every source is topcoded so that no individual income amounts above $100,000 are revealed.

Summary income figures are simple sums of the components shown on the file after topcoding, and are not independently topcoded. Thus, a person with high income from several sources (jobs, businesses, property) could have aggregate monthly income well over the topcode for each source. Families and households with a number of high income members could theoretically have aggregate income shown well over $100,000.

Estimation of Person Characteristics

Some basic types of SPD Bridge longitudinal estimates that can be constructed using SPD

Bridge longitudinal weights are described below in terms of estimated numbers. Of course, more complex estimates, such as percents, averages, ratios, etc., can be constructed from the estimated numbers. The fullest potential of the SPD data is achieved when data users match SPD Bridge data to the 1992 and 1993 SIPP longitudinal panels. SPD Bridge longitudinal weights can be used to construct the following types of longitudinal estimates:

1. The number of people who have ever experienced a characteristic or situation during a given period of time.

2. The amount of a characteristic accumulated by people during a given time period.

Since the Bridge survey is one component of a longitudinal survey, young children will not have weight—due to the fact they were born after the inception of the panel. For the 1992 panel, children under 5 will not have a longitudinal weight. For the 1993 panel, children under 4 will not have a longitudinal weight. If users wish to explore estimates of young children, then an exploratory weight, assigning the mother's weight to each child without their own longitudinal weight may be used. However, the effect of these exploratory weights on estimates is unknown. Users should use extreme caution when interpreting results using these weights.

Longitudinal Research Using This File

The SPD is designed exclusively to support longitudinal analysis of the impact of welfare reform. The data on the 1997 experimental file can be linked to the 1992 and 1993 SIPP files using the identification number system and match key variables discussed earlier in this chapter.

The longitudinal weight is assigned to each original sample member that participated in at least Waves 1 and 9/10 of the 1992 and 1993 SIPP Panels. Three-fourths of the 1992 SIPP sample were eligible for ten waves.

The SPD longitudinal data represents the behavior and characteristics of people in two fixed cohorts over a period up to 10 years. One cohort represents the population as it existed in March 1992 (from the 1992 panel of the SIPP) and the other population as of March 1993 (from the 1993 panel of the SIPP). This is not a traditional longitudinal survey in that it does not repeat the same measure throughout the period.

The core information common throughout the data collection consists of basic demographics, labor force activity, income, and program participation. However, the reference periods, recall length, and question phrasing vary over the three components of this survey (1992/1993 SIPP, 1997 SPD, and 1998-2002 SPD). Since common elements of these three components correspond to the March CPS, variable names and reference periods for this file are based on the public-use March CPS data.

As noted above, this file does not support cross-sectional estimates of the population in the spring of 1997, even though it looks a lot like the March 1997 CPS. It is designed only to be used in conjunction with earlier SIPP data from the 1992 and 1993 longitudinal files, as extensions of the reference period covered in those surveys.

Chapter 8. Using The Unedited 1998 Calendar Year File

The 1998 SPD is an unedited file that provides socioeconomic data for calendar year 1997. It is intended as an *experimental file* to allow experienced users to conduct longitudinal analyses by matching to the SPD 1997 Bridge and the Survey of Income and Program Participation (SIPP) 1992 and 1993 Panel Longitudinal files. To link the files, at a person level, users will need to match on the following keys: SIPP panel number, sample unit identification number, person number and entry ID. (See table on the following page).

For the 1997 SPD Bridge survey, the regional office staff interviewed 29,619 households from retired 1992 and 1993 SIPP panels. From the 29,619 households interviewed in 1997, we selected a subsample of 19,129 households for the continuing SPD sample because of budget limits. Low income households and households with children were selected with certainty or near certainty. Middle and high income households without children had a 1 in 4 chance of being selected for the sample.

Subject Matter

The file contains basic demographic, economic, and social characteristics data for each member of the households. Although there is significant content overlap with that of the 1997 SPD Bridge Survey, the actual questions changed substantially. The subject matter is described as follows:

Demographic:

Data on age, sex, race, ethnic origin, marital status, household relationship, education, and veteran status. Demographic data refer to the time of the interview.

Economic:

Work Experience - Comprehensive work experience information is given on the employment status, occupation, industry, weeks worked and hours per week worked, total income and income components for people 15 years old and older. Most data on employment refer to the preceding year. However, some questions refer to the week before the interview (current employment status).

Income - Data covering income sources such as income from jobs, net income from businesses, farm or rent, pensions, dividends, interest, and social security payments. Data on income refer to the preceding year.

Noncash benefits - Data covering noncash income sources such as food stamps, the school lunch program, employer-provided group health insurance plan, employer-provided pension plans, personal health insurance, Medicaid, Medicare, CHAMPUS or military health care, and energy assistance. Most data on noncash benefits refer to the preceding year. However, some questions refer to the week before the interview.

Social:

Adult - Data on adult-related issues such as educational enrollment, work training, functional limitation and disability, health care utilization, health insurance, and food security. The data do not always reference the same period.

Children - Data on child-related issues such as: school enrollment, enrichment activities, disability, health care utilization, mother's work schedule, child care, child support agreement, and contact with absent parent. The data do not always reference the same period.

Geographic Coverage

The geographic coverage area is the United States. The file contains codes for 41 individual states, plus the District of Columbia. *However, the sample is not designed to produce state estimates***.** The SPD sample in the nine remaining states are identified in three groups for confidentiality reasons. The three groups are as follows: Maine and Vermont; Iowa, North Dakota, and South Dakota; and Alaska, Idaho, Montana, and Wyoming.

Technical Description

The file is a rectangular person-level file. Household-level variables are included on the record for every person in the household. There are no family-level variables in the file. There are 56,497 records.

Identification Number System and Match Key Variables

The 1998 SPD identification scheme uses match key variables designed to uniquely identify individuals, provide a means of linking data for the same individuals across files, and grouping individuals into households and families across files over time. The various components of the identification scheme are listed below:

Variable	1998 SPD	1992/1993 SIPP		
		Core	Topical	Longitudinal
SIPP Panel	SIPP_PNL	Panel		
Sample Unit ID Number	PP_ID	SUID	ID	PP_ID
Entry Address ID	PP_ENTRY	ENTRY	ENTRY	PP_ENTRY
Person Number	PP_PNUM	PNUM	PNUM	PP_PNUM

The SIPP panel number identifies the panel in which the respondent participated. For the 1998 SPD, the sample person should either have an entry of 1992 or 1993 for their panel number. For the 1992/1993 SIPP, the sample person should have a 92 or 93 for their panel number.

The sample unit identification number was created by scrambling together the primary sampling unit, segment, and serial numbers used for Census Bureau administrative purposes. These identifiers are constructed in the same manner as the 1992 and 1993 SIPP panel files, to enable matching to these files.

The entry address ID represents the address of the person at the time he or she was first interviewed and does not change even if the person moves. It is used in conjunction with the person number to uniquely identify people within the sample unit. This variable is the number 011 for all original sample members. For additional sample people, this can be 011 or greater, depending on the address ID of the unit that the new sample person joined. For example, a person who moves into a household with an ADDIDU8 of 011 will receive a PP_ENTRY of 011. Whereas, a person who enters a household spawned in 1998 (ADDIDU8=121) will have a PP_ENTRY of 121.

The person numbers represent the wave the person entered the sample. The first two digits of the person number signify the SIPP wave during which the individual first entered the sample or when an interview took place at a particular address (usually 01 for Wave 1, but not always). Person numbers such as 0101 and 0102 are assigned in Wave 1 of the SIPP. Person numbers such as 0201 and 0202 are assigned to people added to the roster in Wave 2 of the SIPP. People added to the roster in the 1998 SPD have person numbers 1201 and 1202 and following sequentially as needed.

Use the following variables to match back to the SIPP at the person level:

SIPP Panel Number	SIPP_PNL
Sample Unit Identification Number	PP_ID
Entry Address ID	PP_ENTRY
Person Number	PP_PNUM

Caution Statement

The 1998 SPD calendar year file data have undergone no editing and limited review and should be used carefully. A note of caution about this calendar year file. The file appears to be a cross-sectional snapshot of the population, but it is not. This file represents one segment in an array of longitudinal data being produced for the SPD. The weights in the SPD 1998 calendar year file are longitudinal, therefore, they are only valid for estimates of the characteristics of the cohorts of people who were represented by the original sample member from the SIPP 1992 and 1993 panels. An original sample member is a sample person who was a self or proxy respondent in an interviewed household in Wave 1 of the SIPP 1992 or 1993 panel. In the SPD 1998 calendar year file, the person number of an original sample member is in the 100 level. Children born since the beginning of the SIPP panel do not receive longitudinal weights in this first file. Consequently, the distribution of population by age will be skewed toward the older population. Users can construct an approximate weight for these children by assigning the weights of the mother to each child.

When matching to the 1992 or 1993 SIPP panel, core, or topical module files, note the field length differences in the following variables:

Field Length Differences							
	SIPP 1992 Panel Files			SIPP 1993 Panel Files			1998 SPD File
Short Variable Description	Panel	Wave	Topical Modules	Panel	Wave	Topical Modules	
Person Number	4	3	3	3	3	3	4
Entry Address ID	3	2	2	2	2	2	3

Edits

The purpose of this file is to provide policy analysts with timely data. To expedite the process, the variables in this data set are unedited.

Recodes were performed only for the following purposes:

- To provide a single economic estimate
- To maintain confidentiality of response by moving low-response items to an "other" category

The unedited variables have:

- "D" and "R" codes for Don't Know and Refused
- Blanks for non-response

C Not been verified for consistency

Topcoding of Variables

To protect against the possibility that a user may recognize the identity of an SPD respondent with a very high income, income from every source is topcoded so that no individual amounts above $100,000 are revealed. This topcode amount is consistent with the topcoding for the 1992 and 1993 SIPP panels. Other economic variables are topcoded at the 97 percentile level, meaning the top 3 percent of values are not disclosed. Variables that have been topcoded will have a "T" in the second to the right position.

The SPD topcodes age by bottom coding year of birth. For the 1998 SPD file no age will be older than 88.

Estimation of Person Characteristics

Some basic types of the 1998 SPD longitudinal estimates, which can be constructed using the 1998 SPD longitudinal weights, are described below in terms of estimated numbers. Of course, more complex estimates, such as percents, averages, and ratios can be constructed from the estimated numbers. The fullest potential of the SPD data is achieved when data users match the 1998 SPD data to the SPD Bridge data and the 1992 and 1993 SIPP longitudinal panel data to form a full set of longitudinal data. The 1998 SPD longitudinal weights can be used to construct the following types of longitudinal estimates:

1. The number of people who have ever experienced a characteristic or situation during a given time.

2. The amount of a characteristic accumulated by people during a given time.

In this file, young children born into the household after the Wave 1 interview will not have weight. For the 1992 panel, children under 6 will not have a longitudinal weight. For the 1993 panel, children under 5 will not have a longitudinal weight. If users wish to explore estimates of young children, then an exploratory weight, assigning the mother’s weight to each child without their own longitudinal weight may be used. However, the effects of these exploratory weights on estimates is unknown. Users should use extreme caution when interpreting results using these weights. Future files will contain more carefully constructed weights for these children.

Chapter 9. Using the Longitudinal Files

This chapter describes using the Survey of Program Dynamics longitudinal files. A valuable companion volume for analysts using the SPD longitudinal files is the *Survey of Program Dynamics (SPD) First Longitudinal File Technical Documentation*.

The first SPD Longitudinal file is a fully edited file that provides socioeconomic data for calendar years 1992–1997, except 1995. It is intended for longitudinal analyses of effects of welfare reform on individuals, families and households. The file can be linked to the SPD 1998, the SPD 1997 Bridge and the 1992 and 1993 SIPP panel, core wave, and topical module files.

The file contains basic demographic, economic, and social characteristics data for each member of the households for four or five years depending on the panel. The 1992 panel contains data for 1992, 1993, 1994, 1996 and 1997. The 1993 Panel contains data for 1993, 1994, 1996 and 1997. There is no data available for 1995 because SIPP did not complete a year's worth of data collection in either panel.

The SPD data represent the behavior and characteristics of people in two fixed cohorts. One cohort represents the population as it existed in March 1992 from the 1992 panel of the SIPP and the other population as of March 1993 from the 1993 panel. This is not a traditional longitudinal survey in that it does not repeat the same measure throughout the period. Each round of the SPD interviewing, beginning with the Bridge in 1997, does not represent cross-sectional snapshots of the U.S. population. It does offer insight into what the current condition is of the people in the U. S. population in the early 1990s just prior to welfare reform.

The core information common throughout the data collection (although reference periods and question phrasing vary) consists of basic demographics, labor force activity, income, and program participation. The longitudinal file consists of data collected using three different instruments, each with variations in wording and context.

For the sake of consistency, data for 1992 through 1994 has been converted from a monthly accounting period (reflecting how it was collected in the SIPP) to an annual accounting period (to match data collected for 1996 and 1997). This conversion was carried out by the Urban Institute, under contract to the Social Security Administration. The *Survey of Program Dynamics First Longitudinal File Technical Documentation* includes an appendix which cross references SIPP variables to CPS variables.

Demographic Information

The SPD longitudinal file includes data on age, sex, race, ethnic origin, marital status, household relationship, education, and veteran status. Demographic data refer to the time of the interview for years 1996 and 1997 and December for years 1992, 1993, and 1994.

Economic Information

Work Experience - Comprehensive work experience information is given on the employment status, occupation, industry, weeks worked and hours per week worked, total income, and income components for people 15 years old and older. The data on employment refer to the preceding year.

Income - Data covering income sources such as income from jobs, net income from businesses, farm or rent, pensions, dividends, interest, and social security payments. Data on income refer to the preceding year.

Noncash benefits - Data covering noncash income sources such as food stamps, the school lunch program, employer-provided group health insurance plan, employer-provided pension plans, Medicaid, Medicare, CHAMPUS or military health care, and energy assistance. Most data on noncash benefits refer to the preceding year. However, some questions refer to the week before the interview.

Geographic Coverage

The geographic coverage area is the United States. The file contains codes for 41 individual states, plus the District of Columbia. *However, the sample is not designed to produce state estimates***.** The SPD sample in the nine remaining states are identified in three groups for confidentiality reasons. The three groups are as follows: Maine and Vermont; Iowa, North Dakota, and South Dakota; and Alaska, Idaho, Montana, and Wyoming.

Identification Number System/Match Key Variables

For the longitudinal file, the SPD identification scheme uses match key variables designed to uniquely identify individuals, provide a means of linking data for the same individuals across files, and grouping individuals into households and families across files over time. The various components of the identification scheme are listed below:

Variable	1998 SPD Longitudinal	1992/1993 SIPP		
		Core	Topical	Longitudinal
SIPP Panel Number	SIPP_PNL	Panel		
Sample Unit ID Number	PP_ID	SUID	ID	PP_ID
Entry Address ID	PP_ENTRY	ENTRY	ENTRY	PP_ENTRY
Person Number	PP_PNUM	PNUM	PNUM	PP_PNUM

The SIPP panel number identifies the panel in which the respondent participated. On the SPD longitudinal file, the sample person should either have an entry of 1992 or 1993 for their panel number. On the 1992 or 1993 SIPP, the sample person should either have an entry of 92 or 93.

The sample unit identification number was created by scrambling together the primary sampling unit, segment, and serial numbers used for Census Bureau administrative purposes. These identifiers are constructed in the same manner as the 1992 and 1993 SIPP panel files, to enable matching to these files. To uniquely identify a household, you must use the sample unit identification (ID) number and the address ID code. The sample unit identification number and

the address ID can be used to link all households back to the original household.

The entry address ID represents the address of the person at the time he or she was first interviewed and does not change even if the person moves. It is used in conjunction with the person number to uniquely identify people within the sample unit. This variable is the number 011 for all original sample members. For additional people, this can be 011 or greater depending on the address ID of the unit that the new sample person joined. For example, a person who moves into a household with an ADDIDE8 of 011 will receive a PP_ENTRY of 011. Whereas, a person who enters a household spawned in 1998 (ADDIDE8=121) will have a PP_ENTRY of 121.

The person number represents when the person entered the sample. The first two digits of the person number signify the SIPP wave or the SPD year during which the individual first entered the sample or when an interview took place at a particular address (for example, 01 for Wave 1). Person numbers such as 0101 and 0102 are assigned in Wave 1 of the SIPP. Person numbers such as 0201 and 0202 are assigned to people added to the roster in Wave 2 of the SIPP. People added to the roster in the 1998 SPD have person numbers 1201 and 1202 and following sequentially as needed. People added to the roster in the 1997 SPD Bridge have person numbers 1101 and 1102 and following sequentially as needed.

Households are defined at each cross-sectional time point in this file: 1992, 1993, 1994, 1997, and 1998. If you would like to look at the household configuration in any given year, use the IHHKEY variable appropriate for that year. For example, to look at 1997 household structure, use IHHKEY97. The IHHKEY variable is a concatenation of SIPP_PNL; PP_ID and ADDID for the specific year. For example, IHHKEY97 is SIPP_PNL; PP_ID and ADDIDE7.

Caution Statement

When matching to the 1992 or 1993 SIPP panel, core, or topical module files, note the field length differences in the following variables:

Field Length Differences							
	SIPP 1992 Panel Files			SIPP 1993 Panel Files			1998 SPD File
Short Variable Description	Panel	Wave	Topical Modules	Panel	Wave	Topical Modules	
Person Number	4	3	3	3	3	3	4
Entry Address ID	3	2	2	2	2	2	3

Topcoding of Variables

To protect against the possibility that a user may recognize the identity of an SPD respondent with a very high income, income from every source is topcoded so that no individual amounts above $100,000 are revealed. This topcode amount is consistent with the topcoding for the 1992 and 1993 SIPP panels. Other economic variables are topcoded at the 97 percentile level, meaning the top 3 percent of values are not disclosed. Variables that have been topcoded will have a "T" in the second to the right position. NOTE: Aggregate amounts (PTOTVLR, PERNVLR, etc.) use topcoded amounts as input.

Age is topcoded by bottom coding year of birth. For the First Longitudinal SPD file no age will be older than 88.

Estimation of Person Characteristics

For the estimation of the person characteristics in the SPD universe, the final longitudinal weights of the sample people in the first SPD longitudinal file can be used. Hereinafter, the term "*the final longitudinal weights of the sample people in the first SPD longitudinal file*" will be simply referred to as "*the longitudinal person weights.*" Some basic types of longitudinal estimates (using the first SPD longitudinal file) can be constructed using the longitudinal person weights are described below in terms of estimated numbers.

- The number of people who have ever experienced a characteristic or situation during a given period of time (for example, the number of people who experience unemployment during 1997). To construct such an estimate, sum the weights over all people who possessed the characteristic of interest at some point during the time period of interest.

- The amount of a characteristic accumulated by people during a given time period (for example, the amount of unemployment compensation received by unemployed people during 1997). To construct such an estimate, compute the product of the weight times the amount of the characteristic and sum this product over all appropriate people.

- The average number of consecutive months or years of possession of a characteristic (i.e., the spell length for a characteristic.) For example, one could estimate the average spell of unemployment that elapsed before a person found a new job. (*Note that the first SPD longitudinal file provides the employment data only in terms of week numbers with and without employment in a given year. Thus, for calculation the average unemployment spell length in a time period of interest, the data user needs to match the sample person's record back to the one on the SIPP longitudinal file to determine the number of spells in the time period or needs to make some justifiable approximation on the number of unemployment spells within the time period of interest.*) To construct such an estimate, first identify the sample people possessing the characteristic at some point during the time period of interest. Then, create two sums of these (longitudinal person) weights: Sum 1 is sum of the products of the weights times the number of months (or years) the spell lasted, and Sum 2 is the sum of the weights only. The average spell length in months (or years) is given by Sum 1 divided by Sum 2. A person who experienced two spells during the time period of interest would be treated as two people and appear twice in Sum 1 and Sum 2. An alternate method of calculating the average can be found in the section "Standard Error of a Mean or an Aggregate."

  Note that spells extending before or after the time period of interest are cut off (censored) at the boundaries of the time period. If they are used in estimating average spell length, a downward bias will result.

- The number of year-to-year changes in the status of a characteristic (i.e., number of transitions) summed over every set of two consecutive years during the time of interest. To construct such estimate, sum the longitudinal person weights each time a change is reported between two consecutive years during the time period of interest. For example, to estimate the number of people who changed from receiving any public assistance in 1996 to not receiving in 1997, add together the longitudinal person weights of each person who had such a change.

Longitudinal Research Using This File

The SPD is designed exclusively to support longitudinal analysis of the impact of welfare reform. The First Longitudinal SPD data can be linked to the 1992 and 1993 SIPP Panel and Cross-sectional files, the 1997 SPD Bridge, and the 1998 SPD file using the following variables:

SIPP Panel Number	SIPP_PNL
Sample Unit Identification Number	PP_ID
Entry ID	PP_ENTRY
Person Number	PP_PNUM

Person-level analysis

In order to match the First Longitudinal data file with the unedited SPD 1998 public use file, or the SIPP core or topical module 1992 or 1993 data not included on the file, create the following matchkey to use when merging the longitudinal file with 1992, 1993 or 1998 data. In the SPD

longitudinal file, concatenate the following variables: SIPP_PNL, PP_ID, PP_PNUM, and PP_ENTRY. For example, in SAS:

matchkey=sipp_pnl||pp_id||pp_pnum||pp_entry;

Matching to the SIPP core data

To create this same matchkey in the 1992 or 1993 SIPP core data, use the following variables (data can be extracted using the Data Extraction System on the Census Bureau's website): PANEL, SUID, PNUM, and ENTRY. To create this matchkey in SAS, for example, do the following:

matchkey='19'||panel||suid||'0'||pnum||'0'||entry;

The addition of the extra characters is necessary in order to match the length of the variables in the SPD longitudinal file.

Matching to the SIPP topical module data

To create the same matchkey in the 1992 or 1993 SIPP topical module data (this data can also be extracted using the Data Extraction System on the Census Bureau's website), use PANEL, ID, PNUM, and ENTRY. ENTRY and PNUM are both one character shorter in the topical modules than in the longitudinal file. In order to make them match the variables in the longitudinal file, you can create the matchkey as in the following example for SAS:

matchkey=panel||id||'0'||pnum||'0'||entry;

Matching to the 1998 SPD unedited public use file

To create the same matchkey in the public use 1998 SPD file, concatenate the following variables: SIPP_PNL, PP_ID, PP_PNUM, and PP_ENTRY.

This matchkey uniquely identifies each individual on the file. Using the matchkey, it is possible to link the person's record from the longitudinal file with the core and topical module SIPP data for 1992 and 1993, as well as the public use unedited 1998 SPD file, and after they are released, the 1999 and 2000 cross-sectional SPD files.

Household-level Analysis

Households are defined at each cross-sectional time point in this file: 1992, 1993, 1994, 1997, and 1998. If you would like to look at the household configuration in any given year, use the IHHKEY variable appropriate for that year. For example, to look at 1997 household structure, use IHHKEY97. The IHHKEY variables are a concatenation of SIPP_PNL, PP_ID and ADDIDE for the specific year. For example, IHHKEY97 is SIPP_PNL, PP_ID and ADDIDE7.

Family-level Analysis

Like households, families are also defined each year. In order to look at everyone who is in the same family for a given year, concatenate the IHHKEY variable for that year with the variable indicating family number. For example, to look at families in 1993, concatenate IHHKEY93 with FAMNUME3.

Weighting

A longitudinal weight is assigned to original sample members with full panel weights in the 1992/1993 SIPP file who were successfully interviewed in 1998. Note the full panel weights on the SIPP files were assigned to original sample members who were interviewed for the entire time they remained in the SIPP universe or who had at most one missing interview bounded by successful interviews.

Obtaining Access to SAQ Data

The SAQ data will only be available through the Census Bureau's Research Data Centers. Contact the Research Data Center staff for the requirements for reviewing the SAQ data.

Chapter 10. Analytic Uses of the Data

One attractive feature of the Survey of Program Dynamics (SPD) is that it produces ten years of longitudinal data with welfare reform legislation enacted near the middle of those ten years. This places analysts using the SPD in a unique position to evaluate the impacts of welfare reform. While the construction of these data results in many advantages for the researcher, it also introduces special challenges. The combination of data from the 1992/1993 SIPP, the 1997 SPD Bridge, and the 1998-2002 SPD raises several issues of concern for researchers—including recall periods, missing data and varying levels of aggregation. Although these issues are of concern, they are not crippling to would-be analysts.

This chapter contains information on conducting analysis using the SPD longitudinal data. First, we discuss an example using food stamp receipt, addressing data concerns as well as possible solutions to problems encountered. Then, we describe characteristics of spell data on the 1998 SPD longitudinal file. Finally, we provide instruction for various levels of analysis: person-level, family-level, and household-level. An additional example of using the first longitudinal file to measure the effect of welfare reform (Hess 2001) is available on the Internet at this address: <http://www.census.gov/prod/2001pubs/spd2001-1.pdf>.

Food Stamp Receipt

Food stamp receipt has decreased since the passing of welfare reform legislation in 1996. The debate remains open for several issues for which an analysis examining individual receipt patterns over time might offer some insight. Two such issues are "cream-skimming" and the effect of time limits. "Cream-skimming" addresses the question of whether or not the declining food stamp caseload was solely driven by individuals with a briefer history of food stamp receipt. Has welfare reform targeted the "easiest" cases or have individuals with persistent food stamp receipt been equally affected? The effect of time limits for food stamp receipt remains an open question. Are receipt spells becoming shorter? Are individuals "stockpiling" their eligibility or have receipt patterns remain relatively unchanged? The unique structure and timing of the SPD might offer insight to the answers to these and other policy questions. A discussion of a concrete example, such as food stamp receipt, can also illuminate the general data issues of recall periods, missing data and varying levels of aggregation.

The components comprising the SPD (the 1992/1993 SIPP, the 1997 SPD Bridge, and the 1998-2002 SPD), have different recall periods and levels of aggregation. Respondents in the SIPP panels were interviewed three times per year and, as a result, faced a recall period of four months. Respondents in the 1997-2002 SPD are interviewed only once per year and may face recall periods up to fifteen months. Food stamp receipt is asked at the monthly level for both the SIPP panels and the 1998-2002 SPD. These responses may be aggregated by the data analyst to obtain annual totals. The SPD Bridge file has food stamp receipt information at only the annual level. Receipt is summed for the entire year and specific months of receipt are not available. Data from late 1995 is missing from the SIPP panels, and the amount of missing data depends upon the rotation group of the respondent. For more information regarding rotation groups, see the *SIPP Users Guide*.

Consider the example of examining patterns of food stamp receipt before and after welfare reform. Individual analysts may opt to focus on total months of receipt per calendar year (in a sense treating each year as one observation) or look at individual months of food stamp receipt (treating each month as one observation). The SPD will report total number of months of receipt per year from 1992 to 2002 but will not differentiate which months receipt did or did not occur. If this level of analysis is sufficient then the only concern faced by the researcher is how to handle the missing data for a portion of 1995. Several options are available. One might simply use the partial count available for 1995 or treat all of 1995 as a missing observation. If one feels that any adjustment or imputation of food stamp receipt severely compromises the quality of the data, this may be the best option. An alternative is to conjecture that data obtained before and after the missing months sheds light as to the likely receipt for the missing months.

For example, suppose an individual received food stamps for all twelve months in 1994 and 1996. If the 1995 data show nine months of receipt with three months of missing data, it might be reasonable to assume that receipt would have occurred during the missing months. Other cases may involve more ambiguity and may require a greater level of an analyst's judgement.

In general, one can think of the following structure:

Let X = total number of months of food stamp receipt in 1994
Let Y = total number of months of food stamp receipt in 1995 (with missing data)
Let Z = total number of months of food stamp receipt in 1996

There are three possible cases:

1. $X = Z$. If assigning receipt to any, all or none of the missing months can result in $X = Y = Z$, then adjust the data to make all three equal.

2. $X < Z$. If assigning receipt to any, all or none of the missing can result in $X \leq Y \leq Z$, then adjust the data to fit that range. Whether the adjusted value of Y is closer to X or Z is left to the discretion of the analyst. One might consider examining receipt totals from 1993 and 1997 to better establish consistent patterns.

3. $X > Z$. If assigning receipt to any, all or none of the missing can result in $X \geq Y \geq Z$, then adjust the data to fit that range. Whether the adjusted value of Y is closer to X or Z is left to the discretion of the analyst. One might consider examining receipt totals from 1993 and 1997 to better establish consistent patterns.

In cases 2 and 3, if adjustments to Y cannot result in fitting into the desired range of values, then one might consider using the total for 1995 without making adjustments to 1995. Finally, one could probabilistically estimate receipt in any given month and then determine how many missing months are "likely" to have food stamp receipt.

Characteristics of Spell Data on the First Longitudinal File

As a longitudinal survey, one of the strong attributes of the SPD is to provide a collection of data that renders itself to the estimation of spell durations for participation in various transfer

programs and unemployment. The methodology for spell duration estimates using the SPD data generally entails the following three components:

- Non-sampling errors—particularly the bias induced by the seam phenomenon.
- Definitions of a spell.
- The statistical approaches used for the spell duration estimates.

This section does not discuss the methodology for spell duration estimates per se. The objective of this section is to discuss the characteristics of the spell data on the first longitudinal file associated with the spell duration estimates. The relationship between the spell data on the first longitudinal file and those on the SIPP Panel 1992 and 1993 longitudinal files is also included in the discussion.

All the time-varying data on the first longitudinal file are yearly instead of monthly like those on the SIPP Panel 1992 and 1993 longitudinal files. The yearly data on the first longitudinal file generally cover 1992, 1993, 1994, 1996 (the SPD Bridge), and 1997 (SPD 1998). For example, on the first longitudinal file the variable PAWMONE7 represents the number of months in 1996 in which a sample person received public assistance payments; the variable LKWKSE4 represents the number of weeks in 1994 in which a sample person was looking for work or on layoff from a job.

For the 1992, 1993, and 1994 data, the user can decompose the yearly data on the first longitudinal file into monthly data by linking the sample people back to the SIPP Panel 1992 and 1993 longitudinal files. For 1997 data, the yearly data on the first longitudinal file can be decomposed into monthly data; however, at present, these monthly data are available to the public only by special request to the Census Bureau. For the 1996 data, the yearly data cannot be directly decomposed into monthly data because the SPD Bridge did not ask the respondents for month by month recalls. Therefore, if needed, the user has to use an analytical approach to decompose the 1996 yearly data into the monthly data based on the monthly data for 1992, 1993, and 1994 on the SIPP Panel 1992 and 1993 Longitudinal Files, and 1997 monthly data from the SPD 1998 available for the cohort of sample people under consideration.

On the basis of the above discussion, if the first longitudinal file is used alone for spell duration estimates, the time unit of a spell duration may be more advantageously expressed in years and then the spell duration treated as a continuous yearly random variable instead of a discrete weekly or monthly variable. For example, “a sample person receiving 23 weeks of public assistance in 1997" will be converted to “a person receiving 23÷52 = 0.4423 years of public assistance in 1997.” Similar to the SIPP, the SPD sample data were subject to the preselected starting and ending points for data collection and recall period specified by the sample design. Consequently, the spells reported in the SPD panel (including the SIPP Panels 1992 and 1993) will generally cover the following four situations:

- A spell may start and end during the panel (an uncensored spell—a spell observed at its entirety).
- A spell may start during the panel and be still ongoing at the end of the panel (a right censored spell).
- A spell may start before the beginning of the panel and end during the panel (a left censored spell).

- A spell may start before the beginning of the panel and be still ongoing at the end of the panel (a doubly censored spell).

Since the SPD data collected prior to the SPD Bridge were extracted from the SIPP Panels 1992 and 1993, the SPD spell data inherently carried over a type of non-sampling error commonly referred to as "the seam effect." In the SIPP, the seam is the boundary between the four-month reference periods for interviews in successive waves of the panel. Namely, for participation in various programs, the number of spell starts or stops reported for the four-month recall (reference month one) was substantially higher than those reported for the one, two, or three month recalls (reference months four, three, and two). This is contrary to the expectation that, after the first wave, the distribution for reported spell starts or stops by month of recall is a uniform one—with approximately 25% of spell starts or stops being reported at each month of recall. As indicated in the *SIPP Quality Profile* (1998), the bias in the spell data due to the seam effect is significant in the SIPP panels and cannot be ignored in the spell duration estimates. In the SIPP, the cause of the seam bias in the spell data has not been identified with certainty, but it has been commonly suggested that questionnaire wording and design, length of recall, and the interaction between them play an important role. For the SPD, the seam effect between the combined SIPP Panels 1992 and 1993 and the SPD Bridge, and the SPD Bridge and the SPD 1998 on the spell data have not been studied.

Applications of the SPD Longitudinal Weights for Analyses

Each SPD sample person was assigned four weights: two are crude longitudinal panel weights (*LGTPERWT* on the 1997 SPD file, and *LGTPERW8* on the 1998 SPD file); the other two are the refined longitudinal panel weight (*SPDLNWGT)* and the longitudinal annual weight (ANNUALWT) on the SPD first longitudinal file. A sample person on the 1997 SPD file, the 1998 SPD file, and the SPD first longitudinal file will have either a positive weight or a zero weight assigned to *LGTPERWT*, *LGTPERWT*, *SPDLNWGT*, and *ANNUALWT* according to his or her longitudinal interview status (as described in Chapter 5). The SPD first longitudinal file contains annual data for 1992, 1993, 1994, 1996, and 1997 while the 1998 calendar year file contains only annual data for 1997. Therefore, by using the first longitudinal file to obtain data for longitudinal analyses, analysts can avoid the burden of linking files.

On the 1997 SPD file, the original sample members with positive longitudinal panel weights (*LGTPERWT* > 0) collectively provide a crude representation of the characteristics of the noninstitutionalized civilian population in March 1993 (the SPD panel universe) for the time span between 1992 and 1996. Similarly, the original sample members with *LGTPERW8* > 0 on the 1998 SPD file collectively provide a crude representation of the characteristics of the noninstitutionalized civilian population in March 1993 for the time span between 1992 and 1997. The weight, *LGTPERWT* or *LGTPERW8*, of a sample person quantitatively represents the number of people in the survey universe who have the demographic and economic characteristics similar to those of the sample person. To use the *LGTPERWT* or *LGTPERW8* for any estimates covering multiple years requires matching the sample persons on the 1997 SPD file or the 1998 SPD file back to the 1992/1993 SIPP longitudinal files. The crude longitudinal panel weights, *LGTPERWT* and *LGTPERW8*, were produced to be used for preliminary estimates and research at the early stage of the SPD when the SPD first longitudinal file and the refined longitudinal panel weight, *SPDLNWGT* were not available. However, the *LGTPERWT* and *LGTPERW8* are

superseded by the *SPDLNWGT* on the SPD first longitudinal file.

On the SPD first longitudinal file, the longitudinal panel weight, *SPDLNWGT*, should be used for any estimates covering multiple years within 1992 to 1997, and the longitudinal annual weight, *ANNUALWT*, should be just for any annual or calendar year estimates. However, the *SPDLNWGT* is also recommended to be used for any annual or calendar year estimates if the estimates do not concern the characteristics of the children born after the first interview of the 1992/1993 SIPP panels. Some caution should be taken when using the *ANNUALWT* for estimating the characteristics of children aged six and less. Because of the approach used to assign the weights to the sample children born after the first SIPP interview, the estimates for the children in this age group are generally 2.2 percent higher than the corresponding 1998 benchmark estimates. By race for the children in this age group, the estimates are 3.6 percent higher than the benchmark estimates for non-Black, and 5.4 percent lower the benchmark estimates for Black children. Since the 1992 data are available only for the sample units from the SIPP Panel 1992 (which is approximately half of the SPD sample size), the weights used any 1992 estimates must be twice the longitudinal weights on the file (i.e., 2 × *SPDLNWGT* or 2 × *ANNUALWT*.) The variances of the estimates for this year will need to be inflated by two as well.

All the weights, *LGTPERWT*, *LGTPERW8*, *SPDLNWGT*, and *ANNUALWT* can be used for the following three levels of analyses:

- Person-level analysis
- Family-level analysis
- Household-level analysis

Since all the four weights can be used in the same manner for above three levels of analyses; without the loss of generality, the discussion of the levels of analysis provided below will be made based only on the longitudinal panel weight, *SPDLNWGT* on the SPD first longitudinal file.

Person-Level Analysis

For longitudinal analysis at the person level, the sample person weights (*SPDLNWGT*) provided on the first longitudinal file can be used directly, as shown in the following illustration. Suppose you want to assess the poverty levels of the people in the SPD panel universe (the 1993 population) before and after welfare reform. The assessment can begin by constructing a transition matrix classifying how many people in the SPD panel universe retained or changed their original (1993) poverty status in 1997:

Poverty Status of People in the SPD Panel Universe (the 1993 population) in 1993 (before welfare reform) and 1997 (after welfare reform).		
	1993 Poverty Status	
	Not in Poverty (denoted by 0_)	In Poverty (denoted by 1_)

1997 Poverty Status	Not in Poverty (denoted by _0)	Cohort 00—People who were not in poverty in both 1993 and 1997 (i.e., stayed out of poverty).	Cohort 10—People who were in poverty in 1993 but were not in poverty in 1997 (i.e., left poverty).
	In Poverty (denoted by _1)	Cohort 01—People who were not in poverty in 1993 but were in poverty in 1997 (i.e., enter poverty).	Cohort 11—People who were in poverty in both 1993 and 1997 (i.e., stayed in poverty).

As indicated in the table above, the people in the SPD panel universe are classified into four cohorts:

- Cohort 00 consists of the people in the SPD panel universe who were not in poverty in 1993 and were also not in poverty in 1997 (i.e., stayed out of poverty).
- Cohort 10 consists of the people in the SPD panel universe who were in poverty in 1993 but were not in poverty in 1997 (i.e., left poverty).
- Cohort 01 consists of the people in the SPD panel universe who were not in poverty in 1993 but were in poverty in 1997 (i.e., entered poverty).
- Cohort 11 consists of the people in the SPD panel universe who were in poverty in 1993 and were also in poverty in 1997 (i.e., stayed in poverty).

Since the panel universe is adequately represented by the original sample persons on the SPD first longitudinal file who have a positive longitudinal panel weight (*SPDLNWGT* > 0), only these sample people need to be considered in estimating the numbers of people in Cohorts 00, 10, 01, and 11. To estimate the numbers of the people in each cohort, identify the family poverty status of the original sample persons (with positive *SPDLNWGT*) in 1993 and 1997 based on the family poverty status indicators (*FAMLISE3* and *FAMLISE7*, respectively). Suppose you define a low income family as "a family with the total family income below the low income threshold." Then, *FAMLISE3*=1 would imply the family is a low income family in 1992, and *FAMLISE7*=1 would imply the family is a low income family in 1993. A person living in a low income family in a given year is in poverty for that year, and not in poverty for that year otherwise. Assign the poverty status of a person as 1 if in poverty and 0 if not in poverty.

Based on the above definition of the poverty status of a person, classify the original sample persons (with positive *SPDLNWGT*) as belonging to Cohorts 00, 10, 01, 11—in accordance with their poverty statuses in 1993 and 1997. The estimate of the number of the people in each of the four cohorts in the SPD panel universe can be calculated by summing the weights (*SPDLNWGT*) of the original sample people in the same cohort.

The poverty levels of the people in the SPD panel universe (the 1993 population) before and after welfare reform can be assessed using the estimates of the number of the people in Cohorts 00, 10, 01, and 11. For example, if the estimate of the number of people in Cohort 10 (in poverty in 1993 but not in 1997) is statistically significantly larger than the estimate of the number of people in Cohort 01 (not in poverty in 1993 but in poverty in 1997), then you can infer that more people left poverty than entered poverty after the welfare reform. This suggests that the welfare reform has a positive effect in reducing the poverty level in the pre-welfare-reform population. (The statistical significance test for the comparison can be made using the procedure provided in

Chapter 6.)

Family-Level Analysis

While families are not defined longitudinally in the SPD, it is feasible to create a time series of family estimates based on these data. For analyses at the family level, the weight (*SPDLNWGT*) of the sample person who is the reference person of her/his family can be used to represent the weight of that sample family on the first longitudinal file. An illustration would be to suppose that a user wants to estimate the proportions of the low income families in 1994 and 1997 in the SPD panel universe. Based on the above discussion, the user can calculate the estimates based on the six step procedure provided below.

Step 1. Let F_{94} denote the 1994 estimate of the number of all the families in the SPD panel universe. As discussed above, the weight of a sample family is represented by the weight of the reference person of that sample family on the first longitudinal file. Therefore, F_{94} can be expressed as the sum of the weights (*SPDLNWGT*) of all the original sample members with positive weights who were the family reference people in 1994. A family reference person on the first longitudinal file can be identified by the categorical value of the variable *FAMRELE4* equal to one.

Step 2. Let F_{97} denote the 1997 estimate of the number of all families in the SPD panel universe. In the same manner as Step 1, F_{97} can be calculated as the sum of the weights (*SPDLNWGT*) of all the original sample members with positive weights who were the family reference people in 1997. A family reference person on the first longitudinal file can be identified by the categorical value of the variable *FAMRELE7* equal to one.

Step 3. Let F_{L94} denote the 1994 estimate of the number of low income families in the SPD panel universe. On the first longitudinal file, a low income family can be identified by the categorical value of the variable *FAMLISE4* equal to one. In the same token as Step 1, the weight of a low income family is represented by the weight of the reference person of that family. Thus, F_{L94} can be expressed as the sum of the weights (*SPDLNWGT*) of all the original sample members with positive weights who were the reference people (*FAMRELE4 = 1*) of a low income family (*FAMLISE4 = 1*) in 1994.

Step 4. Let F_{L97} denote the 1997 estimate of the number of low income families in the SPD panel universe. In the same manner as Step 3, F_{L97} can be expressed as the sum of the weights (*SPDLNWGT*) of all the original sample members with positive weights who were the reference people (*FAMRELE7 = 1*) of a low income family (*FAMLISE7 = 1*) in 1997.

Step 5. Let P_{L94} and P_{L97} be the 1994 and 1997 estimates of the proportions of the low income families among all the people in the SPD panel universe, respectively. By definition, P_{L94} and P_{L97} can be expressed in terms of F_{94}, F_{97}, F_{L94}, and F_{L97} (calculated in Steps 1 to 4) as follows.

$$P_{L94} = \frac{F_{L94}}{F_{94}} \qquad P_{L97} = \frac{F_{L97}}{F_{97}}$$

Step 6. A methodology for estimating the standard errors of the estimates F_{94}, F_{97}, F_{L94}, F_{L97}, P_{L94} and P_{L97}, and a methodology for testing the statistically significant difference between P_{L94} and P_{L97} are provided in Chapter 6.

Household-Level Analysis

Although households are not defined longitudinally in the SPD, it is feasible to create a time series of household estimates based on these data. For analyses at the household level, the weight (*SPDLNWGT*) of the sample person who is the reference person of the household can be used to represent the weight of that sample household on the first longitudinal file. An illustration would be to suppose that an analyst wants to estimate the 1994 and 1997 proportions of households headed by females with their own children, but with no spouse present—in a cohort of all the households headed by householders living with relatives in the SPD panel universe. The analyst can calculate the estimates based on the six steps below.

Step 1. Let H_{94} denote the 1994 estimate of the number of all the households headed by householders living with relatives in the SPD panel universe. As discussed above, the weight of a sample household is represented by the weight of the household reference person on the first longitudinal file. Thus, H_{94} can be expressed as the sum of the weights (*SPDLNWGT*) of all the original sample members with positive weights who were the household reference people living with relatives on the first longitudinal file in 1994. A household reference person living with relatives on the first longitudinal file can be identified by the categorical value of the variable *RRPE4* equal to one.

Step 2. Let H_{97} denote the 1997 estimate of the number of all the households headed by householders living with relatives in the SPD panel universe. In the same manner as Step 1, H_{97} can be calculated as the sum of the weights (*SPDLNWGT*) of all the original sample members with positive weights who were the household reference people living with relatives (*RRPE7*=1) on the first longitudinal file in 1997.

Step 3. Let H_{F94} denote the 1994 estimate of the number of the households headed by female householders with own children but with no spouse present. On the first longitudinal file, a female can be identified by the categorical value of the variable *SEX* equal to two, a householder (reference person) living with relatives in 1994 can be identified by the categorical value of the variable *RRPE4* equal to one, no spouse present in 1994 can be identified by the categorical value of the variable *MARITLE4* not equal to one or two, and having own children in 1994 can be identified by the categorical value of the variable *RRPE4* for someone in her household equal to five. Thus, H_{F94} can be expressed as the sum of the weights of all the original sample members with positive weights who were a female household reference person living with relatives but no spouse present and had own children in 1994.

Step 4. Let H_{F97} denote the 1997 estimate of the number of the households headed by female householders with their own children but with no spouse present. In the same manner as Step 3, H_{F97} can be calculated as the sum of the weights of all the original sample members with positive weights who were a female (*SEX*=2) household reference person living with relatives (*RRPE7*=1) but no spouse present (*MARITLE7* … 1 or 2) and had their own children (*RRPE7*=5 for someone in the household) in 1997.

Step 5. Let P_{H94} and P_{H97} be the 1994 and 1997 estimates of the proportions of the households headed by female householder with own children but with no spouse present in a cohort of all the households headed by householders living with relatives in the SPD panel universe, respectively. By definition, P_{H94} and P_{H97} can be expressed in terms of H_{94}, H_{97}, H_{F94}, and H_{F97} (calculated in Steps 1 to 4) as follows:

$$P_{H94} = \frac{H_{F94}}{H_{94}} \qquad P_{H97} = \frac{H_{F97}}{H_{97}}$$

Step 6. A methodology for estimating the standard errors of the estimates H_{94}, H_{97}, H_{F94}, H_{F97}, P_{H94} and P_{H97}, and a methodology for testing the statistically significant difference between P_{H94} and P_{H97} are provided in Chapter 6.

References

Creighton, K., K. King, and E. Martin. The Use of Monetary Incentives in Census Bureau Longitudinal Surveys. Paper presented at the Federal Committee on Statistical Methodology Statistical Policy Seminar, hosted by the Council of Professional Associations on Federal Statistics, 8-9 November, Bethesda, MD.

Fisher, G.1992. The Development and History of the Poverty Thresholds. *Social Security Bulletin*, 55: 3-14. <http://www.ssa.gov/history/fisheronpoverty.html>

Hess, J. 2001 Preparing to Measure Welfare Reform Using the Longitudinal Survey of Program Dynamics: 2001. SPD Analytic Report No. SPD-2001-1. U.S. Census Bureau. <http://www.census.gov/prod/2001pubs/spd2001-1.pdf>

Lamas, E., J. Tin, and J. Eargle. 1994. The Effect of Attrition on Income and Poverty Estimates from the Survey of Income and Program Participation. Paper presented at the Conference on Attrition in Longitudinal Surveys, 24-25 February, Washington, D.C.

Mack, S. and R. Petroni. 1994. Overview of SIPP Nonresponse Research Data. SIPP Working Paper No. 9414. U.S. Census Bureau.

Martin, E. and M. de la Puente. 1993. Research on Sources of Undercoverage Within Households. *American Statistical Association 1993 Proceedings of the Section on Survey Research Methods*, Alexandria, VA: American Statistical Association, pp. 1262-1267.

Petroni, R. 1991. Current Non-response Research for the Survey of Income and Program Participation. Paper presented at the Second International Workshop on Household Survey Non-response, October.

Singh, R. and R. Petroni. 1988. *Non-response Adjustment Methods for Demographic Surveys at the U.S. Bureau of the Census*. SIPP Working Paper No. 8823. U.S. Bureau of the Census.

U.S. Census Bureau. 1997. *Survey of Program Dynamics (SPD). 1997 Experimental File. Technical Documentation*. SPD-97.

——— 1998. *Survey of Income and Program Participation: SIPP Quality Profile*. 3rd ed.

——— 1998. *Survey of Program Dynamics (SPD). 1998 Public Use File. Technical Documentation*. SPD-98.

——— 2001. *Survey of Income and Program Participation Users' Guide*.

——— 2001. *Survey of Program Dynamics (SPD). First Longitudinal File. Technical Documentation*.

U.S. Department of Labor. Bureau of Labor Statistics. 2000. *Current Population Survey. Technical Paper 63. Design and Methodology*. <http://www.census.gov/prod/2000pubs/tp63.pdf>

Zabel, J. 1993. An Analysis of Attrition in the PSID and SIPP with an Application to a Model of Labor Market Behavior. SIPP Working Paper Series No. 9403. U.S. Census Bureau.

Appendixes

Acronyms and Abbreviations

AFDC	Aid to Families with Dependent Children
BLS	Bureau of Labor Statistics
CAPI	Computer-assisted personal interviewing
CHAMPUS	Civilian Health and Medical Program Uniformed Service
CMSA	Consolidated Metropolitan Statistical Area
CPS	Current Population Survey
DES	Data Extraction System
FERRET	Federal Electronic Research and Review Extraction Tool
FR	Field representative
GAO	General Accounting Office
GED	General equivalency diploma
GVF	Generalized variance functions
ISDP	Income Survey Development Program
LQ	Living quarters
MSA	Metropolitan Statistical Area
NHIS	National Health Interview Survey
NSAF	National Survey of American Families
OMB	Office of Management and Budget
PRWORA	Personal Responsibility and Work Opportunity Reconciliation Act
PSID	Panel Study of Income Dynamics
PSU	Primary Sampling Unit
RHC	Residential History Calendar
RO	Regional Office
SIPP	Survey of Income and Program Participation
SPD	Survey of Program Dynamics
SSI	Supplemental Security Income
TANF	Temporary Assistance for Needy Families
WIC	Women, Infants, and Children (nutrition program)
WPA	Work Projects Administration

Glossary

Address Unit. A person or group of persons living at the same address at the time of an interview. The address unit may consist of one person living alone, a group of unrelated individuals, or one or more families.

Cold Deck Imputation. Procedures which use group estimates (such as means) for a sample as a whole or for subgroups within it as the source of information for the values to assign to those cases for which data are missing. See also *logical imputation*, *hot deck imputation*, and *longitudinal edits*.

Cross-Sectional Survey. Data collected for a single time period from a single sample.

Data Editing. The use of related information to replace missing or inconsistent data in the survey. See also *imputation*.

Hot Deck Imputation. Statistical method used to replace missing values with data from records with similar characteristics. See also *logical imputation*, *cold deck imputation*, and *longitudinal edits*.

Household. People living in a housing unit at the time of an interview.

Housing Unit. Living quarters with its own entrance and cooking facilities.

Imputation. Procedures for replacing missing values with statistical estimates that are based on the best relevant information available. See also *logical imputation*, *hot deck imputation*, *cold deck imputation*, and *longitudinal edits*.

Imputation Flag. An identifier associated with a questionnaire item to indicate whether information has been imputed.

Item Nonresponse. A source of missing data that occurs when a respondent does not answer one or more questions.

Logical Imputation. A procedure for inferring a missing value, based on other characteristics on a person's record or within a household.

Longitudinal Edits. A procedure for assigning values based on previously collected data.

Longitudinal Survey. Data collected at different times over an extended period from a single sample.

Mover. An original sample member who changed residence during the life of a panel.

Original Sample Member. A person who was interviewed in the first wave of a panel.

Panel. All households selected for a single sample.

Primary Sampling Units (PSUs). Geographic units (typically counties and or their equivalent) based on Census data and used in developing a sample.

Reference Period. The period of time to which interview questions relate.

Reference Person. An owner or renter of record who can reasonably be expected to answer questions about the household and about other household members (should they be unavailable for an interview).

Sample Attrition. Loss of sample members.

Seam Effect. The tendency of respondents to report a disproportionate number of changes as occurring at the "seam" between the end of one reporting period and the beginning of another.

Topcoding. The practice of recoding variables (like income) to protect against the possibility that the identity of a respondent with an extreme value might be discernible.

Type Z Nonresponse. An eligible person in an interviewed household from whom the interviewer could not get an interview, or for whom the interviewer could not obtain a proxy interview.

Wave. One round of interviewing in a longitudinal survey.

Weighting. Calculation of the number of units in a target population that a given sample unit represents.

Index

Address Unit, 11, 89
Adolescent Self-Administered Questionnaire, 7, 15, 18, 24, 25
Aid to Families with Dependent Children (AFDC), 9, 27, 87
Bureau of Labor Statistics (BLS), 87
Cross-Sectional Survey, 9, 89
Current Population Survey, 3, 7, 8, 17, 19, 36, 53, 54, 84, 87
  March Supplement, 3, 5, 7, 17, 35, 53-55, 58
Data Editing, 9, 31, 32, 89
Data Extraction System (DES), 8, 70, 87
Direct Variance Estimation, 44
Enumerative Check Census, 3
Error Estimation, 43
Federal Electronic Research and Review Extraction Tool (FERRET), 8, 87
General Accounting Office (GAO), 9, 87
Generalized Variance Functions (GVFs), 44, 87
  to Approximate the Standard Error of an Estimated Aggregate, 46
  to Approximate the Standard Error of an Estimated Difference, 47
  to Approximate the Standard Error of an Estimated Mean, 45, 47
  to Approximate the Standard Error of an Estimated Number, 45
  to Approximate the Standard Error of an Estimated Percentage, 46
  to Approximate the Standard Error of an Estimated Ratio of Means, 49
Household, 12, 89
Housing Unit, 89
Imputation, 31, 89
  Flag, 89
  Logical imputation, 31, 33, 34, 89
  Longitudinal edits, 33, 89
  Variance Estimation, 49
  "Cold deck" imputation, 33, 89
  "Hot deck" imputation, 33, 89
Income Survey Development Program (ISDP), 3, 87
Living Quarters (LQs), 11, 43, 87
Longitudinal Research, 57, 69
Longitudinal Survey, 8, 11, 57, 58, 65, 75, 89
Manpower Demonstration Research Corporation, 9
Match Key Variables, 54, 57, 60, 66
Missing Data, 31
Movers, 15, 90
National Health Interview Survey (NHIS), 21, 87
National Survey of American Families (NSAF), 8, 87
Nonresponse
  Item Nonresponse, 31-33, 49, 89
Nonsampling errors, 43
Original Sample Member, 11, 12, 15, 16, 40, 54, 57, 61, 62, 71, 77, 80, 81, 90
Panel, 90
Panel Study of Income Dynamics (PSID), 8, 87
Personal Responsibility and Work Opportunity Reconciliation Act , iii, 4
Primary Sampling Units (PSUs), 11, 44, 61, 67, 87, 90
Reference Period, 17, 19, 58, 65, 76, 90
Reference Person, 55, 56, 79-81, 90
References, 83
Residential History Calendar (RHC), 7, 18, 24, 26, 87
Sample Attrition, 14, 90
Sampling Errors, 43
Seam Effect, 76, 90
Spell Data, 75
Supplemental Security Income (SSI), 27, 87
Survey of Income and Program Participation (SIPP)
  Quality Profile, 11, 76
  Sample Design, 11
  Users' Guide, i, 11, 17, 31, 83
Survey of Program Dynamics (SPD)
  1997 Experimental File, 53
  1998 Calendar Year File, 59
  Analytic Uses, 73
  Content, 7, 17, 25
  Data Products, 8
  First Longitudinal File, 65

Primary Goals, 4
Sample, 6
Sample Design, 44
Survey Design, 12
Universe, 6
Uses, 5
Temporary Assistance for Needy Families (TANF), 27, 87
Topcoding, 34, 56, 57, 63, 68, 90
Type Z nonresponse, 31, 49, 90
Urban Change Study, 9
Wave, 90
Weighting, 35, 90
1997 SPD File, 36, 37
1998 SPD File, 37, 38
First Longitudinal File, 39, 41
SIPP Final Panel Weight, 35
Summary, 41
Women, Infants, and Children (WIC), 27, 87
Work Projects Administration (WPA), 3, 87