

ICPSR 3927

## **National Survey of America's Families (NSAF), 1999**

*Urban Institute*

*Child Trends*

User Guide

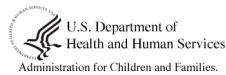
**ICPSR** | INTER-UNIVERSITY  
CONSORTIUM FOR  
POLITICAL AND  
SOCIAL RESEARCH

P.O. Box 1248  
Ann Arbor, Michigan 48106  
[www.icpsr.umich.edu](http://www.icpsr.umich.edu)

## About *Research Connections*

These data are made available by the Child Care and Early Education *Research Connections* (CCEERC) project, which promotes high quality research in child care and early education and the use of that research in policymaking. Our vision is that children are well cared for and have rich learning experiences, and their families are supported and able to work.

*Research Connections* is supported by the Child Care Bureau, Administration for Children and Families of the United States Department of Health and Human Services through a cooperative agreement with the National Center for Children in Poverty, Mailman School of Public Health at Columbia University, and its partner, the Inter-university Consortium for Political and Social Research, Institute for Social Research at the University of Michigan.



# 1999 NSAF Public Use File User's Guide

Report No. 11

Prepared by:

Nate Converse  
Adam Safir  
Fritz Scheuren  
Rebecca Steinbach  
Kevin Wang  
Urban Institute



Assessing  
the New  
Federalism

*An Urban Institute  
Program to Assess  
Changing Social Policies*

Methodology Reports

## **Preface**

The National Survey of America's Families (NSAF) is part of Assessing the New Federalism (ANF), a multi-year Urban Institute research project to analyze the devolution of responsibility for social programs from the federal government to the states, focusing primarily on health care, income security, job training, and social services. In collaboration with Child Trends, researchers from the Urban Institute monitor program changes and fiscal developments, along with changes in the well-being of children and families. Data collection for the NSAF was carried out in 1997, 1999 and 2002 by Westat. Further information about the ANF project is available on the internet at:

<http://www.urban.org/Content/Research/NewFederalism/AboutANF/AboutANF.htm>

or by writing to:

Assessing the New Federalism  
The Urban Institute  
2100 M St. NW  
Washington, DC 20037

The project is funded by a consortium of private foundations including the Annie E. Casey Foundation, the Robert Wood Johnson Foundation, the W.K. Kellogg Foundation, the Ford Foundation, the David and Lucile Packard Foundation, the John D. and Catherine T. MacArthur Foundation, the Henry J. Kaiser Family Foundation, the Charles Stewart Mott Foundation and others.

The NSAF is a household survey that can be used to produce cross-sectional estimates for a wide variety of child, adult and family well-being indicators at the state level for 13 states and the nation as a whole. The NSAF questionnaire content can be found in Report No. 1: 1999 NSAF Questionnaire. Other reports in the NSAF Methodology Series cover sample design, weighting, variance estimation procedures, data collection and data editing.

This report provides documentation for using the 1999 NSAF Public Use Files. It provides an overview of the survey, including descriptions of the sample design and data collection procedures. Chapter 2 provides information on the content and structure of the 1999 NSAF Public Use Files. Chapter 3 gives instructions and examples on how to use NSAF weights to obtain survey estimates. Chapter 4 describes procedures for calculating variances and standard errors of survey estimates.

## **Table of Contents**

<b><u>Chapter</u></b>	<b><u>Page</u></b>
<b>CHAPTER 1: OVERVIEW OF THE NATIONAL SURVEY OF AMERICA'S FAMILIES (NSAF).....</b>	<b>1-1</b>
1.1 INTRODUCTION.....	1-1
1.2 ABOUT THE SURVEY .....	1-1
1.2.1 OVERVIEW .....	1-1
1.2.2 GOALS FOR THE NSAF .....	1-2
1.3 LIMITATIONS OF EXISTING SURVEYS.....	1-4
1.4 NSAF SAMPLE DESIGN .....	1-5
1.5 NSAF INTERVIEWING PROCESS .....	1-6
1.6 WITHIN HOUSEHOLD SAMPLING.....	1-7
1.7 NSAF QUESTIONS .....	1-7
<b>CHAPTER 2: NSAF DATA OVERVIEW.....</b>	<b>2-1</b>
2.1 GENERAL DESIGN .....	2-1
2.2 DATA SETS .....	2-2
2.2.1 HOUSEHLD.....	2-2
2.2.2 SOCFAM .....	2-3
2.2.3 CPSFAM.....	2-3
2.2.4 FAMRESP .....	2-3
2.2.5 PERSON .....	2-3
2.2.6 FOCALCHD .....	2-3
2.2.7 ADULT_PR.....	2-4
2.2.8 ADULT_RN .....	2-4

2.2.9	ADULT_RB.....	2-5
2.3	COMPARISON TO 1997 PUBLIC USE FILE STRUCTURE.....	2-5
2.4	USING THE PUBLIC USE FILES.....	2-5
2.5	MISSING VALUES.....	2-6
2.6	USING THE DATA DICTIONARY.....	2-6
<b>CHAPTER 3: USING WEIGHTS WITH NSAF DATA.....</b>		<b>3-1</b>
3.1	OVERVIEW OF THE NSAF WEIGHTS.....	3-1
3.2	DESCRIPTION OF THE NSAF WEIGHTS.....	3-2
3.3	SELECTING THE UNIT OF ANALYSIS.....	3-6
3.4	APPLYING THE CORRECT WEIGHT FOR ADULT ESTIMATES .....	3-7
3.5	COMBINING ADULT DATA FILES .....	3-11
3.6	COMBINING ADULT AND CHILD DATA FILES .....	3-12
3.7	SUBGROUP ANALYSES .....	3-14
3.8	CHANGE ESTIMATES .....	3-15
<b>CHAPTER 4: CALCULATING STANDARD ERRORS.....</b>		<b>4-1</b>
4.1	LIMITATIONS OF STANDARD STATISTICAL PACKAGES.....	4-1
4.2	USING DESIGN EFFECTS FOR APPROXIMATE STANDARD ERRORS.....	4-2
4.3	METHODS FOR OBTAINING CORRECT STANDARD ERRORS .....	4-5
4.4	USING REPLICATE WEIGHTS TO CALCULATE STANDARD ERRORS.....	4-6
4.5	WESVAR.....	4-7
4.6	SAS AND STATA MACROS .....	4-8
4.7	ESTIMATING VARIANCES FOR CHANGES ESTIMATES .....	4-9

<b>REFERENCES .....</b>	<b>R-1</b>
<b>APPENDIX A    SAMPLE SIZES .....</b>	<b>A-1</b>
<b>APPENDIX B    DESIGN EFFECT TABLES.....</b>	<b>B-1</b>
<b>APPENDIX C    ROUND 1 - ROUND 2 CORRELATIONS .....</b>	<b>C-1</b>
<b>APPENDIX D    SAS MACROS FOR USE WITH NSAF DATA.....</b>	<b>D-1</b>
<b>APPENDIX E    COMPUTING JRR STANDARD ERRORS .....</b>	<b>E-1</b>

#### List of Tables

<u>Table</u>	<u>Title</u>	<u>Page</u>
TABLE 1-1	SUMMARY OF WELL-BEING MEASURES IN THE 1997 AND 1999 NSAF .....	1-9
TABLE 2-1	SUMMARY DATA SET DESCRIPTIONS .....	2-2
TABLE 2-2	MISSING VALUES .....	2-6
TABLE A-1	ROUND 1 EXTENDED INTERVIEW SAMPLE SIZE.....	A-1
TABLE A-2	ROUND 2 EXTENDED INTERVIEW SAMPLE SIZE.....	A-1
TABLE B-1	AVERAGE DEFF AND DEFT FOR ESTIMATES FROM 1999 NSAF CHILD FILE FOR ALL CHILDREN AND LOW-INCOME CHILDREN, BY SITE.....	B-2
TABLE B-2	AVERAGE DEFF AND DEFT FOR ESTIMATES FROM 1999 NSAF CHILD FILE FOR ALL HISPANIC CHILDREN AND LOW-INCOME HISPANIC CHILDREN, BY SITE.....	B-3
TABLE B-3	AVERAGE DEFF AND DEFT FOR ESTIMATES FROM 1999 NSAF CHILD FILE FOR ALL BLACK CHILDREN AND LOW-INCOME BLACK CHILDREN, BY SITE.....	B-3
TABLE B-4	AVERAGE DEFF AND DEFT FOR ESTIMATES FROM 1999 NSAF ADULT PAIR FILE FOR ALL ADULTS AND LOW-INCOME ADULTS, BY SITE.....	B-4
TABLE B-5	AVERAGE DEFF AND DEFT FOR ESTIMATES FROM 1999 NSAF ADULT PAIR FILE FOR ALL HISPANIC ADULTS AND LOW-INCOME HISPANIC ADULTS, BY SITE .....	B-4

TABLE B-6	AVERAGE DEFF AND DEFT FOR ESTIMATES FROM 1999 NSAF ADULT PAIR FILE FOR ALL BLACK ADULTS AND LOW-INCOME BLACK ADULTS, BY SITE .....	B-5
TABLE B-7	AVERAGE DEFF AND DEFT FOR ESTIMATES FROM THE 1999 NSAF ADULT PAIR FILE FOR ALL ADULTS IN HOUSEHOLDS WITH CHILDREN AND LOW-INCOME ADULTS IN HOUSEHOLDS WITH CHILDREN, BY SITE.....	B-5
TABLE B-8	AVERAGE DEFF AND DEFT FOR ESTIMATES FROM THE 1999 NSAF ADULT PAIR FILE FOR ALL ADULTS IN HOUSEHOLDS WITH NO CHILDREN AND LOW-INCOME ADULTS IN HOUSEHOLDS WITH NO CHILDREN, BY SITE .....	B-6
TABLE C-1	CORRELATIONS FROM 1997 AND 1999 NSAF CHILD FILE FOR ALL CHILDREN , BY STUDY AREA .....	C-3
TABLE C-2	CORRELATIONS FROM 1997 AND 1999 NSAF CHILD FILE FOR LOW-INCOME CHILDREN, BY STUDY AREA .....	C-3
TABLE C-3	CORRELATIONS FROM 1997 AND 1999 NSAF ADULT PAIR FILE FOR ALL ADULTS, BY STUDY AREA .....	C-4
TABLE C-4	CORRELATIONS FROM 1997 AND 1999 NSAF ADULT PAIR FILE FOR LOW-INCOME ADULTS, BY STUDY AREA.....	C-4
TABLE C-5	CORRELATIONS ESTIMATED FROM HOUSEHOLDS SAMPLED IN BOTH ROUNDS FOR ALL CHILDREN, BY STUDY AREA .....	C-5
TABLE C-6	CORRELATIONS ESTIMATED FROM HOUSEHOLDS SAMPLED IN BOTH ROUNDS FOR ALL ADULTS (ADULT PAIR FILE), BY STUDY AREA .....	C-5
TABLE C-7	CORRELATIONS ESTIMATED FROM CHILDREN SAMPLED IN BOTH ROUNDS FOR ALL CHILDREN, BY STUDY AREA .....	C-6
TABLE C-8	CORRELATIONS ESTIMATED FROM ADULTS SAMPLED IN BOTH ROUNDS FOR ALL ADULTS (ADULT PAIR FILE), BY STUDY AREA .....	C-6

### **List of Figures**

<b><u>Figure</u></b>	<b><u>Title</u></b>	<b><u>Page</u></b>
FIGURE 1	13 TARGETED NSAF STATES .....	1-3



# **Chapter 1: Overview of the National Survey of America's Families (NSAF)**

## **1.1 Introduction**

This report presents basic information on the 1999 National Survey of America's Families (NSAF) public use data. The goal of this report is to provide users with enough information on the survey itself and data files to be able to use the data. The NSAF is part of *Assessing the New Federalism (ANF)*,<sup>1</sup> a multi-year Urban Institute research project to analyze the devolution of responsibility for social programs from the federal government to the states, focusing primarily on health care, income security, job training, and social services. In collaboration with Child Trends, researchers from the Urban Institute monitor program changes and fiscal developments, along with changes in the well-being of children and families. Data collection for the NSAF was carried out in 1997, 1999 and 2002 by Westat. Further information about the ANF project is available on the internet at:

<http://www.urban.org/Content/Research/NewFederalism/AboutANF/AboutANF.htm>

or by writing to:

Assessing the New Federalism  
The Urban Institute  
2100 M St. NW  
Washington, DC 20037

This introductory chapter gives an overview of the survey, including descriptions of the sample design and data collection procedures. Chapter 2 provides information on the content and structure of NSAF public use files for the 1999 NSAF.<sup>2</sup> Chapter 3 gives instructions and examples on how to use NSAF weights to obtain survey estimates. Chapter 4 describes procedures for calculating variances and standard errors of survey estimates.

## **1.2 About the Survey**

### **1.2.1 Overview**

During this period of devolution, the Urban Institute began a project entitled *Assessing the New Federalism (ANF)*. The project's goals are to give policymakers, state administrators, and advocates information they need to make better decisions and to help the nation determine the consequences of devolution.

---

<sup>1</sup> Assessing the New Federalism is funded by a consortium of private foundations including the Annie E. Casey Foundation, the Robert Wood Johnson Foundation, the W.K. Kellogg Foundation, the Ford Foundation, the David and Lucile Packard Foundation, the John D. and Catherine T. MacArthur Foundation, the Henry J. Kaiser Family Foundation, the Charles Stewart Mott Foundation and others.

<sup>2</sup> In the near future, we hope to re-release public use data from the 1997 NSAF in the same structure that we are releasing public use data from the 1999 NSAF.

The project is premised on the notion that better information yields better policies. With increased state-level authority, state data becomes increasingly more important; yet such data are currently very limited. Of course, devolution should not be “evaluated” with a pronouncement of success or failure at the end, but should be monitored, with continuous input into the policy and implementation process. This implies that a new relationship between research and practice is necessary and a long-term effort needed.

One of the project’s defining features is the breadth of topics it covers:

- Welfare Reform
- Employment and Training
- Health Care
- Health Insurance Coverage
- Child Care
- Child Support
- Child Welfare
- Child Well-Being

While the focus is on states, the scope is national—with a primary emphasis on low-income families with children.

Data collection and analysis are extensive and varied, with wide dissemination a major component. The project has employed four major data sources:

- (1) The compilation and integration of existing state databases;
- (2) In-depth, state-specific baseline case studies with follow-up monitoring;
- (3) Special surveys of states; and
- (4) Our topic here, the National Survey of America’s Families.

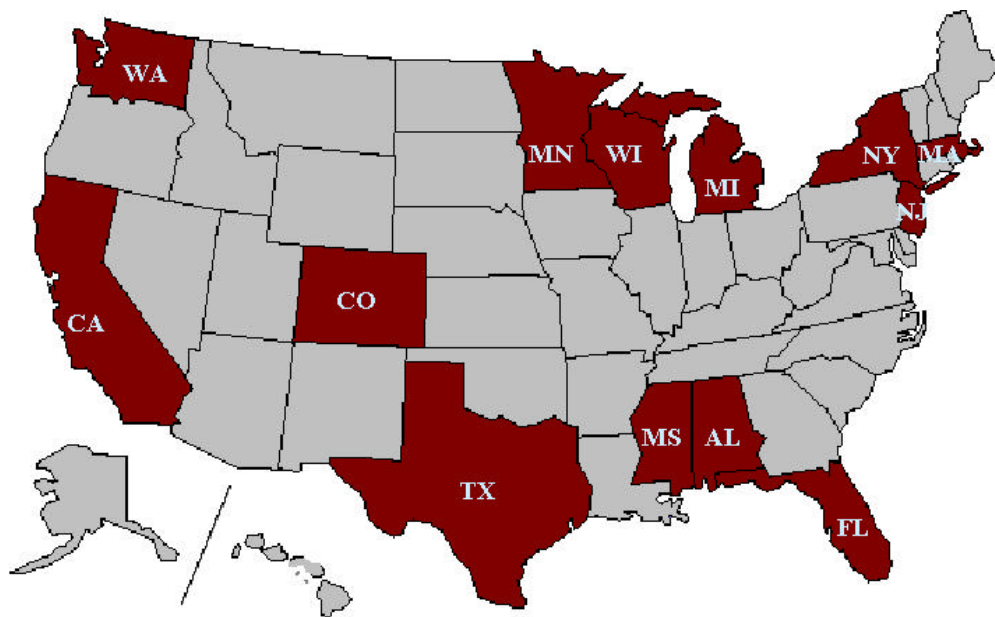
### **1.2.2 Goals for the NSAF**

The unfolding New Federalism can be viewed as 51 “natural experiments.” Ideally, the effects of devolution should be examined by conducting intensive case studies and tracking changes in well-being in *all* states, as well as by controlling for substate variation in program implementation. However, this would be prohibitively expensive. Instead, the Urban Institute decided to focus on just 13 states, shown in Figure 1. These study areas (see Figure 1) were selected by the Urban Institute to vary in terms of their size, geographic location, fiscal capacity, citizens' needs, and traditions of providing government services.

Intensive case studies were conducted in these 13 states during 1997 and 1999 to obtain an in-depth understanding of each state's responses to the New Federalism. Monitoring continues as each state's circumstances change.

NSAF data from all rounds are used to explore linkages between state policy and child and family well being. The data were collected with the purpose of allowing analysts at the Urban Institute and Child Trends to look at changes over time—with the NSAF used to make comparisons across states at a point in time, within a state over time, and across states over time.

**Figure 1: 13 Targeted NSAF States**



Given these objectives for the ANF project, we identified four key requirements for the design of the survey. These requirements were:

- Large, representative samples of families with children in the intensively studied states and the nation;

- Large, representative samples of low-income families with children in the intensively studied states and the nation;
- Observations at multiple points in time; and
- A broad range of well-being indicators that are uniformly measured across states.

Because of the survey's breadth, which encompasses the economic, health, and social dimensions of well-being, the NSAF permits comprehensive assessments of child and family well-being and program participation, while controlling for important socioeconomic characteristics.

### **1.3 Limitations of Existing Surveys**

The ANF Project determined that the NSAF was necessary after we reviewed existing surveys and found them to be lacking in several aspects for meeting the survey objectives:

In 1997, many major national household samples—such as the Survey of Income and Program Participation (SIPP)—were not state representative, which meant that reliable state-specific estimates could not be produced from them. At that time, exceptions were the Current Population Survey (CPS) and the National Health Interview Survey (NHIS). However, as will be described shortly, neither the CPS nor the NHIS met the project's needs because of limitations with respect to content and sample size.

State-representative surveys either focused on narrow aspects of well-being or did not include variables that relate to the anticipated policy changes. For example, the CPS focuses mainly on employment, and at the time of the 1997 NSAF it did not include information on health services use or access to care. The NHIS has the needed health questions, but lacks both information on receipt of AFDC and food stamps, and detailed income information. In addition, neither the CPS nor the NHIS contains information on the need for and use of social services or child care.

Finally, even with state-representative sample frames, the existing surveys' sample sizes in most states were too small (particularly for the low-income population). Without some supplementation, the assessment of changes over time in individual states would be very imprecise.

In summary, with the data sources available in 1997, it was simply not possible (without the NSAF) to make reliable state-specific estimates on a wide range of well-being indicators, either because the samples were not state representative or because the survey content was too narrow. In addition, even if one were willing to narrow the scope of well-being measures to just a few dimensions, the sample sizes available for examining low-income households would make comparisons over time very imprecise for most states.

There have now been two rounds of NSAF, fielded during 1997 and 1999. The data collection for the survey was done by Westat for the Urban Institute and Child Trends. A third round is being conducted

in 2002. Each round of the survey collects information on the economic, health, and social dimensions of the well being of children, adults under the age of 65, and their families. While the survey is national, 13 states are targeted for special study. Wisconsin was targeted for particularly intensive study, with separate large samples for Milwaukee<sup>3</sup> and the balance of the state. Data are also collected in the balance of the nation to permit national estimates.

## **1.4 NSAF Sample Design**

The NSAF is designed to produce estimates that are representative of the civilian, noninstitutionalized population under age 65. As with virtually all household surveys, some segments of the population (e.g. the homeless) could not be sampled because of their living arrangements and are not included in the survey results.

The NSAF draws households from two separate sampling frames. The first frame consists of a households from a random-digit dial (RDD) sample of households with telephones. The RDD approach was adopted because it is a cost-effective means to collect the data. However, because households without telephone service contain a disproportionate number of low-income children, a supplementary area sample was conducted in person for those households without telephones. Nationally, Giesbrecht et al. (1996) estimate that about 20 percent of families in poverty have no telephone and that about 10 percent of families with one child age three or under have no telephone. The area sample provides data for these and other families without current phone service. A sizable area sample was screened to find such households. No other households were interviewed from the area sample.

In addition to the use of RDD sampling to reduce costs (compared to pure area sampling), costs were further reduced through the use of screener-based subsampling of households contacted in the RDD component. With this approach, a single income question was asked during the RDD screening interview. Those households that reported an absence of children or reported incomes above 200 percent of the poverty threshold were subsampled. More detailed and reliable income questions were asked during the extended interview for those not “subsampled out.” In Round 1, there was less consistency between the screener and extended versions of the income question than had been anticipated. This led to a serious review of how to re-optimize the subsampling rates for the households that initially report high income for Round 2 (see section 3.3 of chapter 3 in the 1999 NSAF Methodology Report No. 2).

For the 1997 NSAF the RDD telephone sample started out with 483,260 randomly generated phone numbers in 100-banks<sup>4</sup> with at least one listed residential phone number. Out of these, 48 percent were found or imputed to be working residential phone numbers. Screening data were obtained from 76 percent of the working residential numbers, yielding a total screened sample of 176,791 households. Given the sampling rules described above, 32,474 households with children were selected for extended interviews, as well as another 19,800 adult-only households. From these, 40,400 children, 3,500 other

---

<sup>3</sup> Milwaukee County will not be oversampled in the 2002 design.

<sup>4</sup> A 100-bank is a block of 100 consecutive phone numbers with the same first eight digits. Thus, the telephone numbers between 212-555-5500 and 212-555-5599 would make up one 100-bank.

adults in households with children, and 22,100 adults in adult-only households were selected. The nontelephone sample for 1997 was significantly smaller than the telephone sample. It consisted of 1,682 interviews from about 1,500 eligible nontelephone households. These households were found by screening an initial sample of 44,400 dwelling units across 1,388 segments in 114 primary sampling units (PSUs). Note that one MKA interview can provide data on up to two sample children and that there were a few MKA interviews with parents under the age of 18.

In contrast to the 1997 NSAF, the second round of the survey built in part on the RDD sample selected in the initial round, augmenting it with an additional sample of newly selected telephone numbers. This overlap was quite large and led to both cost savings and some reduction in the variance of change measures from round to round (See Report No. 2 and Report No. 4 in the 1999 Methodology Series). For the 1999 sample, a total of 383,653 telephone numbers were used. Of these numbers, about 217,421 were drawn from the round 1 sample of telephone numbers. These numbers were subsampled from round 1 numbers at different rates based on the round 1 screener result code. Overall for round 2, 147,623 households were screened, and detailed extended telephone interviews were conducted in 40,874 households. For a complete description of the NSAF telephone survey methods, see Report No. 9, in both the 1997 and 1999 NSAF Methodology Series.

The nontelephone sample for 1999 also built on the earlier 1997 selections, in that for 1999, we returned to most of the same sample segments used in 1997. There was, however, some supplementation in the balance of the US to reduce the large variances that were found to exist across PSUs. In the 1999 NSAF in-person component, 1,676 extended interviews were conducted in 1,486 households.

Overall, the samples for 1997 and 1999 were roughly of equal size. Appendix A provides a state-by-state summary of the achieved sample sizes for 1997 and 1999, for children and adults separately—both in total and for low-income individuals.

## **1.5 NSAF Interviewing Process**

In the RDD portion of the NSAF, the interviewing consisted of a short screener to determine if the household was to be selected for an extended interview. There were two types of extended interviews. The longer type of extended interview (45 minutes), referred to as an Option A interview, was administered in households with children under 18. The shorter interview (27 minutes), called Option B, was given to adults under 65 without children under 18. Option A interviews asked questions about both children and their families, while Option B interviews contained only the questions from Option A that were relevant to adults. The questionnaire was divided into several sections, including the following topics: education, health care coverage and access, child care, employment and earnings, family income, welfare participation, housing and economic hardship, social services, problems, race, ethnicity, and nativity.

The interviews and screener were programmed into Westat's CATI system to facilitate administration and data editing. Data collection for the screener and extended interview started on February 15,

1999. For the 1997 survey, the screening began before February 15 with the extended interviewing beginning at approximately the same time. In both rounds, therefore, the timing of data collection was planned to allow respondents a chance to receive all of their 1998 tax information documents (W-2s, Forms 1099, etc.) before the interview and thus be able to answer questions about their prior year's income in a more informed way. The survey was completed on early November for the 1997 NSAF and in early October for the 1999 NSAF. For more information about the interview process, see Report No. 5, on the in-person interviews, and Report No. 9, on telephone interviews, in the both the 1997 and 1999 Methodology Series

Households without telephones were administered an altered version of the screening instrument. Since all nontelephone households with at least one age-eligible person (under 65) were interviewed, this version of the screener did not screen out any households because of income. The screener merely verified that there was not a working telephone in the household and, if not, continued with the same series of sample person selection questions described above for telephone households.

## **1.6 Within Household Sampling**

For the entire NSAF sample of households, there was a decision to subsample household members to reduce the respondent burden. If there were multiple children under age 6, one was randomly selected. The same was done for children ages 6 to 17. No more than two children were sampled from each household. For example, if a household had three children all under the age of five, then only one child was sampled and there was not a second focal child. Furthermore, if there were two families in the household and each had two children (one between 0 and 5 years old and one between 6 and 17 years old), only one child age 0 to 5 and one child age 6 to 17 were sampled. Both children could be from the same family or there could be one child from each family.

Data were collected about each of these sample children through the most knowledgeable adult in the household for that child. In choosing the MKA, interviewers asked to speak to the person in the household who knew the most about the sampled child's education and health care. Therefore, selection of MKAs was not a random process; rather, the interviewer sought to obtain the highest quality information possible for each child. In families with two sampled children, the MKA was not necessarily the same person for both children. Consequently, there were cases in which one family had two MKAs.

For households without children, up to two childless adults between the ages of 18 and 64 were selected for interviewing. If two childless adults were selected for interviewing, they could not be spouse/partners of each other. For households with children, up to two adults between the ages of 18 and 64 who were not identified as having children under the age of 18 in the household could also have been selected for interviewing. As in households without children, these additional adults selected for interviewing could not be spouse/partners of each other.

## **1.7 NSAF Questions**

A summary table of survey content is shown in Table 1-1. All items for the 1997 and 1999 questionnaires are provided in Report No. 12 (1997 NSAF Questionnaire) and Report No. 1 (1999 NSAF Questionnaire). For interviews with MKAs, most items about adults were asked about both the MKA and his or her spouse/partner if the spouse/partner also lived within the household. Items on current health insurance status, employment, education and training were asked about both the MKA and spouse/partner. However, some questions were asked only about the MKA. These questions concerned feelings, religious and volunteer participation, and attitudes and opinions. Other questions concerning past year's health insurance coverage and health care utilization were randomly asked about one of the two when both were present. The subject for the latter questions—the MKA and his or her spouse/partner—was randomly selected by the survey instrument. The concern was that collecting information about the child, the MKA, and the spouse of the MKA all by proxy through the MKA would tire the MKA excessively. By asking these questions about only the MKA or his or her spouse or partner, the burden on the MKA was reduced. This protocol was applied identically in the RDD and area components.

For interviews with childless adults, as with MKA interviews, many items were asked of both the childless adult respondent and his or her spouse/partner. Items on religious and volunteer participation and attitudes towards welfare and raising children were asked only of the respondent. However, unlike in MKA interviews, in childless adult interviews, both the respondent and the spouse/partner were subjects of questions on health insurance and health care utilization.

It is important to note that most NSAF items are not asked of all persons in families or households. This means that for many measures of adult or child well-being, it will not be possible to construct aggregate measures of well being based on individual information for families or households that will be valid for all families or households. For example, current employment status is only ascertained for respondents and their spouse/partners. There may be other adults in the household for whom current employment status is not determined. Therefore, trying to build a measure of the number of adults in the family or household who are currently employed will yield an invalid measure, unless one were to restrict the population to families or households with only two adults where the adults are spouse/partners of each other. Similarly, most of the items on child well-being are only determined for focal children and not all children in the household or family.



**Table 1-1.**  
**Summary of Well-Being Measures in the 1997 and 1999 NSAF**

<i>Well-Being Construct/ Items Measured</i>	<b>Person/Unit for Whom Measured:</b>		
	<b>Child</b>	<b>Parent/ Adult</b>	<b>Family/ Household</b>
<i>Economic Security</i>			
Poverty/family income			X
Parent/Adult employment/earnings/work stability		X	
Health insurance coverage (includes Medicaid)	X	X	
Parent/Adult use of education and training		X	X
Child support	X	X	X
Use of public assistance (includes AFDC, SSI)	X	X	X
Use of food assistance (includes Food Stamps, WIC, school lunch and school breakfast)	X	X	X
Economic hardship			X
Food Security		X	X
Use of housing assistance			X
Housing adequacy/stability/crowding	X	X	X
<i>Health and Health Care</i>			
Health status/limitations	X	X	
Hospitals stays and Physician visits	X	X	
Health care access, use, and satisfaction	X	X	
Health care monitoring (includes dental visits, well- care/preventive care)	X	X	
Ability to afford medical/dental care, medicine	X	X	
<i>Child's Education/Cognitive Development</i>			
Grade for age	X		
Problem doing well in school, with school work	X		
Whether parents read or tell stories to child	X		
Whether parents take child on outings	X		
Child care use (including amount, type, quality, stability)	X		X
<i>Child's Social Development and Positive Development</i>			
Employment and participation in training programs	X		
Participation in recreational activities: Teams, clubs, scouts, religious groups	X		
<i>Child's Behavior Problems</i>			
Behavior Problems Index	X		
Cut classes/suspended expelled from school	X		
<i>Family Environment</i>			
(A) <i>Family Structure</i>			
Whether two-parent family, whether biological parents present	X	X	
Visitation with noncustodial parent (if relevant)	X		
Stability/turbulence (includes changes in family composition, housing, child care)	X	X	X
(B) <i>Parent/Adult Psychological Well-Being</i>			
Depression		X	
Attitudes toward Parenting		X	
Participation in volunteer/religious activities		X	
(C) <i>Family Stress</i>			
Problems in family (includes mental health, family conflict)	X	X	X

(D)	Immigration Status	X	X	X
(E)	Child-Rearing Practices			
	Monitoring; well-child care; dental visits	X		
	Community Environment			
	Knowledge of community services availability		X	

## Chapter 2: NSAF Data Overview

### 2.1 General Design

The NSAF data are organized into rectangular public use files, including household-, family-, person-, adult-, and child-level data sets. This hierarchical structure was adopted for two reasons. First and foremost, this structure is consistent with the manner in which the data are collected in the NSAF. Second, this structure corresponds to the different weights that are applied to the various data elements. Organizing the variables into multi-level data sets with the appropriate weights attached facilitates analysis of the data.

Each data set contains five types of variables:

**Survey variables** give information obtained directly from questions asked on the survey. Each of these variables begins with the letter corresponding to the section of the survey from which it was obtained.

**Constructed variables**, which begin with the letter "U," were created by Urban Institute staff for analysis purposes. Some of these are straightforward recodes of individual survey items. Others involve aggregating information from several or many survey variables to create more complex measures, such as family income as a percentage of the poverty threshold (e.g. UINCRPOV, U\_SOCPOV). Users should check to see if a constructed variable has been created that meets their analysis needs before going directly to the use of survey variables, especially if they believe that the measure they are trying to create will involve a large number of survey items.

**Administrative variables** provide information that was not obtained from the interview, such as the geographic location of the household and information about the interviewing process itself. Also included are identifier variables.

**Imputation flags** indicate which observations had data imputed for a given variable. Imputation flags begin with an "X," generally followed by the name of the variable for which data was imputed.

**Weights**, as mentioned above, are included on the appropriate files. Weight variables begin with the letter "W". For more information on weights, see chapter 3 of this report.

**Table 2-1.**  
**Summary Data Set Descriptions**

<b>Data Set Name</b>	<b>Data Set Description</b>	<b>Weights</b>
HOUSEHLD	One record for each household. Household-level data items.	(None)
SOCFAM	One record for each social family. Family-level data items (all survey sections plus created variables).	WGSOCAD0 - WGSOCAD60
CPSFAM	One record for each CPS family. Family-level data items (created variables).	WCPSAD0 - WCPSAD60
FAMRESP	One record for each respondent. Family / respondent level data items from all sections.	(None)
PERSON	One record for each person living in the household. Person-level data items from sections E (Health Insurance Coverage), J (Income), and O (Demographic).	(None)
FOCALCHD	One record for each focal child. Person-level data items from extended interviews.	WGFCAD0 - WGFCAD60
ADULT_PR	One record for each respondent (both Option A and Option B). One record for each spouse/partner of the respondent. Information from sections I ( Employment and Earnings) and L (Education and Training).	WGPRAD0 - WGPRAD0
ADULT_RN	One record for each Option A interview (either for the respondent or his/her spouse/partner depending on which one is chosen). One record for each Option B respondent. One record for spouse/partner (if one exists) of Option B respondents. Person-level data items from sections E (Health Insurance Coverage) and F (Health Care Access and Utilization).	WGRNAD0 - WGRNAD60
ADULT_RB	One record for each Option B respondent only (not spouse/partners). Person-level data items from sections N (Issues, Problems and Social Services) and P (Closing section).	WGRBAD0 - WGRBAD60

## 2.2 Data Sets

This release of the NSAF Public Use Files includes one household-level data set, HOUSEHLD. There are two family-level data sets, SOCFAM and CPSFAM; each aggregates family information based on a different definition of a family. In addition, one respondent-level data set, FAMRESP, is included. There are five person-level data sets: PERSON, FOCALCHD, ADULT\_PR, ADULT\_RN, and ADULT\_RB. The PERSON data set has one observation for each person living in the household, and the remaining four contain information collected about sampled people during the extended interview. That is, these data sets contain information about focal children, respondents (Option A and Option B), and spouses/partners of respondents.<sup>5</sup>

### 2.2.1 HOUSEHLD

---

<sup>5</sup> Emancipated minors appear on the FOCALCHD data set as well as on the ADULT\_PR, ADULT\_RN, and ADULT\_RB data sets along with their data items from the Option B interviews.

There is one household-level data set called HOUSEHLD. This data set contains household and administrative variables. The HOUSEHLD data set has one linking variable, HHID, which should be used in order to merge between HOUSEHLD and any other data set.

### **2.2.2 SOCFAM**

The SOCFAM data set contains items that were asked about the social family as well as variables that were aggregated at the social family level using the UFAMID variable (social family ID). Among the survey items included in the SOCFAM data set are those variables indicating whether anyone in the social family had a particular type of income (e.g., JAFDC) or a particular type of health insurance (e.g., EEMP1COV). The SOCFAM data set also contains created variables that summarize information across all social family members, such as UNFAMILY (indicating the number of family members). The SOCFAM file includes the linking variable UFAMID, which is the social family ID.

### **2.2.3 CPSFAM**

Because the social family definition was used in fielding the NSAF, the CPSFAM data set includes only variables created using the UCPSID. The CPS family ID variable UCPSID is also the linking variable for this data set.

### **2.2.4 FAMRESP**

There is one family/respondent-level data set called FAMRESP. The term “family-level” is accurate, to some extent, in that information collected about the respondent’s family is contained in this data set. However, observations are actually at the respondent level. Because in the NSAF there may be two respondents within a family and some family-level questions are asked of each respondent, it is possible to have two different answers for the same family-level variable. Ideally, the answers from both respondents will be the same, but this is not true all of the time. The FAMRESP data set has two linking variables, HHID and RESPID. The RESPID variable should be used in order to merge between the FAMRESP and any person-level data set. The HHID variable should be used to merge to the HOUSEHLD data set.

### **2.2.5 PERSON**

This data set contains one observation for each person living in the household. Records for non-resident household members have been removed from this data set, because very little information, if any, was collected about these individuals. This data set contains demographic information (age, sex, marital status of adults, race/ethnicity, etc.) as well as information on current health insurance status and income. The primary linking variable on this file is PERSID.

### **2.2.6 FOCALCHD**

This data set contains data elements from the extended interview that are specific to either the FC1 or the FC2. Select data items asked only of MKAs are also included in this data set because the unit of analysis with respect to these questions will usually be children. These items are primarily from sections N and P. The PERSTYPE variable indicates whether the observation is an FC1 or an FC2 observation. The value when PERSTYPE is equal to '1' means the observation is for a younger focal child (FC1) and the value when PERSTYPE is equal to '2' means the observation is for an older focal child (FC2). A PERSTYPE value of '23' indicates the observation is for an FC2 who is also the spouse/partner of an MKA. Similarly, a PERSTYPE value of '24' indicates that the observation is for an FC2 who is also an MKA (of the FC1). There are several linking variables on the FOCALCHD data set; these include the person ID of the child (PERSID), the respondent ID of the MKA who provided information about the child (RESPID), the CPS family ID (UCPSID), the social family ID (UFAMID), and the household ID (HHID).

### **2.2.7 ADULT\_PR**

The structure of the ADULT\_PR, ADULT\_RN, and ADULT\_RB data sets was driven by the NSAF questionnaire design and sampling. In the survey, some questions are asked about both the respondent and his/her spouse/partner (if one exists), others are asked about either the respondent or his/her spouse/partner (randomly chosen), and still others are asked about only the respondent. Thus, three different person level adult weights are needed for the three different types of data elements. To make it less confusing for users to match data elements to the appropriate weight, these three data sets have been created so that weights and data items in the files are consistent.

The ADULT\_PR data set contains data elements from the extended interview that are collected about both the respondent (Option A and Option B) and his/her spouse/partner. There is one observation per respondent and one per spouse/partner (if one exists) in this data set. This data set contains information primarily from sections I (Employment and Earnings) and L (Education and Training). The PERSTYPE variable indicates whether the observation belongs to a respondent or to the respondent's spouse/partner. A value of 4 (MKA) or 6 (Option B respondent) means the observation is for a respondent, while a value of 3 (spouse/partner of MKA) or 5 (spouse/partner of Option B respondent) means the observation is for the respondent's spouse/partner. The linking variables on the ADULT\_PR data set include PERSID, RESPID, UCPSID, UFAMID, and HHID.

### **2.2.8 ADULT\_RN**

The ADULT\_RN data set contains data elements from the extended interview that are specific to a randomly selected adult (either the respondent or the spouse/partner). This situation occurs only in sections E (Health Insurance Coverage) and F (Health Care Access and Utilization) for Option A interviews. In Option B, questions are asked from sections E and F about both the respondent *and* the spouse/partner. However, rather than splitting observations for variables across two data sets, information from sections E and F for Option B adults is included in this data set. Thus, for Option A interviews, there is one observation per interview, whereas for Option B interviews, there may be one or two records, one for the respondent and one for the spouse/partner (if one exists). Here again, the

PERSTYPE variable indicates whether the observation belongs to a respondent or to the respondent's spouse/partner. The linking variables on the ADULT\_RN data set include PERSID, RESPID, UCPSID, UFAMID, and HHID.

### **2.2.9 ADULT\_RB**

The ADULT\_RB data set contains data elements from the extended interview that are asked only of the respondent in Option B interviews. These are items from sections N (Issues, Problems, and Social Services) and P (Closing section). There is one record per Option B interview in this data set. As described above, similar data elements collected only about the respondent in Option A interviews are included in the FOCALCHD data set, because the unit of analysis for these data items will be the focal child. Thus, some "respondent only" variables are contained in both the FOCALCHD and the ADULT\_RB data sets. The linking variables on the ADULT\_RB data set include PERSID, RESPID, UCPSID, UFAMID, and HHID.

## **2.3 Comparison to 1997 Public Use File Structure**

Researchers who have previously used the 1997 NSAF Public Use Files will note that the structure of this data release is significantly different from previous releases. In Round 1, data were made available in pieces—focal child, MKA, non-MKA, and family files—as work was finished. As a result, the structure of the Public Use Files differed from that of the internal Urban Institute data files. Work on the 1999 Public Use Files was completed more quickly, allowing the public release of files in a structure identical to that of the internal files. In the interests of uniformity, the 1997 NSAF data is also being re-released in the nine file format. The new format should facilitate analysis of both Round 1 and Round 2 data by researchers, while not affecting in any way conclusions drawn from the previously released 1997 NSAF Public Use Files.

## **2.4 Using the Public Use Files**

Each NSAF Public Use File is available as a compressed ASCII file contained in a self-extracting program that must be downloaded and uncompressed. To download the file and save it to your hard drive, click on the file name. A window will appear asking for the location where the file will be saved. Enter the location and choose "Save." To unzip the file, go to the File Manager or Windows Explorer and double-click the downloaded file. The extraction program will unzip the ASCII file into the same directory and create a new subdirectory. The new subdirectory will contain six files: (1) *filename.pdf*, the data dictionary for the file, in Adobe PDF format; (2) *filename.txt*, the record description, including variable names, types (i.e., character or numeric), positions (i.e., the columns the variable occupies), and labels; (3) an ASCII copy of the Public Use Data File the researcher has selected; (4) a SAS sample read-in statement for use with the data; (5) an SPSS input statement for use with the data; and (6) *readFN.me*, a read me file which describes the content of the files.

To convert the ASCII file to an SAS data set, use the SAS sample read-in data step and change the infile statement to refer to the location of the downloaded, uncompressed file. To convert the ASCII file to an SPSS data set, use the SPSS sample read-in data step and change the file statement to refer to the location of the downloaded, uncompressed file.

## 2.5 Missing Values

There are four types of missing values in NSAF data: Inapplicables, Refusals, "Don't Know" responses, and Not Ascertained. The first, Inapplicables, occurs when a sampled person is not eligible to receive a certain question. Refusals take place when a respondent refuses to answer a question. "Don't Know" values occur when a respondent does not know or cannot answer a question. Finally, Not Ascertained values occur when a respondent is eligible for a certain question but for some reason or other the question was not asked. In this release missing values are represented in different ways to accommodate both SAS and SPSS users.

SAS Users: The SAS read in statement contains an array that converts all missing values to character format. Inapplicables are denoted by (.I), refusals (.R), "Don't Know" responses (.D), and not ascertained (.N).

SPSS Users: In this release missing values are signified by negative numbers. Inapplicables are represented by (-1), refusals (-7), "Don't Know" responses (-8), and not ascertained (-9).

**Table 2-2. Missing Values**

Missing Value	SAS	SPSS
Inapplicable	.I	-1
Refused	.R	-7
"Don't Know"	.D	-8
Not Ascertained	.N	-9

## 2.6 Using the Data Dictionary

The data dictionary, or codebook, provides information on the variables released on the NSAF Public Use Files. Previous users of the NSAF Public Use Files may note that this data dictionary has a slightly different format than that used in the codebooks released as part of the 1997 NSAF Methodology Series. Each entry in this data dictionary has eight fields: variable name, label, type, length, question number, question text, description, and frequencies/means. When the Round 1 files are re-released, two additional fields, 'R1-R2 Changes' and 'R1-R2 Availability,' will be added to the data dictionary for a total of ten fields. In the meantime, researchers interested in analyzing changes across time are encouraged to check the appropriate Round 1 data set codebook for any differences in question text or allowable non-missing values.



Each of the data dictionary fields is described in more detail below:

**Variable.** For each entry, a mnemonic string of characters is provided as the variable name. The first letter of the variable name indicates the section of the questionnaire from which the variable was obtained, while the remaining characters (up to seven more) are a short description of the variable. For example, the variable CCHGSC comes from section C of the questionnaire, which deals with the child's education. The remaining characters are a mnemonic reference to the variable description "Changed (shortened to CHG) school (shortened to SC) past 12 months."

Notable exceptions to this naming convention include created variables and demographic variables. Variables beginning with a U have been created by Urban Institute analysts and do not come directly from a survey question. To avoid confusion, demographic variables, most of which were obtained during the initial household screening, consist only of one relevant word. For example, the variable indicating gender is simply SEX. Flags for imputed variables begin with letter "X," replacing the section identifier that normally occupies the first letter of a variable name. Finally, weights begin with the letters "W."

**Label.** The label is a short description of the variable. In some cases, the label contains slight abbreviations. For example, the variable BDISBL has the label "Has hlth condition that limits activity," in which the word "health" has been shortened to "hlth."

**Type.** Variables can either be character variables, designated by 'C', or numeric, designated by 'N'. Researchers should note that values for character variables are case-sensitive, so that while 'AL' is an allowable value for the character variable STATE, 'al' is not.

**Length.** This field indicates the length of the variable.

**Question Number.** The question number is provided if the variable was obtained directly from the survey. This field is left blank for all imputation flags, weight variables, and analysis variables created using other variables from the survey.

**Question Text.** Text from the questionnaire is provided if the variable was obtained directly from the survey. This field is left blank for all imputation flags, weight variables, and analysis variables created using other variables from the survey.

**Description.** The description field lists any special instructions given to the interviewer for the question. In addition, the description field clarifies some of the terms used in the question and variable label, and, when possible, relates these terms to those that are used by the U.S. Census Bureau in its Current Population Survey (CPS). Comments on the appropriateness of the variable for analysis and changes between round one and round two are located in the description field.

**Frequencies/Mean.** Weighted and unweighted variable counts, along with value labels, are provided here for categorical variables. Means are provided here for continuous variables, along with the variable's range of values. SPSS users should note that frequencies of variables denote missing values in character format.

## **Chapter 3: Using Weights with NSAF Data**

### **3.1 Overview of the NSAF Weights**

Responses to NSAF items were weighted to provide approximately unbiased aggregate estimates for each study area and for the country as a whole. The weights were applied to all survey items in an effort to:

- Compensate for differential probabilities of selection for households and persons;
- Reduce biases occurring where nonrespondents have different characteristics than respondents;
- Adjust, to the extent possible, for undercoverage in the sampling frames and in the conduct of the survey; and
- Reduce the variance of the estimates by using auxiliary information.

The weighting can be described as involving three stages for both the random-digit dial and in-person components of the NSAF to produce person and family weights:

- The first stage was the computation of the base weight. The base weight is the inverse of the probability of selection, which accounts for the unequal screening rates. This weight also includes an adjustment for the planned exclusion of nontelephone households from the area sampling frame and for the subsampling of persons in selected households.
- The second stage was an adjustment for unit nonresponse (entire households and persons who did not respond to the survey). This was done by adjusting the weights of respondents in particular groups to account for the nonrespondents in those groups.
- In the third stage, the nonresponse-adjusted weights were post-stratified so that the NSAF sample estimates agreed with independent population totals derived from U.S. Census Bureau sources<sup>6</sup> on the number of persons by age, education, ethnicity, gender, race, and housing tenure. This was done for each study area and for the nation as a whole.

These three stages incorporate screener data to create household weights and extended interview data to create the person and family weights. The weights account for the unequal probability of sampling (at both the household and person levels) and include adjustments for nonresponse and undercoverage. The final result is a series of estimates consistent with Census Bureau population totals that reduce biases due to undercoverage and nonresponse. In some cases, the adjustment to Census Bureau population controls may also reduce the sampling error of the estimates.

---

<sup>6</sup> In Round 1, two sets of weights were released: One that adjusted for the Census undercount and one that did not. In Round 2, only undercount-adjusted weights will be released.

Weights were attached to each responding case to facilitate producing approximately unbiased estimates for each study area and for the entire nation using the 1999 NSAF data. Estimates for 1999 are called cross-sectional because they provide a mechanism for making inferences about the units in the population at that point in time. The 1997 NSAF Snapshot Survey Weights, Report No. 3, described the weights and procedures for making cross-sectional estimates for that year. (See also Report No. 14 in the 1997 NSAF Methodology series.) In addition to these estimates at two specific points in time, estimates of changes in the population between 1997 and 1999 can be produced using data collected in both rounds of the NSAF.

Because, as we have seen, households and persons were sampled with differential probabilities, the use of weights is essential to produce cross-sectional and change estimates that are representative of the population. The next section describes procedures that should be followed to produce cross-sectional and change estimates from the NSAF, focusing on which weights should be used in different circumstances. In order to make appropriate statements based on the NSAF data, researchers should also be aware that some questions are asked of subsets of respondents, and this must be taken into account to make valid statements from the data.

Approximate unbiased estimates of characteristics of persons and families in the study areas and for the nation can be produced by appropriately weighting the survey responses. The estimates are of the noninstitutionalized<sup>7</sup> population of persons under age 65 in the study areas and in the nation. For families, the estimates are limited to those families with at least one person who is under 65 years old. The weights used to produce study area estimates are the same as those used to produce national estimates.

### 3.2 Description of the NSAF Weights

Five categories of weights are currently available with the NSAF data, each appropriate for a different set of respondents or group of questions from the survey. The five categories of weights are: the focal child weights, the adult pair weights, the random adult weights, childless adult weights, and family weights. These weights are described briefly in the next section, with further discussion and examples of using the weights in subsequent sections.

The **focal child (FOCALCHD) weights** are the weights developed to enable users to produce estimates of the number and characteristics of children less than 18 years old. These weights, which should be used to produce most estimates of children, include:

- **WGFCAD0:** This is the full sample weight for focal child variables.

---

<sup>7</sup> Most persons living in group quarters (housing with many unrelated persons such as boarding houses) were also excluded.

- WGFCAD1-WGFCAD60: These replicate weights for focal child variables can be used in variance estimation, as described in Report No. 4 of both the 1997 and 1999 NSAF Methodology Series.

The child weights include factors that adjust for the probability of selecting the child (including differential factors by reported poverty level on the screener and the number of children per household), and nonresponse at the household and person level. Furthermore, the weights are adjusted to be consistent with known totals of the number of children by race, Hispanic ethnicity, age, sex, and tenure (renting or owning a home) for each study area and the nation.

While the WGFCAD0, WGFCAD1 – WGFCAD60 weights can be used for most child based estimates, users interested in child based estimates on many of the child care item in the survey should use an alternative set of weights. Note that most of the questions on child care arrangements ask about care during the last month. In both rounds of the NSAF, some interviews were conducted during the summer months, when care arrangements could differ from those during the school year. Furthermore, questions about child care that were asked during the summer months differed between the two rounds.

Analysts should use the focal child school year weight, WSFCAD0 and replicate weights, WSFCAD1 – WSFCAD60, for both rounds, to get estimates on child care arrangements during the school year.

- WSFCAD0: This is the basic weight for obtaining estimates about child care arrangements during the school year.
- WSFCAD1-WSFCAD60: These are the replicate weights that that can be used for variance estimation for estimates about child care arrangements during the school year.

The **adult pair (ADULT\_PR) weights** are the weights developed to produce estimates of all adults 18 to 64 years old, for most of the questions relevant for adults in the NSAF. When an adult was sampled from a household, most of the questions about adults were asked about both the sampled adult and his/her spouse/partner, if the spouse/partner lived in the household. Because both the adult and the spouse/partner were effectively sampled as a pair for these questions, these weights are called the adult pair weights. For each study area and the nation, these weights adjust for the probability of selection (of the pair of adults), nonresponse, and adjustments to known totals of the number of adults by race, Hispanic ethnicity, age, sex, tenure, and educational attainment. ADULT\_PR weights include:

- WGPRAD0: This is the basic weight for ADULT\_PR variables.
- WGPRAD1-WGPRAD60: These replicate weights for ADULT\_PR variables can be used in variance estimation, as described in Report No. 4 of both the 1997 and 1999 NSAF Methodology Series.

Analysts should use the MKA school-year weight, WSMKAD0 and replicate weights, WSMKAD1 – WSMKAD60, for both rounds, to get estimates on adult-level child care measures during the school year (e.g., child care expenses).

- WSMKAD0: This is the basic weight for obtaining estimates about child care measures during the school year.
- WSMKAD1- WSMKAD60: These are the replicate weights that that can be used for variance estimation for estimates about child care measures during the school year.

The **random adult (ADULT\_RN) weights** are the weights developed to produce estimates of 18- to 64-year-old adults for questions in sections E and F dealing with health insurance over the last 12 months (*items E37-E43 and section F*). In households with children, these items were asked only of either the respondent or his or her spouse/partner, but not both. In households without children, these questions were asked about both members of the pair. The probability of selection of the adult for the ADULT\_RN variables is the same as for the ADULT\_PR, except for MKAs with a spouse/partner in the household. In this latter case, the probability of selection for the adult randomly selected is generally half the pair selection probability. Thus, the ADULT\_RN weight is generally twice the ADULT\_PR weight. The weighting procedures and control totals are the same for the ADULT\_RN weights as for the ADULT\_PR weights. The ADULT\_RN weights include:

- WGRNAD0: This is the basic weight for ADULT\_RN variables.
- WGRNAD1-WGRNAD60: These replicate weights for ADULT\_RN variables can be used in variance estimation, as described in Report No. 4 of both the 1997 and 1999 NSAF Methodology Series.

The third category of adult weights consists of the **childless adult weights**, used to produce estimates of adults without children in the household for a few questions. These questions were asked only of respondents and never of the respondent's spouse/partner. Adults without dependent children were randomly selected from among all childless adults, so they (and their spouse/partners) all had a chance of being asked the questions. The childless adult weights are appropriate for these adults, for the questions asked only of respondents. The childless adult weights were produced by modifying the adult pair weights for the probability of selecting the particular member of the pair, without further adjustment to control totals. The childless adult weights include:

- WGRBAD0: This is the basic weight to be used with adults without children.
- WGRBAD1-60: These replicate weights for childless adults can be used in variance estimation, as described in Report No. 4 of both the 1997 and 1999 NSAF Methodology Series.

Note that there is no weight that can be used to generalize to all adults with children for items asked only of respondents. Since MKA respondents were not randomly selected, the spouse/partner of the MKA

(often the male spouse/partner) had no chance of being sampled and thus no estimates for adults with children are possible for these questions.

The **family weights** are appropriate for making estimates of the number and characteristics of families that have at least one member who is under 65 years old. Families in which all members are over 65 are not included in the NSAF. Families studied by the NSAF were defined using two alternate sets of rules. A social family includes not only married partners and their children, but also unmarried partners, all of their children, and members of the extended family (anyone related by blood to the MKA, his/her spouse/partner, or their children). The second family definition, the CPS family is limited to the householder, spouse of the family householder, children in the family, and other relatives of the family household respondent.

The family weights are derived from the ADULT\_PR weights, with adjustments for the probability of sampling the family. No family-level control totals are applied, but the weights using the control totals from the ADULT\_PR weights are the basis for this weight. In most cases, the ADULT\_PR weights for the MKA and the family weight are identical. However, in some families more than one person is an MKA and in these cases the family weight may not be equal to the ADULT\_PR weights for either of the sampled MKAs. Even in common situations, such as a family with one parent and a child who is over 17 years old, the ADULT\_PR weights and the family weight will not be identical. The family weight should only be used for estimates of the number and characteristics of families. Family weights include:

- WCPSAD0: This is the basic weight for CPS family variables.
- WCPSAD1-WCPSAD60: These replicate weights for CPS family variables can be used in variance estimation, as described in Report No. 4 of both the 1997 and 1999 NSAF Methodology Series.
- WSOCAD0: This is the basic weight for social family variables.
- WSOCAD1-WSOCAD60: These replicate weights for social family variables can be used in variance estimation, as described in Report No. 4 of both the 1997 and 1999 NSAF Methodology Series.

Even though most households have no more than one family, households and families are not equivalent. For example, households of exactly one person are not family households by definition. Thus, the number of families is substantially less than the number of households. No household weight is available for analysis, although one was created as a part of the weighting process. Rather, for the few questions that are asked at the household level, the data can best be analyzed by changing the unit of analysis to a different population. For example, rather than estimate the percentage of households with at least one person born outside the United States, estimate the number of persons living in households in which at least one person was born outside the United States.

### 3.3 Selecting the Unit of Analysis

One of the first tasks facing the researcher is to determine the appropriate unit of analysis or the population of interest. Once the unit of analysis is determined, the choice of the appropriate weight is relatively simple. If the child is the unit of analysis, the child weight is appropriate; if the family is the unit, then the family weight is used; and, if the adult (or MKA) is the unit, then the adult pair weight, the random adult weight, or the childless adult weight is used.

#### *Example 1. Characteristics of Children*

The focal child weights are used for virtually all estimates dealing with children. Several examples are the number of children who are male, the percentage of children who are in a specific grade in school, the percentage of children who live in a family that owns a car, the percentage of 14- to 17-year-olds who work, and the percentage of children with an MKA who reports the family has problems paying for food. Notice that some of the examples were of subgroups of children, and no special consideration is needed for these types of estimates. The following SAS programming example demonstrates how to estimate the number of children living in various family structures.

```
PROC FREQ DATA=focal chd;  
  TABLES ufamstr;  
  WEIGHT wgfcd0;  
RUN;
```

Generates the following table:

#### Living arrangements of children

UFAMSTR	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	2801410	3.9	2801410	3.9
2	17798569	24.8	20599979	28.7
3	5597077	7.8	26197056	36.4
4	45693947	63.6	71891003	100.0

Frequency Missing = 73147.033602

For many statistics available from the NSAF, researchers must choose between family-level estimates and person-level estimates. For example, it is possible to estimate either the number of children who live in a family that is below FPL, or the number of families with children that are below FPL. In many situations, the former is the preferred statistic because it gives information about the number of children



irrespective of the number of children per family. If the researcher chooses to present the child estimate, the child weight is appropriate rather than the family weight.

A related situation arises for estimates about characteristics related to the MKA of a child. For example, the percentage of children who have an MKA who reports worrying about having enough food (*M9A*) or the percentage of children with an MKA who feels calm and peaceful (*item P1b*) could be estimated by using the child weight.

### 3.4 Applying the Correct Weight for Adult Estimates

For producing simple estimates for adults (18 to 64 years old), it is necessary to take into account the person for whom the survey item of interest was asked. NSAF items about adults are asked by design of 1) the respondent and the spouse/partner, 2) the respondent only, or 3) the respondent or the spouse/partner (for MKA interviews).

#### *Example 2 - Characteristics of All Adults*

For questions asked of both the respondent and the spouse/partner, the adult pair weight is used. Examples are: the percent of adults with a high school education, the number of adults born outside the U.S., and the percent of adults who live in households with children. Subgroup estimates such as the percent of Hispanic adults who are currently employed can also be made using this weight. The following programming example demonstrates how to estimate the percentage of adults with a high school education.

```
PROC FREQ DATA=adult_pr;
  TABLES l_hsdip;
  WEIGHT wgprad0;
  RUN;
```

Generates this table:

Earned high school diploma				
LHSDIP	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	7792609	63.7	7792609	63.7
2	4447724	36.3	12240334	100.0

Frequency Missing = 154384747.33

For questions asked randomly of either the respondent or the spouse/partner (*items E37-E43 and F1-F29*), the ADULT\_RN weight is used. Some examples of such estimates are

the number of times the individual visited the doctor in the last 12 months and the percentage of adults who had health insurance continuously for the last 12 months. Once again, subgroup analysis presents no special problems, but the ADULT\_RN, and not the ADULT\_PR weight, should be used whenever a question is asked only of the randomly selected adult. The following programming example demonstrates how to estimate the percentage of adults who postponed getting medical care at some point during the last 12 months.

```
PROC FREQ DATA=adult_rn;
  TABLES fwhmed;
  WEIGHT wgrnad0;
RUN;
```

Generates the following table:

Postponed medical care last year				
	FWHMED	Frequency	Percent	Cumulative Frequency
				Cumulative Percent
1	12645694	7.6	12645694	7.6
2	1.5398E8	92.4	1.6663E8	100.0

### Example 3. Characteristics of MKAs

Questions asked of all MKAs can be analyzed using the adult pair weight. This includes questions asked of the MKA and the spouse/partner of the MKA and questions asked only of the MKA. Two examples in which the adult pair weight is the correct weight are: the percent of MKAs that reported arguing with their children, and the percent of MKAs that never attended religious services in the last 12 months. Estimates of any subgroup of MKAs such as MKAs that are under 40 years old can also be made using this weight. Note that MKAs who are outside the 18- to 64-year-old age range are not represented in these estimates.

If the characteristic of the MKA is asked only for the randomly selected adult (*items E37-E43, F1-F29*), then the random adult weight should be used. For example, the number of MKAs, aged 18–64, that had more than one visit to the emergency room in the last 12 months is estimated using the random adult weight. In the following example, the ADULT\_PR weight is used to estimate the number of MKAs who were working at more than one job at the time of the survey. Note that a WHERE statement is used to subset out those adults who are MKAs.

```

PROC FREQ DATA=adult_pr;
  TABLES ijobs;
  WHERE perstype in (4, 34, 47, 48);
  WEIGHT wgprad0;
RUN;

```

Produces the following table:

**More than one job now**

IJOBS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	1485330	6.2	1485330	6.2
2	22450772	93.8	23936102	100.0

Frequency Missing = 15532013.876

#### *Example 4. Characteristics of Adults in Households With No Children*

For most questions, adults living in households with no children can be handled as a subgroup and the ADULT\_PR weight or ADULT\_RN weight is appropriate, as discussed in example 2. However, a few questions were asked only of respondents and estimates of adults without children in the household. For these items, the childless adult weight must be used. For example, the percentage of adults without children in the household who never attended religious services in the last 12 months can be estimated using childless adult weight. For estimating characteristics of subgroups that are based on these questions, even if other questions involved in the estimates are asked of all adults, the childless adult weight must be used.

The following SAS example uses the ADULT\_PR weight to estimate the number of adults from childless households who took college courses. Note that to separate the adults living in households with no children, a SAS proc merge by household ID is employed. Those adults whose household IDs have no match on the focal child dataset live in households with no children. The PERSTYPE variable cannot be employed as in the previous example, since it does not differentiate childless adults living in households with children (referred to as Option B stragglers) and childless adults living in households with no children.

```

PROC SORT DATA=adult_pr OUT=aprttemp;

```

```

BY hhi d;
RUN;

PROC SORT DATA=focal chd OUT=fctemp;
BY hhi d;
RUN;

DATA wki ds wki ds fcnotapr;
MERGE aprtemp(in=a) fctemp(in=b);
BY hhi d;
IF a AND b THEN OUTPUT wki ds;
IF a AND not b THEN OUTPUT wki ds;
IF b AND not a THEN OUTPUT fcnotapr;
RUN;

PROC FREQ DATA=wki ds;
TABLES lwhcrdt;
WEIGHT wgprad0;
TITLE 'Adults from Households without Children Took College
Courses';
RUN;

```

Produces the following table:

Took college courses				
LWHCRDT	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	12164683	82.9	12164683	82.9
2	2510555	17.1	14675238	100.0

Frequency Missing = 72350818.753

#### Example 5. Characteristics of Families

Possible estimates of families include the percentage of families with children and the percentage of families with children and two parents. Subgroup analysis of families can also be conducted. As described in earlier examples, sometimes the preferred method is to make estimates of children (focal child weight) in families with a certain characteristic or adults (ADULT\_PR weight or ADULT\_RN weight) in families with a certain characteristic instead of making family estimates directly. However, caution is needed when making family estimates so that characteristics of a specific individual are not presumed to hold for the entire family. For example, even if the sampled adult did not have any emergency room

visits in the last year, this does not mean that no one in the family did. The following programming example demonstrates how to estimate the percentage of CPS families with income below the poverty level.

```
PROC FREQ DATA=cpsfam;
  TABLES uincrpov;
  WEIGHT wcpsad0;
RUN;
```

Generates the following table:

CPS family income as % of poverty				
UINCRPOV	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0. 5	3029254	4. 5	3029254	4. 5
1	4673934	6. 9	7703188	11. 3
1. 5	5778429	8. 5	13481617	19. 8
2	5565786	8. 2	19047403	28. 0
3	11429562	16. 8	30476965	44. 8
4	37573648	55. 2	68050614	100. 0

### 3.5 Combining Adult Data Files

When choosing among the three adult weights, bear in mind that if the characteristic can only be obtained from a subset of the sampled adults (respondents and spouse/partners), then the weight for that subset must be used in the analysis.

#### *Example 6. Characteristics of Adults from Merged Files*

A more complex example is the percentage of adults who are currently employed and have had continuous health coverage for the past 12 months. Current employment is available for all adults and could be estimated using the ADULT\_PR weight, but continuous health coverage over the past 12 months is only available for randomly sampled adults and requires the use of the ADULT\_RN weight. Since the smallest subset is the randomly sampled adults, the ADULT\_RN weight must be used in this analysis. Thus, the following SAS code:

```
PROC SORT DATA=adult_rn OUT=arntemp;
  BY persid;
RUN;
```

```
PROC SORT DATA=adult_pr OUT=aprtemp;
  BY persid;
RUN;
```

```
DATA adult anotb bnota;
  MERGE arntemp(in=a) aprtemp(in=b);
  BY persid;
  IF a AND b THEN OUTPUT adult;
  IF a AND NOT b THEN OUTPUT anotb;
  IF b AND NOT a THEN OUTPUT bnota;
RUN;
```

```
PROC FREQ DATA=adult;
  TABLES eccovt*iempnow/nofreq;
  WEIGHT wgrnad0;
RUN;
```

Generates this table:

Percent Row Pct Col Pct	1	2	Total
1	68.88 77.72 88.21	19.74 22.28 90.08	88.62
2	9.21 80.90 11.79	2.17 19.10 9.92	11.38
Total	78.08	21.92	100.00

Frequency Missing = 26968310.401

The same principle holds for doing subgroup analysis. If any of the characteristics that define the subgroup or are involved in the estimates within the subgroup are from the randomly sampled adult, then the random adult weight must be used.

### 3.6 Combining Adult and Child Data files

Obtaining estimates for all persons under 65 is fairly straightforward, in that the weight chosen should correspond to whether or not the survey items of interest include those among items *E37–E43* and *F1–F29*. If so, the random adult weight (and accompanying data items) must be concatenated with the child weight (and data) to produce the appropriate estimates. If the adult survey item is not among items *E37–E43* or *F1–F29* and can be asked of both the respondent and spouse/partner, the adult pair weight should be used.

*Example 7. Characteristics of Persons Under 65 from Merged Files*

To estimate the number of persons under 65 (children as well as adults) with a given characteristic, concatenate the child and appropriate adult files, rename the weights so that they are the same, and then carry out the analysis. For example, to estimate the number of persons under 65 who live in a family with income below 200 percent of the poverty level, merge together the child and adult pair files and rename the focal child weight and the ADULT\_PR weight to be, say, TOTW. The analysis can then proceed with TOTW. The SAS code for this process is:

```
DATA focal chd;
  SET focal chd (RENAME=(wgfcad0=totw));
RUN;

DATA adul t_pr;
  SET adul t_pr (RENAME=(wgprad0=totw));
RUN;

DATA all;
  SET focal chd adul t_pr;
RUN;

PROC FREQ DATA=all;
  TABLES ui ncrpov;
  WEIGHT totw;
RUN;
```

This code produces the following table:

CPS family income as % of poverty				
UINCRPOV	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0. 5	13328121	5. 6	13328121	5. 6
1	18353705	7. 7	31681826	13. 3
1. 5	21607979	9. 1	53289805	22. 3

2	21707700	9. 1	74997506	31. 4
3	41810227	17. 5	1. 1681E8	49. 0
4	1. 2178E8	51. 0	2. 3859E8	100. 0

If the analysis was to be of persons under 65 who were uninsured at some time in the last year, then the child (with the focal child weight) and the random adult (with the ADULT\_RN weight) files would have to be concatenated. This follows because the insurance question is only asked of random adults.

### 3.7 Subgroup Analyses

The other principle in the choice of which weight is appropriate depends on the subgroups the questions are about—in other words the questionnaire skip pattern. In the examples above, we have discussed items that are asked of major subgroups, such as the randomly selected adult or the respondent or focal children. But for NSAF survey items, the skip pattern will further determine the items that are asked of more distinct subgroups.

An example that brings this point out more clearly involves the few questions that are asked only about the male member of an adult respondent and spouse/partner pair.<sup>8</sup> Estimates of adult males can generally be considered as a simple subset and the rules regarding the use of the adult pair weight or the random adult weight can be applied as described earlier, (i.e. it depends on the questions being analyzed).

Some estimates involve questions asked only about the male in a spouse/partner situation, but these items can still be analyzed using the ADULT\_PR weight. These include items identifying whether or not the male respondent or male spouse/partner has any children under 18 living outside the household and whether or not they made payments to support their children outside the household. To estimate the number of males who have children under 18 years old living outside the household, the adult pair weight is used. Similarly, the number of adult males paying child support can also be estimated using the adult pair weight.

#### *Example 8. Characteristics of Adult Males Under 65*

Only adult males are asked if they have any children under 18 years old who live outside the adult's household. To estimate the number of males who have children under 18 years old living outside the household, the ADULT\_PR weight is used. This easily extends to questions only asked of the randomly selected adult. The following example estimates the percentage of males who have children under 18 years old living outside the household who have had insurance continuously for the past 12 months; the ADULT\_RN weight is used.

---

<sup>8</sup> In 37 cases both the MKA and spouse/partner were male and these questions were only asked about the respondent. This results in a very small bias in these estimates.



```

PROC SORT DATA=adult_rn OUT=arntemp;
  BY persid;
RUN;

PROC SORT DATA=adult_pr OUT=aprtemp;
  BY persid;
RUN;

DATA adult anotb bnota;
  MERGE arntemp(in=a) aprtemp(in=b);
  BY persid;
  IF a AND b THEN OUTPUT adult;
  IF a AND NOT b THEN OUTPUT anotb;
  IF b AND NOT a THEN OUTPUT bnota;
RUN;

PROC FREQ DATA=adult;
  TABLES eccovt/nofreq;
  WEIGHT wgrnad0;
  WHERE dki dohh = 1;
  TITLE 'Adult Males with Children Outside Household Covered by
  Insurance Continuously During Last 12 Months';
RUN;

```

Generating the following table:

**Adult Males with Children Outside Household Covered by Insurance Continuously  
During Last 12 Months**

	ECCOVT	Frequency	Percent	Cumulative Frequency	Cumulative Percent
	1	4679353	87.0	4679353	87.0
	2	701971.5	13.0	5381324	100.0

Frequency Missing = 2269856.5608

### 3.8 Change Estimates

The NSAF data collected in 1997 and 1999 is a valuable source for estimating changes over the two-year time period. Estimates of change such as the difference in the percentage of children under age 6 who live in families below the poverty level can be produced using the data from Rounds 1 and 2. These are called estimates of net change. Estimates of net change can be produced by calculating separate

estimates for each round of NSAF and then the difference between the two rounds. No special adjustments to the weights are required.<sup>9</sup>

As noted earlier in the discussion of the NSAF sample design, the round 2 sample of telephone numbers consisted of a partial overlap of telephone numbers that were used in round 1 as well as a new sample of telephone numbers. Furthermore, the telephone numbers from round 1 that were used in round 2 were based only on the screener result code. This means that in most cases where an extended interview was obtained in round 2 from a telephone number that was used in round 1, an extended interview was not completed in round 1. Furthermore, we did not follow up with persons interviewed in round 1 as you would in a true panel survey. The overlap of telephone numbers for the RDD sample (and segments for the area sample) was done for the purpose of improving the precision of estimates of change over time, not for the purpose of estimating gross change, or change at the individual level.<sup>10</sup> As such, we do not provide longitudinal weights that would allow you to produce gross change estimates.

---

<sup>9</sup> Estimating the standard error of estimates of net change are covered in the next chapter.

<sup>10</sup> We have decided not to release information that would allow analysts to link persons between rounds of the NSAF in public use files for confidentiality reasons.

## Chapter 4: Calculating Standard Errors

The sample of households and persons obtained in a survey is just one of many possible samples that could have been obtained. Sampling error refers to error in survey estimates that arise due to the fact that estimates are based on a sample of observations rather than the population of observations. This form of error is usually expressed in terms of the standard error of an estimate, or the square of the standard error, the sampling variance. Standard errors are also required to conduct hypothesis tests or tests of statistical significance. Any honest presentation of estimates from a survey or hypothesis testing should include measures of uncertainty associated with using a sample for inference as opposed to the entire population.

In this chapter, we show how to obtain standard errors for estimates using NSAF data. The NSAF sample design features stratification, clustering and oversampling. Specialized software or procedures must be used in order to obtain standard errors that reflect these aspects of the sample design. While survey estimates obtained from standard statistical packages will be correct, standard error estimates from these packages will be incorrect (and most often understate the true standard error). For researchers without the technical ability to calculate standard errors appropriately, we provide instructions on how to use design effects to obtain approximate standard errors for some survey estimates. Finally, we briefly describe how standard errors can be calculated correctly using several statistical packages, including Wesvar, SAS and STATA.

### 4.1 Limitations of Standard Statistical Packages

Note that standard statistical packages such as SAS or SPSS can still be used to obtain approximately unbiased estimates from NSAF data. You can use the WEIGHT option for any number of SAS PROC steps (as shown in the previous chapter) to obtain correct estimates in SAS. In SPSS, you can use the Weight Cases option under the Data menu to obtain weighted estimates. So, if all you are interested in is getting weighted estimates, standard statistical packages can be used in a straightforward manner.

However, you may want to get some sense of the uncertainty around these estimates due to the fact that you have only one sample out of many possible samples that could have been drawn. Or, you may want to construct confidence intervals or conduct hypothesis tests. Most standard statistical packages such as SAS, SPSS, or STATA use formulas to calculate standard errors which assume that the data are from a simple random sample. More formally, observations are assumed to be independent of one another and identically distributed (iid). Several features of the NSAF sample design however should convince you that NSAF data are not from a simple random sample, including oversampling by study area, oversampling by household type and household income, clustering in the area sample of nontelephone households, subsampling within households, having observations (sample persons) within the same household, and for the second round, differential sampling by round 1 screener result codes.

## 4.2 Using Design Effects for Approximate Standard Errors

Researchers without the technical ability to calculate variances on their own can consider employing design effects, an alternate method of describing the variability of an estimate from a survey. The term design effect (DEFF) is used to describe the variance of sample estimates for a particular sample design, relative to the corresponding variance of a simple random sample with the same sample size. DEFFs are used to evaluate the efficiency of the sampling design and estimation procedure used to develop the estimates.

The concept of the design effect was popularized by Kish (see, for example, Kish 1965) to deal with complex sample designs involving stratification and clustering, designs like that of the NSAF. Stratification generally leads to a gain in efficiency over simple random sampling. On the other hand, clustering usually leads to deterioration in efficiency. This latter effect arises due to positive intraclass correlation among the subunits in the clusters. For example, the DEFF is larger for children because we sometimes sampled more than one child from the same household. This clustering effect increases the variance over that which would pertain in a simple random sampling of children. There is also a stratification effect to consider in the NSAF. By oversampling Mississippi, for instance, we obtain excellent results for that state—roughly as good as those for the much larger California. However, this oversampling means that our estimates of the nation as a whole are not as good as if we had drawn a simple random sample of the country as a whole.

In order to determine the total effect of any complex design on the sampling variance in comparison with the alternative simple random sampling, one calculates a ratio of variances associated with an estimate, namely:

$$\text{DEFF} = \frac{\text{sampling variance of a complex sample}}{\text{sampling variance of a simple random sample}}$$

This ratio is called the design effect of the sampling design for the estimate. This ratio measures the overall efficiency of the sampling design and the estimation procedure used to develop the estimate. At the analysis stage, the DEFF is useful because most statistical analysis software (such as SAS and SPSS) assumes the data are from a simple random sample when computing sampling errors of estimates. The DEFF can, in some circumstances, indicate how appropriate this is and can be used to adjust these simple estimates to produce ones that are closer to the actual sampling errors of the estimates (Skinner, Holt, and Smith 1989).

For example, the design effect for a proportion can be expressed as:

$$\text{DEFF} = \frac{\text{Var}_{des}(p)}{\text{Var}_{srs}(p)}$$

Where:

$p$  denotes the weighted estimate of the population proportion  $P$ ,

$\text{Var}_{srs}(p)$  is the estimated simple random variance  $v(p)_{srs} = \frac{p(1-p)}{n}$ , and

$\text{Var}_{des}(p)$  is the variance of the complex sample calculated appropriately

In the NSAF and in most other large-scale surveys, a large number of data items or variables are collected from respondents. Each variable has its own design effect. One way to represent all of these is to compute design effects for a number of similar variables and then try to generalize about the impact of the complex sample design. Appendix tables B-1 through B-8 enable us to do this by showing the average, maximum, and minimum design effects for 33 estimates of children and adults from the 1999 NSAF (For the corresponding tables from the 1997 NSAF, see 1997 NSAF Methodology Report No. 4).

In most cases, design effects for complex samples are larger than one. This is true of the NSAF, with some design effects considerably greater than one, especially those for the nation as a whole, where the DEFFs range from 0.81 to 10.02. The most important factors that result in design effects larger than one in the NSAF include:

1. **Oversampling by Study Area.** The need for both study area and national estimates required oversampling to produce stable separate estimates for the 13 specified study areas. This oversampling increased the design effect for national estimates.
2. **Household Screening.** Additional variability comes from the subsampling of households without children and those above 200 percent of the federal poverty level (FPL). The misclassification of households as above 200 percent FPL when they actually fell below 200 percent of the poverty level also increases the variance of estimates restricted to persons at or below 200 percent of the poverty level. See Flores-Cervantes et al. (1998) for a discussion of this topic.
3. **Within-Household Subsampling.** Differential sampling rates at the person level also contribute to increases in design effects. Children and adults were subsampled within households for both the RDD and area sample components.<sup>11</sup>
4. **Differential Sampling Rates of Round 1 Telephone Numbers.** Telephone numbers from Round 1 were subsampled at different rates in Round 2, depending on their result codes in that round of data collection. This increases the design effects for Round 2 estimates somewhat but improves precision for estimates of change and reduces overall data collection costs.

---

<sup>11</sup> Households without children and households above 200 percent of the poverty level were not subsampled in the area component of the study, so design effects associated with such subsampling were not incurred in the area sample but were for the RDD sample.

5. **Clustering of Households in the Area Sample.** For the area sample component of the survey, households were clustered within segments and segments were clustered within primary sampling units (PSUs). Design effects increase to the extent that respondents in the same cluster are similar in their responses to survey items. Larger design effects resulting from the area sample clustering are more likely to affect low-income households because a larger percentage of low-income persons are in nontelephone households.

A final point about the DEFF tables is that they show Milwaukee and the balance of Wisconsin separately. Using the publicly available site variable (SITE), researchers can treat Milwaukee and the balance of Wisconsin as separate study areas. Alternately, researchers can simply use the larger of the two sets of design effects shown, or employ the replicate structure of the NSAF files to calculate variances directly.

For some of the files in this public use release, the average DEFTs shown above can be used directly by calculating from the file an unbiased estimate of the simple random sampling error. Below, we have carried out an extended example in detail.

We begin by modifying a conventional 95 percent confidence interval for the population proportion  $P$ . This modification is of the form

$$p \pm 1.96(\text{DEFT}) (\text{Var}_{\text{SRS}}(p))^{1/2},$$

where  $p$  is the estimate from NSAF of the true population value  $P$  obtained (as in section 5 above) by calculation of the weighted total. Because we are using a conventional 95 percent confidence interval and under the assumption of normality, the confidence coefficient is 1.96. The DEFT will depend on the particular  $P$  we try to estimate, as set out in the NSAF Appendix tables mentioned above.

$(\text{Var}_{\text{SRS}}(p))^{1/2}$  is an estimate of the standard error of  $p$  under simple random sampling (SRS). It can be useful to think of the SRS standard error as

$$\text{SRS standard error} = (\text{population standard deviation})/(\text{unweighted sample size})^{1/2}.$$

For a proportion, this is the familiar  $v(p)_{\text{SRS}} = \frac{p(1-p)}{n}$  that was used above. Notice that for proportions, all that is needed is to properly calculate the weighted estimate  $p$ , then the SRS standard error is immediate and the adjusted confidence intervals follow readily.

Consider the following example, worked out here in detail. In particular, consider estimating average earnings in the previous year, U\_EARN. We first use the SAS PROC MEANS statement

```
PROC MEANS DATA=adul t_PR VARDEF=WDF N SUMMGT MEAN VAR STD;
  VAR u_earn;
  WEIGHT wgprad0;
```

**TITLE 'Total Earnings Last Year Using (Sum of Weights) 1 to  
Calculate the Variance';  
RUN;**

to obtain

n	Sum Wgt	Mean	Variance	Std Dev
27, 599	37, 298, 973. 02	18, 873. 51	901, 697, 957	30, 028. 29

The simple random sampling standard error is then:

$(\text{population standard deviation})/(\text{unweighted sample size})^{1/2} = (30,028.29)/(27,599)^{1/2}$ .

This calculation yields 181.55. Since U\_EARN average = 18,873.51 and from Appendix table B-4, DEFT = 2.16, the final confidence interval is:

$$18,873.51 \pm 1.96 * 2.16 * (181.55)$$

or

$$18,873.51 \pm 768.10$$

It might be worth noting that our basic approach here is similar to that taken in Census Bureau publications from the CPS (e.g., see P-60, No. 198, which is the CPS publication most comparable to the 1997 NSAF study).

### 4.3 Methods for Obtaining Correct Standard Errors

The design effect approach can be used to obtain approximate standard errors for percentages, proportions and means. However, they cannot be used for estimates such as ratios, regression coefficients and totals. In addition, design effects have only been provided for some specific subgroups. Finally, the average design effects shown in Appendix B are based on a relatively small number of estimated variances from the survey, and these variances are also sample estimates. The particular estimates used to generate the tables in Appendix B also affect the averages. The estimates used were specifically selected, and many of them are related to lower income status. Other choices of estimates would give different averages. It is recommended that you use the methods described in this section to obtain optimal estimates of standard errors.<sup>12</sup>

---

<sup>12</sup> A good starting point for learning about variance estimation in complex design samples with links to software can be found at the website of the Survey Research Methods Section of the American Statistical Association (<http://www.amstat.org/sections/SRMS/index.html>).

There are basically two methods for obtaining standard errors for NSAF estimates. You can either use only the 0 weight (e.g. for child estimates, use only the WGFCAD0 weight) along with variables that describe the structure of the replicate weights (VARUNIT and VARSTRAT) or you can use the 0 weight along with the replicate weights (i.e. WGFCAD0, WGFCAD1, WGFCAD2, ... WGFCAD60).

Users who wish to obtain standard errors using packages that rely upon linearization methods (Taylor series approximations) such as in SUDAAN, the “svy” commands in STATA and the PROC SURVEYMEANS and PROC SURVEYREG commands in SAS do not need to use the replicate weights. Instead, you will only use the ‘0’ weight along with the VARUNIT and VARSTRAT variables. We will not deal with using these packages for estimating standard errors in this report. A discussion of using these methods to calculate standard errors is covered in chapter four of Report No. 4 of the 1999 NSAF Methodology Series.<sup>13</sup>

#### 4.4 Using Replicate Weights to Calculate Standard Errors

The basic idea behind replication is to draw subsamples from the sample, compute the estimate from each of the subsamples, and estimate the variance from the variability of the subsample estimates. Specifically, subsamples of the original *full* sample are selected to calculate subsample estimates of a parameter for which a *full-sample* estimate of interest has been generated. The variability of these subsample estimates, around the estimate for the full sample, provide an estimate of the standard error of the estimate. The subsamples are called replicates and the estimates from the subsamples are called replicate estimates. Balanced repeated replication (BRR) and jackknife replication are two approaches to forming subsamples. Rust and Rao (1996) discuss these and other replication methods, show how the units included in the subsample can be defined using variance strata and units, and describe how these methods can be implemented using weights.

Replicate weights are created to derive the corresponding set of replicate estimates. Each replicate weight is derived using the same estimation steps as the full sample weight but using only the subsample of cases comprising each replicate. Once the replicate weights are developed, it is a straightforward matter to compute estimates of variance for sample estimates of interest. Estimates of variance take the following form:

$$v(\hat{\mathbf{q}}) = c \sum_{k=1}^G (\hat{\mathbf{q}}_{(k)} - \hat{\mathbf{q}})^2, \quad (2-1)$$

where

---

<sup>13</sup> SUDAAN version 8 also has an option that allows the user to calculate variances using jackknife replicate weights supplied by the user. We have not tested this version of SUDAAN and cannot provide advice or guidance on using NSAF replicate weights with this package.



$\hat{\mathbf{q}}$	is the estimate of $\mathbf{q}$ based on the full sample.
$\hat{\mathbf{q}}_{(k)}$	is the $k$ -th estimate of $\mathbf{q}$ based on the observations included in the $k$ -th replicate.
$G$	is the total number of replicates formed.
$c$	is a constant that depends on the replication method.
$v(\hat{\mathbf{q}})$	is the estimated variance of $\hat{\mathbf{q}}$ .

Thus, imagine using each of the 60 replicate weights, one replicate weight at a time, to obtain 60 separate weighted estimates of the same statistic, such as a mean. Take these 60 estimated means and calculate the sum of the squared deviations from the mean estimated using the full sample weight. This sum of the squared deviations from these 60 means is your estimated variance. In turn, the standard error is the square root of the estimated variance.

This logic applies to any statistic for which you wish to estimate the standard error. For example, if you wish to obtain standard errors of a regression coefficient, you would essentially estimate the same regression model 60 times, once using each of the 60 replicate weights. You would then calculate the sum of the squared deviations of the regression coefficient from the full sample estimate (and then take the square root) to get the estimated standard error of that regression coefficient. Similarly, if you are interested in calculating the standard error of the difference between two means, you would calculate the difference between those means 60 times, once using each replicate weight. You would then calculate the sum of the squared deviations from these 60 differences and take the square root of that to get the standard error of the difference.

Replicate weights were created for both the 1997 and 1999 NSAF using a paired jackknife approach. Full details of creating these replicate weights are provided in Report No. 4 of both the 1997 and 1999 NSAF Methodology Series.

## 4.5 WesVar

WesVar<sup>14</sup> is a package developed by Westat for the personal computer (PC). WesVar uses replication methods such as the jackknife and BRR to compute variance estimates. Through the use of replicates, adjustments made during weighting (nonresponse, raking) can be taken into account by making the same adjustments to each replicate separately. Replication is computer intensive, but powerful PCs have largely eliminated this as an issue. However, it is still possible that for very large data sets the computations will exceed the capacity of the machine or take a long time. Although replication can be used for most estimates, replication techniques are not necessarily appropriate for all sample statistics of interest. Special care is needed when trying to estimate standard errors of medians,

---

<sup>14</sup> The latest version of WesVar is version 4.0. An older version of WesVar that can be used to produce many estimates and standard errors (including regression and logistic regression) is freely available for download at (<http://www.westat.com/wesvar/demo/index.html>).

quartiles, or other quantiles. Direct estimates of quantiles using the jackknife method are not supported, but an alternative method is supported.

WesVar is an interactive program centered on sessions called *workbooks*. A workbook is a file linked to a specific WesVar data set. In a workbook, the user can request descriptive statistics and regression models, as well as analyze and create new statistics. The information about the design is incorporated into the replicate weights when the data file is created. Regression requests support both linear and logistic regression (both dichotomous and multinomial). Outputs include statistics of interest, such as the sum of weights, means, totals, percentages, ratios, regression coefficients, and log odds-ratios, along with their corresponding standard errors, CV, and confidence intervals. Chi-square tests of independence are performed on two-way tables, and goodness of fit statistics are produced for regression models. WesVar can also estimate linear combinations of parameter estimates (e.g. differences and sums of regression coefficients) and perform hypothesis tests. Design effects can be output for all the above statistics except compiled statistics such as ratios.

When you import NSAF data into WesVar, you will not need to create replicate weights since they have already been created. The ‘0’ weight (e.g. WGFCAD0 on the focal child file) is the full sample weight and the weights with names ending in 1 – 60 (e.g. WGFCAD1 – WGFCAD60) are the replicate weights. Specify JK2 as the method of replication.

## **4.6 SAS and STATA Macros**

The SAS macros are programs written by Urban Institute (UI) staff to enable researchers to generate accurate variance estimates from the National Survey of America’s Families (NSAF) without using additional software. By default, lower versions of SAS and other standard statistical packages cannot make ready use of certain complex survey designs for statistical analysis. More specifically, SAS does not have the built-in capability to properly analyze NSAF and its use of replicate weights.

Appendix D describes the macros, their capabilities, and the syntax required to invoke the macros correctly. In the second section of Appendix D, we present examples of sample programs to run the macros, and follow these examples with the output resulting from the submitted statements. The third section contains the actual macro programs. The macro programs in section 3 can be used as is, or can be modified to run other statistical tests (e.g., logistic regression). For basic questions about the macros and their use, please contact [nsaf@ui.urban.org](mailto:nsaf@ui.urban.org). However, please note that while these macros are being made available to external researchers as a convenience, we cannot provide support beyond general technical assistance.

Appendix E describes macros for use in STATA. This section also reviews the commands that are currently developed and their syntax, discusses their limitations, and illustrates their structure using OLS with JRR standard errors as an example. With this example, users of the NSAF with experience programming in STATA should be able to readily extend the method to other regression commands or customize the routines described here.

## 4.7 Estimating Variances for Change Estimates

An important goal of the 1999 NSAF is to estimate changes that have occurred in the population since the 1997 NSAF data collection (e.g., the change in the percentage of children who live in poverty by study area or the percentage of adults without insurance). The retention of a substantial number of telephone numbers and nontelephone sampling areas, as described in 1999 NSAF Methodology Report No. 2 dealing with sample design, was designed to help accomplish this goal by reducing the variance of estimates of change.

The method for computing variances for estimates of change over time in the NSAF is a function of the designs of the 1997 and 1999 surveys. This will be clear once some notation is established. Let  $\hat{\mathbf{q}}_t$  be the estimate of a characteristic or total for time  $t$ , and  $v(\hat{\mathbf{q}}_t)$  be its estimated variance (the square of the standard error). The estimated change between times  $t_1$  (1997) and  $t_2$  (1999) for this characteristic or total is  $\Delta = \hat{\mathbf{q}}_{t_1} - \hat{\mathbf{q}}_{t_2}$ . We are interested in estimating its variance,  $v(\Delta)$ .

If the NSAF sample for Round 1 and Round 2 were selected independently, standard statistical theory could be applied. Under independence, the variance of the difference is the sum of the variances for the two time periods,

$$v(\Delta) = v(\hat{\mathbf{q}}_{t_1}) + v(\hat{\mathbf{q}}_{t_2}) \quad (3.1)$$

The two variances on the right-hand side are computed separately, using the replication procedures described in Report No. 4 of both the 1997 Methodology Series. Similarly, this report (briefly) and the 1999 NSAF Methodology Report No. 4 delineate the procedures for Round 2. The estimated difference and its variance can then be computed simply by using equation (3.1).

This approach is not completely appropriate for the NSAF because the two samples are not independent. In fact, the sample design for Round 2 was deliberately established to make the samples dependent. With dependent samples, the variance of the estimated change has an additional component to account for the correlation in the samples. The estimated variance is

$$v(\Delta) = v(\hat{\mathbf{q}}_{t_1}) + v(\hat{\mathbf{q}}_{t_2}) - 2 \cdot \mathbf{r} \cdot \sqrt{v(\hat{\mathbf{q}}_{t_1})v(\hat{\mathbf{q}}_{t_2})}, \quad (3.2)$$

where the last term accounts for the dependence of the two estimates. When the correlation,  $\rho$ , is large and positive, then the variance of the estimated change may be much smaller than obtained from independent samples. With independent samples, the correlation is zero, and equation (3.2) reduces to equation (2.1). When the samples partially overlap like they do in the NSAF, then correlation is

typically smaller than with a complete overlap (see Kish 1965, section 12.4 for more discussion of this case). Tables of these correlations for NSAF are provided in Appendix C.

One way of estimating change (and the variance of a change estimate) from the NSAF is to concatenate the 1997 and 1999 data files into one file that contains all the variables needed for the analysis, including all of the replicate weights. For example, the variables from the child files for 1997 and 1999 that will be used in the change analysis would be included in the concatenated file. In preparing the concatenated file, the variables for both rounds have been given the same variable name (e.g., poverty status for 1997 and poverty status for 1999 should have the same variable name). The weights for the two rounds are also given the same variable names. In addition, a binary variable that indicates the data collection round should be created.

Estimates of change are the difference between subgroups in the concatenated file, where the subgroups are defined by the round indicator. The issues raised in the previous section that concern making cross-sectional estimates can now be applied directly to producing estimates of change. The choice of the appropriate cross-sectional weight determines the weight to be used to estimate change.

## References

1997 and 1999 National Survey of America's Families Basic Snapshot II Tables. The Urban Institute Web site: <http://newfederalism.urban.org/nsaf>

1997 and 1999 National Survey of America's Families Methodology Reports, The Urban Institute Web site: <http://newfederalism.urban.org/nsaf/methodology.html>.

Black, T., and A. Safir. 2000. "Non-Response Bias in the NSAF," *Proceedings of the Survey Research Methods of the American Statistical Association*. Alexandria, VA.

De Leeuw, E. D., and J. van der Zouwen. 1988. "Data Quality in Telephone and Face-to-Face Surveys: A Comparative Meta-Analysis," *Telephone Survey Methodology*. John Wiley and Sons.

Scheuren, F. 2000. "Quality Assessment of Quality Assessment," *Proceedings of the Survey Research Methods of the American Statistical Association*. Alexandria, VA.

Wang, K, D. Cantor, and A. Safir. 2000. "Panel Conditioning in an Random-digit dial Survey," *Proceedings of the Survey Research Methods of the American Statistical Association*. Alexandria, VA.

## Appendix A Sample Sizes

**Table A-1.  
Round 1 Extended Interview Sample Sizes**

	Children		MKAs and Spouse/Partners		Childless Adults	
State	All	Low Income	All	Low Income	All	Low Income
Alabama	2,098	1,160	2,875	1,386	1,412	569
California	2,060	1,242	2,698	1,471	1,666	596
Florida	2,063	1,176	2,792	1,383	1,231	454
Massachusetts	2,381	1,016	3,314	1,148	2,048	494
Michigan	2,143	960	3,009	1,143	1,756	510
Minnesota	2,360	1,033	3,347	1,250	2,307	634
New Jersey	2,566	1,029	3,577	1,192	2,476	630
New York	2,252	1,230	2,959	1,395	1,381	474
Texas	2,249	1,367	3,079	1,704	1,177	421
Washington	2,469	1,174	3,465	1,443	2,415	777
Mississippi	1,984	1,219	2,612	1,392	1,434	640
Milwaukee	1,804	1,002	2,300	1,053	1,255	420
Wisconsin	2,320	977	3,389	1,217	2,068	563
Bal. of U.S.	3,392	1,808	4,736	2,210	3,322	1,096
Colorado	2,298	1,078	3,258	1,332	2,167	680
Total	34,439	17,471	47,410	20,719	28,115	8,958

**Table A-2.  
Round 2 Extended Interview Sample Sizes**

	Children		MKAs and Spouse/Partners		Childless Adults	
State	All	Low Income	All	Low Income	All	Low Income
Alabama	1,827	891	2,583	1,076	1,373	568
California	1,917	823	2,702	1,019	1,225	432
Florida	1,989	860	2,873	1,050	1,106	400
Massachusetts	2,564	773	3,784	899	1,575	400
Michigan	2,177	735	3,180	883	1,547	467
Minnesota	2,510	738	3,826	906	2,238	530
New Jersey	2,931	812	4,311	949	1,459	305
New York	2,197	966	3,038	1,121	1,125	376
Texas	2,163	1,055	3,068	1,325	665	269
Washington	2,381	822	3,439	988	1,366	366
Mississippi	1,734	927	2,328	1,049	1,186	498
Milwaukee	1,991	802	2,709	850	1,336	436
Wisconsin	2,543	720	3,871	900	1,633	425
Bal. of U.S.	4,557	1,861	6,673	2,349	3,636	1,279
Colorado	2,457	830	3,643	1,041	1,221	355
Total	35,938	13,615	52,028	16,405	22,691	7,106



## **Appendix B**

### **Design Effect Tables**

Each of the eight tables has a row for each state, four columns for all children or adults, and four columns for low-income children or adults. The first column is the average design effect (DEFF), the second is the maximum DEFF, the third is the minimum DEFF, and the fourth is labeled the DEFT. The DEFT is the square root of the DEFF, so it is similar to the DEFF but on the scale of the standard deviation of the estimate rather than the variance. The figures labeled DEFT in the tables are actually the average of the DEFTs.

Table B-1 gives the average DEFTs for all children and for low-income children by study area. The average DEFT for all the study areas is generally in the range of 1.3 to 1.5, or about 30 to 50 percent above what would be found in a simple random sample of the same size. The average DEFT for the national estimate of all children is just under 2.0. For children in low-income families, the average DEFT is almost the same as for all children. These averages (and the maximum and minimum shown in the table) were computed from 25 estimates of all children in each study area and 29 estimates for low-income children.

Table B-2 gives the average DEFTs for Hispanic children. The averages for some of the study areas are based on a very small number of statistics because any estimate with a sample size of less than 10 children was excluded from the computations. For example, in Alabama only five estimates of Hispanic children have sample sizes (the numerator of the estimated percentage) of 10 or more, so the estimated average DEFT for Hispanic children in Alabama is unstable. The small sample size is a consequence of there being so few Hispanics in Alabama. The other study areas where the averages are based on 15 or fewer statistics (of the 25 or 29 for low-income children that could have been used) were Michigan, Minnesota, and Mississippi, and the balance of Wisconsin. The average DEFT for Hispanics is generally slightly lower than the average DEFT for all children shown in table B-1, but it should be remembered that these averages may be unstable.

Table B-3 is the corresponding table of average DEFTs for black children. Only two study areas have averages based on fewer than 15 statistics: Colorado and the balance of Wisconsin. The average DEFTs for all black children and the low-income black children are about the same as the average DEFTs for all children given in table B-1. Given the instability of all these average DEFTs, there does not seem to be a great deal of variability across the all, Hispanic, and black children categories.

Table B-4 shows the average DEFT for estimates of all adults in each study area. The median of these average DEFTs across the study areas is slightly less than 1.5, so the average standard error of the estimate is about 40 to 50 percent greater than expected from a simple random sample. The average DEFT for Colorado is the largest; this is due in part to one statistic with a particularly large DEFT (the maximum DEFF in Colorado is the largest in the table). The average DEFT for the national estimate of adults is 2.2, which is larger than the study area average as expected because of the oversampling by study area. The average DEFT for low-income adults in the table is about the same as for all adults,



with little variation across study areas. All of the averages in the table are based on the same 22 statistics computed by study area and for the nation.

Tables B-5 and B-6 give the estimates for adult Hispanics and blacks, respectively. The average DEFT for Hispanics is very unstable for many of the study areas because the average is computed from only a few statistics. For example, in Alabama and Mississippi, only five of the adult statistics had the sample sizes of 10 or more needed to be included in the average DEFT. Other study areas in which the average DEFT was based on less than 15 statistics for either all adult Hispanics or low-income adult Hispanics are Michigan, Minnesota, and the balance of Wisconsin. For blacks, only Colorado and the balance of Wisconsin have average DEFTs based on less than 15 statistics for all adults or low-income adults. The average DEFTs by study area are relatively consistent for Hispanics and blacks, with some variation by study area.

The last two tables of adults are for those who live in households with and without children (tables B-7 and B-8). The average DEFT for all adults in households with children is very similar to the average DEFT for low-income adults in households with children. In addition, there is little variation in the average DEFTs across study areas. The same holds true for adults in households without children. The average DEFT is smaller for adults in households without children. For example, the average DEFT for all adults in households without children in Florida is 1.2; for adults in households with children, the average DEFT for Florida is 1.6.

**Table B-1.**  
**Average DEFF and DEFT for Estimates from the 1999 NSAF Child File for All Children and Low-income Children, by Site**

Study Area	All Children				Low-income Children			
	Average	Maximum	Minimum	DEFT	Average	Maximum	Minimum	DEFT
Alabama	1.80	2.45	0.86	1.33	1.74	2.85	0.47	1.30
Balance of Wisconsin	2.28	4.45	1.16	1.48	2.25	4.69	0.73	1.45
California	1.89	2.88	0.94	1.36	1.89	3.14	0.55	1.35
Colorado	1.87	2.81	0.79	1.35	1.87	3.00	0.74	1.34
Florida	2.24	6.08	0.86	1.46	2.09	5.84	0.60	1.41
Massachusetts	1.78	2.32	1.06	1.33	1.86	2.76	0.86	1.35
Michigan	2.31	8.67	0.70	1.47	1.96	7.50	0.65	1.35
Milwaukee	1.84	2.69	1.13	1.34	1.83	3.19	0.50	1.33
Minnesota	1.79	5.74	0.86	1.31	1.77	5.61	0.69	1.30
Mississippi	1.78	4.34	1.05	1.32	1.72	4.24	0.55	1.28
New Jersey	1.73	2.56	1.13	1.31	1.74	2.68	0.76	1.31
New York	1.74	2.80	0.84	1.31	1.82	3.22	0.51	1.33
Texas	2.09	3.99	1.03	1.43	1.96	4.11	0.54	1.37
Balance of the U.S.	1.72	3.09	0.93	1.30	1.76	2.78	0.62	1.31
Washington	1.86	3.20	0.90	1.34	1.90	3.18	0.72	1.36
National	3.92	6.04	2.34	1.97	3.91	5.70	1.29	1.96

**Table B-2.**  
**Average DEFF and DEFT for Estimates from the 1999 NSAF Child File for All Hispanic Children and Low-income Hispanic Children, by Site**

Study Area	All Hispanic Children				Low-income Hispanic Children			
	Average	Maximum	Minimum	DEFT	Average	Maximum	Minimum	DEFT
Alabama	2.16	2.52	1.65	1.47	1.56	2.24	0.49	1.21
Balance of Wisconsin	1.48	1.85	1.10	1.21	1.54	1.69	1.43	1.24
California	2.20	3.60	0.87	1.46	1.93	3.53	0.37	1.35
Colorado	2.18	3.65	0.78	1.45	2.01	3.69	0.46	1.37
Florida	1.54	3.20	0.76	1.22	1.46	3.30	0.57	1.18
Massachusetts	1.70	2.53	0.72	1.29	1.66	2.46	0.39	1.26
Michigan	1.74	2.34	0.99	1.31	1.53	2.29	0.91	1.22
Milwaukee	1.89	2.62	0.85	1.36	1.70	2.88	0.27	1.26
Minnesota	1.97	2.45	1.27	1.40	1.33	1.92	0.67	1.14
Mississippi	1.61	2.84	0.99	1.23	1.59	3.10	0.84	1.20
New Jersey	1.80	2.99	1.01	1.33	1.63	2.43	0.46	1.26
New York	1.88	2.66	1.16	1.36	1.71	2.53	0.36	1.29
Texas	2.53	5.07	0.96	1.56	2.18	5.13	0.32	1.43
Balance of the U.S.	1.56	2.20	1.06	1.24	1.45	2.45	0.59	1.19
Washington	1.94	2.83	1.05	1.38	1.80	2.82	0.46	1.32
National	3.46	4.96	1.76	1.84	3.09	4.93	0.81	1.72

**Table B-3.**  
**Average DEFF and DEFT for Estimates from the 1999 NSAF Child File for All Black Children and Low-income Black Children, by Site**

Study Area	All Black Children				Low-income Black Children			
	Average	Maximum	Minimum	DEFT	Average	Maximum	Minimum	DEFT
Alabama	1.75	2.79	0.80	1.31	1.65	2.70	0.26	1.26
Balance of Wisconsin	2.35	2.77	1.69	1.53	1.88	3.01	1.29	1.36
California	1.84	2.47	0.53	1.34	1.92	2.48	1.00	1.38
Colorado	1.72	1.95	1.15	1.31	1.70	2.28	0.57	1.28
Florida	2.34	4.42	0.84	1.51	2.07	3.98	0.78	1.41
Massachusetts	2.01	3.53	1.00	1.40	1.96	3.94	0.98	1.38
Michigan	2.56	9.94	0.50	1.53	1.92	7.23	0.30	1.33
Milwaukee	1.83	2.91	1.06	1.34	1.77	3.19	0.41	1.31
Minnesota	2.23	3.98	1.18	1.47	1.88	3.52	1.16	1.36
Mississippi	1.94	4.41	1.07	1.38	1.77	4.19	0.53	1.31
New Jersey	1.66	2.54	0.97	1.27	1.70	2.61	0.71	1.29
New York	1.93	3.55	0.98	1.37	1.79	4.39	0.60	1.31
Texas	1.75	2.82	0.63	1.30	1.64	2.78	0.50	1.25
Balance of the U.S.	1.73	2.63	0.90	1.30	1.62	2.78	0.71	1.26
Washington	2.36	4.07	1.09	1.52	2.20	4.11	0.82	1.45
National	4.58	7.31	2.35	2.12	4.70	8.43	2.14	2.15

**Table B-4.**  
**Average DEFF and DEFT for Estimates from the 1999 NSAF Adult Pair File for All Adults**  
**and Low-income Adults, by Site**

Study Area	All Adults				Low-income Adults			
	Average	Maximum	Minimum	DEFT	Average	Maximum	Minimum	DEFT
Alabama	2.11	3.24	1.09	1.44	2.14	3.50	0.71	1.45
Balance of Wisconsin	2.15	3.30	1.15	1.45	2.23	3.82	1.41	1.48
California	2.00	3.71	1.12	1.40	2.08	4.12	0.82	1.42
Colorado	3.26	7.19	1.02	1.77	2.75	5.90	1.13	1.63
Florida	2.49	4.09	1.30	1.56	2.49	3.83	1.25	1.57
Massachusetts	2.43	3.96	1.32	1.54	2.42	4.98	1.31	1.54
Michigan	2.25	3.54	1.14	1.48	2.19	3.65	1.09	1.46
Milwaukee	2.29	4.30	0.98	1.49	1.87	2.72	1.12	1.36
Minnesota	1.92	2.80	0.92	1.38	1.92	2.92	1.07	1.37
Mississippi	1.97	4.04	1.03	1.39	2.09	4.81	1.21	1.42
New Jersey	2.77	4.70	1.85	1.65	2.44	4.35	1.12	1.54
New York	2.10	3.46	0.80	1.44	2.38	4.04	0.73	1.52
Texas	3.00	5.21	1.08	1.70	3.07	5.22	1.25	1.73
Balance of the U.S.	2.01	3.36	0.92	1.40	2.07	2.70	0.97	1.43
Washington	2.26	3.13	1.45	1.50	2.14	3.50	1.02	1.45
National	4.92	6.77	2.49	2.20	4.54	7.19	2.52	2.11

**Table B-5.**  
**Average DEFF and DEFT for Estimates from the 1999 NSAF Adult Pair File for All Hispanic**  
**Adults and Low-income Hispanic Adults, by Site**

Study Area	All Hispanic Adults				Low-income Hispanic Adults			
	Average	Maximum	Minimum	DEFT	Average	Maximum	Minimum	DEFT
Alabama	5.58	12.51	2.28	2.26	2.18	2.72	1.46	1.47
Balance of Wisconsin	2.34	3.43	1.21	1.51	1.79	2.50	1.21	1.33
California	1.91	3.68	0.65	1.36	1.90	2.91	0.75	1.36
Colorado	2.57	4.73	1.50	1.58	2.20	4.36	1.34	1.46
Florida	2.81	7.72	1.14	1.64	2.35	3.77	1.09	1.52
Massachusetts	2.32	4.87	0.71	1.49	2.42	4.29	1.33	1.54
Michigan	1.92	4.30	0.76	1.36	1.63	2.49	1.01	1.27
Milwaukee	2.20	3.58	0.97	1.46	2.06	3.16	1.30	1.42
Minnesota	2.03	4.34	0.99	1.40	1.50	2.18	0.77	1.21
Mississippi	1.25	1.62	0.81	1.11	0.95	1.07	0.87	0.97
New Jersey	2.42	4.21	1.35	1.54	2.00	3.29	1.02	1.40
New York	2.32	3.94	1.46	1.51	2.35	3.84	0.95	1.52
Texas	3.56	6.59	0.96	1.84	3.56	7.39	1.51	1.84
Balance of the U.S.	1.89	2.70	0.96	1.36	1.63	3.12	0.88	1.25
Washington	1.87	3.42	0.97	1.34	1.79	3.02	0.94	1.32
National	3.46	6.36	1.93	1.84	3.56	6.02	1.50	1.86

**Table B-6.**  
**Average DEFF and DEFT for Estimates from the 1999 NSAF Adult Pair File for All Black Adults and Low-income Black Adults, by Site**

Study Area	All Black Adults				Low-income Black Adults			
	Average	Maximum	Minimum	DEFT	Average	Maximum	Minimum	DEFT
Alabama	2.03	3.07	1.09	1.41	2.13	3.68	1.15	1.44
Balance of Wisconsin	1.42	2.55	0.99	1.17	1.52	2.21	0.61	1.21
California	2.12	3.85	1.14	1.43	2.28	5.74	1.22	1.48
Colorado	1.77	3.78	0.73	1.28	1.52	3.39	0.98	1.21
Florida	2.52	3.83	1.05	1.57	2.32	3.33	1.37	1.51
Massachusetts	3.28	11.88	0.85	1.73	2.48	5.04	0.86	1.54
Michigan	2.20	4.83	0.98	1.46	2.23	3.61	0.94	1.47
Milwaukee	1.93	2.98	0.93	1.38	1.66	2.37	1.03	1.27
Minnesota	2.86	6.93	1.26	1.63	2.16	5.56	0.92	1.42
Mississippi	2.02	3.32	1.03	1.41	1.97	3.72	1.22	1.39
New Jersey	2.27	3.81	1.21	1.49	2.23	5.54	1.36	1.47
New York	2.34	4.49	1.18	1.51	2.29	3.70	0.90	1.50
Texas	2.51	3.99	1.44	1.56	2.18	3.45	0.84	1.46
Balance of the U.S.	2.25	4.40	0.72	1.48	1.92	2.91	0.81	1.37
Washington	2.23	3.93	1.03	1.46	2.32	3.88	1.04	1.49
National	5.81	10.02	2.36	2.37	4.87	6.69	2.54	2.19

**Table B-7.**  
**Average DEFF and DEFT for Estimates from the 1999 NSAF Adult Pair File for All Adults in Households with Children and Low-income Adults in Households with Children, by Site**

Study Area	All Adults, Households with Children				Low-income Adults, Households with Children			
	Average	Maximum	Minimum	DEFT	Average	Maximum	Minimum	DEFT
Alabama	1.95	2.87	0.93	1.38	1.93	3.29	1.10	1.38
Balance of Wisconsin	2.05	3.24	0.40	1.40	2.23	4.06	0.77	1.47
California	1.95	3.02	1.09	1.38	1.82	3.37	0.75	1.33
Colorado	2.28	3.67	0.91	1.49	2.19	3.50	0.93	1.46
Florida	2.53	4.90	1.00	1.56	2.43	4.71	1.13	1.53
Massachusetts	2.20	5.09	0.78	1.45	2.55	6.23	0.95	1.54
Michigan	2.25	4.70	0.53	1.47	2.08	3.89	1.00	1.42
Milwaukee	2.25	6.08	0.75	1.46	1.93	3.63	0.78	1.37
Minnesota	1.95	3.84	0.56	1.37	1.87	3.76	0.93	1.35
Mississippi	1.98	4.84	0.82	1.37	2.01	5.09	0.64	1.38
New Jersey	2.39	4.22	0.86	1.52	1.98	3.27	0.85	1.39
New York	2.58	4.35	0.98	1.58	2.47	5.17	0.66	1.53
Texas	2.85	5.81	1.12	1.64	2.43	4.54	1.27	1.54
Balance of the U.S.	1.82	2.91	0.50	1.33	1.83	3.24	0.82	1.33
Washington	2.03	4.17	0.81	1.39	2.04	3.91	0.72	1.39
National	4.62	7.80	1.16	2.11	4.04	7.51	1.96	1.98

**Table B-8.**  
**Average DEFF and DEFT for Estimates from the 1999 NSAF Adult Pair File for All Adults in**  
**Households with No Children and Low-income Adults in Households with No Children, by Site**

Study Area	All Adults, Households with No Children				Low-income Adults, Households with No Children			
	Average	Maximum	Minimum	DEFT	Average	Maximum	Minimum	DEFT
Alabama	1.52	2.49	0.66	1.22	1.99	2.67	1.07	1.40
Balance of Wisconsin	1.26	1.80	0.47	1.11	1.38	1.89	0.65	1.16
California	1.51	3.36	0.91	1.22	1.84	3.58	1.17	1.34
Colorado	1.60	2.76	0.67	1.25	1.59	2.83	0.70	1.24
Florida	1.50	2.19	0.55	1.21	1.58	2.61	0.85	1.24
Massachusetts	1.36	1.91	0.72	1.16	1.55	2.47	0.99	1.23
Michigan	1.64	2.97	0.81	1.27	1.93	3.15	0.72	1.37
Milwaukee	1.53	2.16	0.65	1.23	1.61	2.80	0.86	1.26
Minnesota	1.41	2.18	0.69	1.18	1.55	2.36	0.91	1.23
Mississippi	1.55	3.10	0.72	1.23	1.80	3.12	0.96	1.33
New Jersey	1.49	2.29	0.82	1.21	1.52	3.00	0.66	1.22
New York	1.25	1.77	0.80	1.11	1.62	2.45	0.91	1.26
Texas	1.46	2.59	0.70	1.19	1.63	3.93	0.60	1.25
Balance of the U.S.	1.54	3.15	0.68	1.22	1.99	2.76	1.15	1.40
Washington	1.38	1.82	1.00	1.17	1.41	2.15	0.67	1.18
National	3.50	5.52	1.81	1.85	3.76	5.83	1.79	1.93

## **Appendix C**

### **Round 1-Round 2 Correlations**

Table C-1 gives summary measures of correlations computed for 26 statistics nationally and by study area using the child data from the 1997 and 1999 NSAF. The first column is the mean correlation and the other columns are order statistics for the correlations. The median correlation will be used in our discussion as the average correlation.

Table C-2 gives the estimated correlations for a subset of the records, those children who were low-income in Round 1 or Round 2.<sup>15</sup> The subsetting has the effect of reducing the correlation because a common set of children is not included. This method was used because it most closely corresponds to the types of estimates analysts are likely to produce. For example, this is the correlation appropriate for estimates of the change in the percentage of low-income children with health insurance across the two years. The average national correlation is 0.09, nearly equal to the average for all children. By study area, the average correlations for low-income child estimates are lower than the corresponding estimates for all children.

Tables C-3 and C-4 give the correlations for estimates of change from the ADULT\_PR file, the first for all adults and the second for low-income adults. For all adults, the average national correlation is 0.06 and the study area correlations are somewhat lower. The estimated correlations for low-income adult statistics are even lower for the national estimate for all adults and about the same at the study area level.

For the ADULT\_RN file, correlations were computed for national estimates of change and for a few study areas. These estimated correlations were approximately the same for the nation and the study areas examined as the estimated correlations from the ADULT\_PR file.

The correlations summarized in this section show that estimates of change are slightly more precise because of the overlapping design, but the increase in precision is not as great as had been anticipated at the design stage. Some of the reasons for this are discussed in detail in 1999 NSAF Methodology Report No. 4. Nevertheless, some gains were achieved from the overlap. Treating the estimates from the two rounds as if they were independent typically results in slightly larger standard errors for estimates of change than would be realized when the correlation is taken into account.

Appendix tables C-5 and C-6 show the correlations estimated for the subset of individuals sampled in both rounds of the survey. The correlations for this subset are typically larger than the estimated correlations for all children and adults shown in tables C-1 through C-4. For example, the national average correlation for the subset of all children in table C-5 is 0.12, while

---

<sup>15</sup> A record for a child was included if the child was low-income for the particular round. Thus, if a child was low-income in Round 1 but not in Round 2, only the Round 1 record for the child was included in the analysis.

it is 0.08 for all children. Thus, it appears that the newly added sample in Round 2 has fairly a minor role in reducing the correlation.

The overlapping household correlations in tables C-5 and C-6 still do not measure the population correlation because of nonoverlap due to nonresponse and sampling within the household. To eliminate these effects, correlations were also computed for the subset of persons who were interviewed in both rounds of data collection. These matching persons were identified by staff at the Urban Institute using matching techniques because no unique identifier was available to do this at the person level. While there is undoubtedly some error involved in the matching, this person-level overlap is the best method for studying population correlations directly. See Report No. 10 in the 1999 Methodology Series for more details on the matching.

Tables C-7 and C-8 are for children and adults who were the subject of interviews in both the 1997 and 1999 NSAF. The correlations at the person level are about two to three times greater than the household-level correlations for all persons shown earlier. The average person-level correlations for children ranges from .23 to .42 across the study areas. For adult estimates of change, the average estimated correlations go from .13 to .29 across study areas. While these correlations are much larger than the other correlations presented, they still are much lower than the expected correlation of 0.60 used in designing the sample. While some of the difference might be due to errors in matching, it appears that the expected correlations, formulated based on other surveys, were overestimates for these characteristics from the NSAF.

The average correlations presented here smooth out some of the noise that is inherent in estimating correlations. Correlations are second-order statistics, like variances, and are subject to relatively large sampling errors. While the averaging should reduce the error in the estimates, it is useful to remember that these are not particularly stable estimates.

The lower-than-expected correlations in tables C-7 and C-8 have implications for the analysis of estimates of change. The overlap was designed to improve the precision for estimates of change, but the reductions in the sampling errors are not as large as expected. For example, if the correlations had been 0.60 (as expected), child estimates of change might have had sampling errors that were 90 percent of the size expected from independent samples.

Two other observations follow from the lower-than-expected correlations. First, analysts who use the simpler procedure for estimating sampling errors for estimates of change given by equation 1.1 for independent samples will not be overestimating by a large amount. Second, the finding suggests that overlapping the sample, even complete overlap, will not likely produce much more precise estimates of change for the types of statistics considered here.

**Table C-1.**  
**Correlations from 1997 and 1999 NSAF Child File for All Children, by Study Area**

<b>Study Area</b>	<b>Mean</b>	<b>Min</b>	<b>10%</b>	<b>25%</b>	<b>Median</b>	<b>75%</b>	<b>90%</b>	<b>Max</b>
Alabama	0.11	-0.21	-0.09	0.02	0.12	0.19	0.26	0.52
Balance of Wisconsin	0.04	-0.33	-0.17	-0.11	0.01	0.18	0.26	0.47
California	0.08	-0.08	-0.03	0.01	0.07	0.14	0.22	0.34
Colorado	0.06	-0.17	-0.09	-0.03	0.05	0.15	0.22	0.32
Florida	0.06	-0.15	-0.13	-0.10	0.04	0.20	0.28	0.47
Massachusetts	0.10	-0.19	-0.08	0.01	0.11	0.21	0.28	0.30
Michigan	0.08	-0.36	-0.20	-0.06	0.07	0.19	0.29	0.77
Milwaukee	0.06	-0.27	-0.16	-0.04	0.07	0.17	0.24	0.40
Minnesota	0.19	-0.22	-0.06	0.01	0.15	0.32	0.57	0.75
Mississippi	0.13	-0.25	-0.13	-0.01	0.11	0.27	0.42	0.53
New Jersey	0.04	-0.33	-0.15	-0.06	0.05	0.15	0.20	0.28
New York	0.06	-0.27	-0.12	0.00	0.10	0.15	0.18	0.30
Texas	0.13	-0.21	-0.08	-0.03	0.12	0.27	0.34	0.45
Balance of the U.S.	0.08	-0.12	-0.06	0.03	0.10	0.14	0.21	0.24
Washington	0.10	-0.26	-0.06	0.02	0.09	0.21	0.26	0.38
National	0.08	-0.19	-0.09	0.05	0.10	0.16	0.21	0.29

**Table C-2.**  
**Correlations from 1997 and 1999 NSAF Child File for Low-income Children, by Study Area**

<b>Study Area</b>	<b>Mean</b>	<b>Min</b>	<b>10%</b>	<b>25%</b>	<b>Median</b>	<b>75%</b>	<b>90%</b>	<b>Max</b>
Alabama	0.07	-0.29	-0.10	-0.03	0.09	0.11	0.26	0.52
Balance of Wisconsin	0.04	-0.21	-0.16	-0.06	0.02	0.12	0.24	0.40
California	0.05	-0.14	-0.04	-0.02	0.03	0.08	0.19	0.24
Colorado	0.04	-0.28	-0.18	-0.09	0.01	0.17	0.22	0.33
Florida	0.06	-0.20	-0.15	-0.09	0.02	0.18	0.28	0.44
Massachusetts	0.03	-0.44	-0.17	-0.05	0.07	0.12	0.18	0.31
Michigan	0.06	-0.35	-0.22	-0.14	0.03	0.17	0.36	0.77
Milwaukee	0.03	-0.30	-0.18	-0.07	0.01	0.16	0.21	0.41
Minnesota	0.13	-0.28	-0.21	-0.10	0.09	0.36	0.47	0.74
Mississippi	0.10	-0.28	-0.16	-0.06	0.08	0.26	0.38	0.53
New Jersey	0.07	-0.15	-0.04	-0.01	0.05	0.11	0.22	0.28
New York	0.04	-0.17	-0.09	-0.01	0.04	0.09	0.14	0.32
Texas	0.11	-0.31	-0.05	0.05	0.07	0.23	0.32	0.42
Balance of the U.S.	0.10	-0.06	-0.01	0.04	0.11	0.15	0.21	0.28
Washington	0.03	-0.30	-0.18	-0.07	0.01	0.16	0.21	0.41
National	0.07	-0.12	-0.07	0.02	0.09	0.13	0.18	0.30



**Table C-3.**  
**Correlations from 1997 and 1999 NSAF Adult Pair File for All Adults, by Study Area**

<b>Study Area</b>	<b>Mean</b>	<b>Min</b>	<b>10%</b>	<b>25%</b>	<b>Median</b>	<b>75%</b>	<b>90%</b>	<b>Max</b>
Alabama	0.06	-0.24	-0.12	-0.07	0.03	0.20	0.33	0.37
Balance of Wisconsin	0.02	-0.24	-0.15	-0.07	0.02	0.10	0.17	0.31
California	0.07	-0.19	-0.08	0.00	0.10	0.15	0.20	0.27
Colorado	0.05	-0.24	-0.14	-0.03	0.06	0.14	0.20	0.36
Florida	-0.06	-0.25	-0.17	-0.12	-0.03	0.02	0.04	0.06
Massachusetts	0.00	-0.30	-0.15	-0.10	-0.02	0.09	0.17	0.30
Michigan	0.02	-0.20	-0.14	-0.07	0.03	0.08	0.13	0.30
Milwaukee	0.05	-0.20	-0.15	-0.06	0.04	0.19	0.22	0.33
Minnesota	0.04	-0.35	-0.10	-0.07	0.02	0.11	0.20	0.62
Mississippi	0.11	-0.21	-0.13	-0.01	0.09	0.19	0.36	0.48
New Jersey	0.04	-0.26	-0.10	0.00	0.05	0.12	0.21	0.26
New York	-0.05	-0.41	-0.22	-0.14	-0.05	0.07	0.10	0.29
Texas	0.04	-0.24	-0.18	-0.10	0.02	0.19	0.27	0.33
Balance of the U.S.	0.02	-0.17	-0.08	-0.05	0.01	0.11	0.15	0.27
Washington	0.05	-0.16	-0.09	-0.03	0.04	0.11	0.16	0.29
National	0.04	-0.24	-0.10	-0.04	0.06	0.11	0.18	0.40

**Table C-4.**  
**Correlations from 1997 and 1999 NSAF Adult Pair File for Low-income Adults, by Study Area**

<b>Study Area</b>	<b>Mean</b>	<b>Min</b>	<b>10%</b>	<b>25%</b>	<b>Median</b>	<b>75%</b>	<b>90%</b>	<b>Max</b>
Alabama	0.08	-0.19	-0.11	-0.03	0.09	0.16	0.31	0.40
Balance of Wisconsin	0.02	-0.24	-0.16	-0.02	0.04	0.08	0.18	0.24
California	0.08	-0.14	-0.03	0.00	0.09	0.16	0.20	0.37
Colorado	0.09	-0.13	-0.03	0.05	0.08	0.13	0.22	0.31
Florida	0.02	-0.13	-0.09	-0.08	0.01	0.08	0.15	0.19
Massachusetts	0.03	-0.14	-0.12	-0.04	0.02	0.11	0.20	0.26
Michigan	0.01	-0.26	-0.16	-0.05	0.02	0.10	0.17	0.19
Milwaukee	0.04	-0.29	-0.20	-0.09	0.00	0.18	0.30	0.41
Minnesota	0.04	-0.34	-0.18	-0.11	0.00	0.17	0.23	0.57
Mississippi	0.16	-0.13	-0.08	0.07	0.15	0.24	0.40	0.50
New Jersey	0.02	-0.15	-0.10	-0.08	0.03	0.08	0.12	0.16
New York	-0.02	-0.28	-0.21	-0.07	0.01	0.06	0.09	0.26
Texas	0.06	-0.28	-0.11	0.03	0.05	0.16	0.29	0.34
Balance of the U.S.	0.00	-0.20	-0.16	-0.09	0.02	0.08	0.16	0.28
Washington	0.02	-0.17	-0.12	-0.05	0.02	0.07	0.08	0.33
National	0.03	-0.09	-0.07	-0.03	0.02	0.05	0.15	0.23

**Table C-5.**  
**Correlations Estimated from Households Sampled in Both Rounds for All Children, by**  
**Study Area**

<b>Study Area</b>	<b>Mean</b>	<b>Min</b>	<b>10%</b>	<b>25%</b>	<b>Median</b>	<b>75%</b>	<b>90%</b>	<b>Max</b>
Alabama	0.11	-0.17	-0.03	0.03	0.13	0.17	0.21	0.36
Balance of Wisconsin	0.11	-0.27	-0.03	0.03	0.10	0.19	0.30	0.39
California	0.06	-0.15	-0.04	-0.02	0.05	0.13	0.18	0.35
Colorado	0.09	-0.11	-0.03	0.00	0.10	0.16	0.21	0.38
Florida	0.02	-0.29	-0.17	-0.04	0.04	0.11	0.15	0.25
Massachusetts	0.14	-0.26	-0.03	0.04	0.16	0.26	0.30	0.38
Michigan	0.07	-0.38	-0.13	-0.03	0.06	0.18	0.28	0.40
Milwaukee	0.05	-0.18	-0.12	-0.07	0.04	0.13	0.24	0.43
Minnesota	0.20	-0.06	0.00	0.11	0.20	0.34	0.37	0.45
Mississippi	0.10	-0.19	-0.09	-0.03	0.09	0.16	0.28	0.74
New Jersey	0.01	-0.30	-0.08	-0.05	-0.01	0.05	0.16	0.39
New York	0.04	-0.18	-0.16	-0.10	0.02	0.15	0.26	0.30
Texas	0.07	-0.33	-0.08	-0.01	0.06	0.15	0.25	0.43
Balance of the U.S.	0.10	-0.12	-0.03	0.04	0.08	0.17	0.24	0.39
Washington	0.11	-0.10	-0.04	0.02	0.09	0.20	0.25	0.35
National	0.11	-0.09	0.00	0.02	0.12	0.17	0.24	0.32

**Table C-6.**  
**Correlations Estimated from Households Sampled in Both Rounds for All Adults**  
**(Adult Pair File), by Study Area**

<b>Study Area</b>	<b>Mean</b>	<b>Min</b>	<b>10%</b>	<b>25%</b>	<b>Median</b>	<b>75%</b>	<b>90%</b>	<b>Max</b>
Alabama	0.03	-0.32	-0.13	-0.07	0.06	0.15	0.19	0.24
Balance of Wisconsin	0.04	-0.25	-0.07	-0.01	0.03	0.13	0.19	0.24
California	0.07	-0.20	-0.14	-0.04	0.08	0.15	0.29	0.38
Colorado	0.08	-0.28	-0.07	0.02	0.08	0.13	0.29	0.33
Florida	-0.03	-0.23	-0.20	-0.11	-0.01	0.04	0.13	0.14
Massachusetts	0.00	-0.31	-0.14	-0.09	-0.01	0.06	0.22	0.23
Michigan	0.00	-0.28	-0.13	-0.09	0.00	0.05	0.14	0.28
Milwaukee	0.07	-0.18	-0.05	-0.03	0.06	0.14	0.20	0.46
Minnesota	0.00	-0.19	-0.14	-0.06	-0.01	0.04	0.11	0.25
Mississippi	0.02	-0.28	-0.18	-0.05	0.03	0.12	0.18	0.30
New Jersey	0.05	-0.23	-0.08	-0.02	0.05	0.12	0.17	0.27
New York	-0.03	-0.27	-0.17	-0.14	-0.04	0.02	0.11	0.38
Texas	-0.01	-0.27	-0.19	-0.09	0.00	0.06	0.14	0.33
Balance of the U.S.	0.06	-0.13	-0.08	-0.03	0.04	0.15	0.23	0.30
Washington	0.06	-0.20	-0.09	-0.06	0.04	0.17	0.23	0.32
National	0.08	-0.12	-0.08	-0.02	0.07	0.17	0.27	0.34

**Table C-7.**  
**Correlations Estimated from Children Sampled in Both Rounds for All Children, by**  
**Study Area**

<b>Study Area</b>	<b>Mean</b>	<b>Min</b>	<b>10%</b>	<b>25%</b>	<b>Median</b>	<b>75%</b>	<b>90%</b>	<b>Max</b>
Alabama	0.38	0.02	0.06	0.19	0.42	0.54	0.60	0.86
Balance of Wisconsin	0.36	-0.19	0.01	0.29	0.34	0.55	0.62	0.70
California	0.39	0.11	0.17	0.30	0.36	0.48	0.59	0.77
Colorado	0.34	-0.08	0.06	0.20	0.36	0.47	0.62	0.67
Florida	0.27	-0.21	-0.11	0.11	0.35	0.42	0.56	0.64
Massachusetts	0.39	0.10	0.20	0.32	0.37	0.53	0.58	0.67
Michigan	0.27	-0.20	-0.03	0.12	0.28	0.45	0.53	0.68
Milwaukee	0.21	-0.34	-0.01	0.16	0.24	0.32	0.43	0.45
Minnesota	0.34	-0.19	0.13	0.21	0.33	0.46	0.64	0.77
Mississippi	0.30	-0.15	0.03	0.14	0.30	0.45	0.58	0.65
New Jersey	0.25	-0.01	0.09	0.16	0.23	0.33	0.39	0.68
New York	0.29	-0.18	0.01	0.13	0.34	0.42	0.54	0.60
Texas	0.29	-0.07	0.03	0.17	0.26	0.40	0.61	0.76
Balance of the U.S.	0.27	-0.07	0.10	0.16	0.25	0.39	0.49	0.59
Washington	0.33	-0.02	0.08	0.15	0.35	0.50	0.57	0.66
National	0.28	-0.01	0.05	0.19	0.26	0.41	0.49	0.55

**Table C-8.**  
**Correlations Estimated from Adults Sampled in Both Rounds for All Adults (Adult Pair**  
**File), by Study Area**

<b>Study Area</b>	<b>Mean</b>	<b>Min</b>	<b>10%</b>	<b>25%</b>	<b>Median</b>	<b>75%</b>	<b>90%</b>	<b>Max</b>
Alabama	0.17	-0.03	-0.02	0.06	0.15	0.29	0.34	0.39
Balance of Wisconsin	0.22	-0.25	-0.11	0.04	0.25	0.39	0.61	0.63
California	0.16	-0.11	-0.04	0.04	0.13	0.29	0.35	0.57
Colorado	0.21	-0.21	-0.01	0.06	0.20	0.36	0.42	0.68
Florida	0.15	-0.62	-0.05	0.01	0.18	0.29	0.45	0.53
Massachusetts	0.21	-0.15	-0.04	0.05	0.19	0.35	0.49	0.68
Michigan	0.23	-0.12	-0.06	0.08	0.27	0.40	0.46	0.59
Milwaukee	0.22	-0.07	0.00	0.12	0.20	0.29	0.44	0.62
Minnesota	0.26	-0.11	0.00	0.12	0.29	0.40	0.48	0.65
Mississippi	0.19	-0.33	-0.03	0.03	0.15	0.39	0.45	0.63
New Jersey	0.20	-0.27	0.00	0.09	0.22	0.27	0.42	0.88
New York	0.19	-0.16	0.02	0.05	0.17	0.33	0.46	0.52
Texas	0.21	-0.11	-0.06	0.08	0.21	0.37	0.48	0.51
Balance of the U.S.	0.22	-0.09	-0.02	0.05	0.23	0.34	0.45	0.63
Washington	0.24	-0.02	0.02	0.08	0.27	0.38	0.45	0.48
National	0.21	-0.06	0.02	0.09	0.21	0.32	0.43	0.63

## Appendix D. SAS Macros for Use with NSAF Data

### Section 1.1 Overview

The SAS macros are programs written by Urban Institute (UI) staff to enable researchers to generate accurate variance estimates from the National Survey of America's Families (NSAF) without using additional software. By default, lower versions of SAS and other standard statistical packages cannot make ready use of certain complex survey designs for statistical analysis. More specifically, SAS does not have the built-in capability to properly analyze NSAF and its use of replicate weights.

The first section of this appendix describes the macros, their capabilities, and the syntax required to invoke the macros correctly. In the second section, we present examples of sample programs to run the macros, and follow these examples with the output resulting from the submitted statements. The third section contains the actual macro programs. The macro programs in section 3 can be used as is, or can be modified to run other statistical tests (e.g., logistic regression). However, please note that while these macros are being made available to external researchers as a convenience, we cannot provide support beyond general technical assistance.

### Section 1.2 Description of Macros

The **JRRFREQ** macro can be used to generate a one-way to two-way tables showing a distribution of variable values on observation counts, sum of weights, and percents. It also generates corrected standard errors and t-statistics.

The **JRRTTEST** macro can be used to produce a T-test for the difference between two sample means of a categorical variable. It also allows selecting two categorical groups of a class variable for comparison.

The **JRROLS** macro can be used to run an Ordinary Least Squares regression model.

### Section 1.3 Syntax

To use any of the macros, use a *%include* statement to reference the macro from within the SAS environment. After invoking the macro, the data step should specify the input data set (containing your analysis variable and replicate weights), variable list (including dependent and independent variables as necessary), replicate weight variable<sup>i</sup>, and output data set (containing the procedure results). Some of the macros have optional formatting and subsetting parameters. However, note that the SAS macro programs do not require values for these macro parameters: `order=`, `valfmt=`, or `subset=`. These optional parameters can be left blank or not specified at all.

**%jrrfreq**      (**indata**= input SAS dataset name,  
                  **classvar**= class variable names,  
                  **weight**= weight variable name,  
                  **outdata**= output SAS dataset name,  
                  **valfmt**= "FORMAT" statement for class variables,  
                  **subset**= "Where" statement for selecting observations )

**%jrrttest**     (**indata**= input SAS dataset name,  
                  **classvar**= class variable name,  
                  **weight**= weight variable name,  
                  **outdata**= output SAS dataset name,  
                  **subset**= "Where" statement to select two categorical groups of the class  
                              variable for comparison)

**%jrrols**        (**indata**= input SAS dataset name,  
                  **depvar**= dependent variable,  
                  **indvar**= independent variable list,  
                  **weight**= weight variable name,  
                  **outdata**= output SAS dataset name)

## Section 2. Sample Macro Code and Output

### JRRFREQ Code

```
libname m 'M:\Nsaf';

filename jrrfreq 'M:\Nsaf\jrrfreq.sas';
%include jrrfreq;

proc format;
    value valinc .5='LT50%'
                1='GE 50% LT 100%'
                1.5='GE 100% LT 150%'
                2='GE 150% LT 200%'
                3='GE 200% LT 300%'
                4='GE 300%';
    value vallfm 1='UBPIA LE 12' 0='UBPIA GT 12';
run;

data child;
    set m.focalchd;
run;

%jrrfreq(indata=child,
         classvar=uincrpov ubpianeg,
         weight=wgfcad,
         outdata=out,
         valfmt=format uincrpov valinc. ubpianeg vallfm.,
         subset= if age>=6 AND age<=11);
run;
```

### JRRFREQ Output

```
CORRECTED STANDARD ERRORS                      13:41 Wednesday, June 19, 2002    1
DISTRIBUTION USING REPLICATE WEIGHTS:  wgfcad0-wgfcad60

TABLE : uincrpov * ubpianeg  if age>=6 AND age<=11
      0 observations were omitted due to missing data
      11614 observations used
```

CPS family income as %	Negative behavior	Estimate	Number of	Standard
---------------------------	----------------------	----------	-----------	----------

of poverty Prob> T	6-11 years	Type	Observations	Estimate	Error	t-Stat
0.00	.	SUMWGT	11,389	24,251,232.75	182,435.48	132.93
0.00	UBPIA GT 12	SUMWGT	10,636	22,719,194.72	199,414.09	113.93
0.00	UBPIA LE 12	SUMWGT	753	1,532,038.04	130,103.86	11.78
LT50%	.	SUMWGT	723	1,844,248.45	101,419.72	18.18
0.00	UBPIA GT 12	SUMWGT	654	1,628,441.65	84,443.39	19.28
LT50%	UBPIA LE 12	SUMWGT	69	215,806.80	64,829.69	3.33
0.00	GE 50% LT 100%	SUMWGT	1,072	2,706,222.54	132,547.41	20.42
0.00	UBPIA GT 12	SUMWGT	957	2,373,245.73	122,403.09	19.39
0.00	UBPIA LE 12	SUMWGT	115	332,976.80	68,222.11	4.88
GE 100% LT 150%	.	SUMWGT	1,348	2,825,147.84	108,066.06	26.14
0.00	UBPIA GT 12	SUMWGT	1,232	2,608,922.59	102,321.90	25.50
0.00	UBPIA LE 12	SUMWGT	116	216,225.25	37,515.60	5.76
GE 150% LT 200%	.	SUMWGT	1,380	2,836,687.83	135,044.27	21.01
0.00	UBPIA GT 12	SUMWGT	1,276	2,651,654.18	139,401.67	19.02
0.00	UBPIA LE 12	SUMWGT	104	185,033.65	33,974.17	5.45
GE 200% LT 300%	.	SUMWGT	2,207	4,677,104.93	209,972.90	22.27
0.00	UBPIA GT 12	SUMWGT	2,074	4,464,871.55	209,942.19	21.27
0.00	UBPIA LE 12	SUMWGT	133	212,233.38	34,374.31	6.17
GE 300%	.	SUMWGT	4,659	9,361,821.16	212,056.91	44.15
0.00	UBPIA GT 12	SUMWGT	4,443	8,992,059.01	206,231.46	43.60

GE 300%	UBPIA LE 12	SUMWGT	216	369,762.15	49,286.16	7.50	
0.00	.	PERCENT	11,389	100.00	0.00	.	
	UBPIA GT 12	PERCENT	10,636	93.68	0.53	177.33	
0.00	.	UBPIA LE 12	PERCENT	753	6.32	0.53	11.96
0.00		PERCENT	723	7.60	0.42	18.13	
LT50%	UBPIA GT 12	PERCENT	654	6.71	0.35	19.08	
0.00	UBPIA LE 12	PERCENT	69	0.89	0.27	3.33	
0.00		PERCENT	1,072	11.16	0.53	21.01	
GE 50% LT 100%	UBPIA GT 12	PERCENT	957	9.79	0.50	19.74	
0.00							

\*\*\*\*\* Class rows with blanks or . are TOTAL rows \*\*\*\*\*

CORRECTED STANDARD ERRORS

13:41 Wednesday, June 19, 2002

2

DISTRIBUTION USING REPLICATE WEIGHTS: wgfcad0-wgfcad60

TABLE : uincrpov \* ubpianeg if age>=6 AND age<=11  
0 observations were omitted due to missing data  
11614 observations used

CPS family income as % of poverty Prob> T	Negative behavior 6-11 years	Estimate Type	Number of Observations	Estimate	Standard Error	t-Stat
GE 50% LT 100%	UBPIA LE 12	PERCENT	115	1.37	0.28	4.90
0.00	.	PERCENT	1,348	11.65	0.46	25.05
0.00	UBPIA GT 12	PERCENT	1,232	10.76	0.44	24.33
0.00	UBPIA LE 12	PERCENT	116	0.89	0.15	5.77
0.00	.	PERCENT	1,380	11.70	0.54	21.56
0.00	UBPIA GT 12	PERCENT	1,276	10.93	0.56	19.39
0.00						



GE 150% LT 200%	UBPIA LE 12	PERCENT	104	0.76	0.14	5.47	
0.00							
GE 200% LT 300%	.	PERCENT	2,207	19.29	0.83	23.11	
0.00							
GE 200% LT 300%	UBPIA GT 12	PERCENT	2,074	18.41	0.84	22.00	
0.00							
GE 200% LT 300%	UBPIA LE 12	PERCENT	133	0.88	0.14	6.19	
0.00							
GE 300%	.	PERCENT	4,659	38.60	0.84	46.19	
0.00							
GE 300%	UBPIA GT 12	PERCENT	4,443	37.08	0.82	45.46	
0.00							
GE 300%	UBPIA LE 12	PERCENT	216	1.52	0.20	7.52	
0.00							
.	.	COL %	11,389	100.00	0.00	.	.
.	UBPIA GT 12	COL %	10,636	100.00	0.00	.	.
.	UBPIA LE 12	COL %	753	100.00	0.00	.	.
LT50%	.	COL %	723	7.60	0.42	18.13	
0.00							
LT50%	UBPIA GT 12	COL %	654	7.17	0.37	19.49	
0.00							
LT50%	UBPIA LE 12	COL %	69	14.09	3.86	3.65	
0.00							
GE 50% LT 100%	.	COL %	1,072	11.16	0.53	21.01	
0.00							
GE 50% LT 100%	UBPIA GT 12	COL %	957	10.45	0.51	20.43	
0.00							
GE 50% LT 100%	UBPIA LE 12	COL %	115	21.73	3.94	5.52	
0.00							
GE 100% LT 150%	.	COL %	1,348	11.65	0.46	25.05	
0.00							
GE 100% LT 150%	UBPIA GT 12	COL %	1,232	11.48	0.47	24.22	
0.00							
GE 100% LT 150%	UBPIA LE 12	COL %	116	14.11	2.37	5.96	
0.00							
GE 150% LT 200%	.	COL %	1,380	11.70	0.54	21.56	
0.00							
GE 150% LT 200%	UBPIA GT 12	COL %	1,276	11.67	0.60	19.31	
0.00							
GE 150% LT 200%	UBPIA LE 12	COL %	104	12.08	2.10	5.76	
0.00							
GE 200% LT 300%	.	COL %	2,207	19.29	0.83	23.11	
0.00							

\*\*\*\*\* Class rows with blanks or . are TOTAL rows \*\*\*\*\*

CORRECTED STANDARD ERRORS

13:41 Wednesday, June 19, 2002

3

DISTRIBUTION USING REPLICATE WEIGHTS: wgfcad0-wgfcad60

TABLE : uincrpov \* ubpianeg if age>=6 AND age<=11  
0 observations were omitted due to missing data  
11614 observations used

CPS family income as % of poverty Prob> T	Negative behavior 6-11 years	Estimate Type	Number of Observations	Estimate	Standard Error	t-Stat	
GE 200% LT 300% 0.00	UBPIA GT 12	COL %	2,074	19.65	0.89	22.01	
GE 200% LT 300% 0.00	UBPIA LE 12	COL %	133	13.85	1.95	7.10	
GE 300% 0.00	.	COL %	4,659	38.60	0.84	46.19	0.00
GE 300% 0.00	UBPIA GT 12	COL %	4,443	39.58	0.85	46.71	
GE 300% 0.00	UBPIA LE 12	COL %	216	24.14	2.99	8.07	
.	.	ROW %	11,389	100.00	0.00	.	.
.	UBPIA GT 12	ROW %	10,636	93.68	0.53	177.33	
0.00	.	UBPIA LE 12	753	6.32	0.53	11.96	
0.00	.	ROW %	723	68.15	469.03	0.15	
LT50% 0.88	UBPIA GT 12	ROW %	654	60.17	414.10	0.15	
LT50% 0.88	UBPIA LE 12	ROW %	69	7.97	54.93	0.15	
0.89	.	ROW %	1,072	100.00	0.00	.	.
GE 50% LT 100% 0.00	UBPIA GT 12	ROW %	957	87.70	2.37	37.06	
GE 50% LT 100% 0.00	UBPIA LE 12	ROW %	115	12.30	2.37	5.20	
GE 100% LT 150% 0.88	.	ROW %	1,348	99.59	681.16	0.15	
GE 100% LT 150% 0.88	UBPIA GT 12	ROW %	1,232	91.97	629.03	0.15	

GE 100% LT 150%	UBPIA LE 12	ROW %	116	7.62	52.13	0.15	
0.88							
GE 150% LT 200%	.	ROW %	1,380	100.00	0.00	.	.
GE 150% LT 200%	UBPIA GT 12	ROW %	1,276	93.48	1.24	75.51	
0.00							
GE 150% LT 200%	UBPIA LE 12	ROW %	104	6.52	1.24	5.27	
0.00							
GE 200% LT 300%	.	ROW %	2,207	100.00	0.00	.	.
GE 200% LT 300%	UBPIA GT 12	ROW %	2,074	95.46	0.75	128.10	
0.00							
GE 200% LT 300%	UBPIA LE 12	ROW %	133	4.54	0.75	6.09	
0.00							
GE 300%	.	ROW %	4,659	100.00	0.00	.	.
GE 300%	UBPIA GT 12	ROW %	4,443	96.05	0.51	187.26	
0.00							
GE 300%	UBPIA LE 12	ROW %	216	3.95	0.51	7.70	
0.00							

\*\*\*\*\* Class rows with blanks or . are TOTAL rows \*\*\*\*\*

## JRROLS Code

```
libname m 'M:\Nsaf';

filename jrrols 'M:\Nsaf\jrrols.sas';
%include jrrols;

data kids6_11;
  set m.focalchd;
  if 6<=age<=11;
run;

%jrrols(indata=kids6_11,
        depvar=ubpia,
        indvar=age,
        weight=wgfcad,
        outdata=out);

run;
```

## JRROLS Output

UNCORRECTED STANDARD ERRORS

08:22 Wednesday, June 12, 2002 8

Simple OLS procedure using normalized base weight: NORMWGT0 ONLY

The REG Procedure

Model: MODEL1

Dependent Variable: UBPIA Age 6-11 Behavioral Problems Index score

Weight: normwgt0

### Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	265.12248	265.12248	64.74	<.0001
Error	11387	46632	4.09519		
Corrected Total	11388	46897			

Root MSE	2.02366	R-Square	0.0057
Dependent Mean	16.07053	Adj R-Sq	0.0056
Coeff Var	12.59236		



# Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr >  t
Intercept	Intercept	1	16.83513	0.09690	173.74	<.0001
AGE	Age	1	-0.09003	0.01119	-8.05	<.0001
CORRECTED STANDARD ERRORS					08:22 Wednesday, June 12, 2002	

9

Simple OLS using replicate weights: wgfcad0 - wgfcad60

11614 total observations

0 observations were omitted due to missing data

Approximately 11614 observations used in regressions (not adjusting for 0 weights)

DEPENDENT VARIABLE: ubpia

Variable	Label	Parameter Estimate	Estimate Variance	Standard Error	t-Stat	p	Signif.
AGE	Age	-0.090025	.	.	.	.	*** p<.01
age		.	.000601957	0.024535	.	.	*** p<.01
intercep		.	.	.	.	.	*** p<.01

### JRRTEST Code

```
libname m 'M:\Nsaf';

filename jrrttest 'M:\Nsaf\jrrttest.sas';
%include jrrttest;

data child;
    set m.focalchd;
run;

%jrrttest(indata=child,
          classvar=uactpos,
          varlist=uagg,
          weight=wgfcad,
          subset=if 6 <= age <= 11,
          outdata=out);

run;
```

### JRRTEST Output

The SAS System

08:51 Friday, June 21, 2002 4

CORRECTED STANDARD ERRORS

TTEST using replicate weights wgfcad0-wgfcad60

----- Variable=UAGG -----  
----

Child is  
invlvd in at  
least one  
activity

	Label	N	mean	Variance	Standard Error	t-Stat
0	Parent aggravation scale score	1874	13.463	0.012	0.108	124.599
1	Parent aggravation scale score	9584	13.884	0.001	0.030	460.289

5

CORRECTED STANDARD ERRORS

TTEST using replicate weights wgfcad0-wgfcad60

Label	Variable	estimate	Standard Error	T Statistic	Prob> T
-------	----------	----------	-------------------	----------------	---------

diff	UAGG	0.420	0.115	3.670	0.001
------	------	-------	-------	-------	-------



### Section 3. Macro Code

#### JRRFREQ.SAS

```
*****;
* Program: jrrFREQ.sas *;
* This program produces one-way to two-way tables. Tables show the *;
* distribution of variable values include: number of observations, *;
* sum of weights and percents. Two-way tables also have column percents*;
* and row percents. Statistics such as: standard errors, t values and *;
* Prob>|t| are calculated for sum of weights, percents, column percents*;
* and row percents. *;
*****;

=====;
%macro jrrfreq(indata=, classvar=, weight=, outdata=, subset=, valfmt=);

*Datasets used;
*-----;
*_base : SUMWGT using base weights;
*_tmp(i): SUMWGT using replicate weight i;
*_matrix: Holds the matrix of SUMWGT from the replicate weights;

*Define local variables;
*-----;
%local n; *Number of replicate weights;
%local k; *Number of class variables passed;
%local i; *Simple counter variable;
%local tmp; *Temp variable;
%local ncol; *Number of columns;
%local nrow; *Number of rows;
%local var1; *The first variable on the classvar parameter
%local var2; *The second variable on the classvar parameter

*Number of replicate weights. Should be 60;
%let n=60;

*Remove trailing and leading spaces from list of class variables;
*Then count spaces+1 to determine the number of variables passed;
*Also create new string variable contains class variable with ""
*for title of output
*-----;

%do %while (%substr(&classvar,1,1)=" ");
  %let &classvar=%substr(&classvar,2,%length(&classvar)-1);
%end;
%do %while (%substr(&classvar,%length(&classvar),1)=" ");
  %let &classvar=%substr(&classvar,1,%length(&classvar)-1);
%end;

%let k=1;
%let crssth=; * for crosstab title;
%let delim=;
%let tmp=%qscan(&classvar,&k,%str( ));
%do %while (&tmp ne);
  %let crssth=&crssth&delim &tmp;
  %let k=%eval(&k+1);
  %let tmp=%qscan(&classvar,&k,%str( ));
%end;
```

```

%let k=%eval(&k-1);
%let len=length(&crssth);
%let crssth=substr(&crssth,3,%eval(&len-2));

*-- Check the "SUBSET" parameter for subsetting --*;
*-- the sas file provided on the "INDATA" parameter --*;

%let chkwhere=%eval(%length(%left(&subset)));

%if &chkwhere GT 0 %then %do;
  data _temp;
    set &INDATA;
    &subset;
    %let indata=_temp;
%end;

*-- Identify class variables in the classvar parameter--*;

%let var1=%qscan(&classvar,1,%str( ));
%let var2=XXX;
%if &k > 1 %then %do;
  %let var2=%qscan(&classvar,2,%str( ));
%end;

*-- Count missing class variables --*;

data _null_;
  set &indata(keep=&classvar ) end=last;
  length count cntmiss miss 3;

  count+1;
  miss=0;

  x=put(&var1,3.);

  if x=' ' or trim(left(x)) = '.' then miss=1;

  if &k > 1 then do;
    x=put(&var2,3.);
    if x=' ' or trim(left(x)) = '.' then miss=1;
  end;
  cntmiss+miss;

  if last then do;
    call symput('nmiss',cntmiss);
    call symput('ncount',count);
  end;

run;

%let nused=%trim(%eval(&ncount-&nmiss));
%let nmiss=%trim(&nmiss);
%let ncount=%trim(&ncount);

*****;
* Compute weighted counts *;
*****;

```

```

*-- Get sum weight using base weights(0) --*;

proc summary data=&indata(keep=&classvar &weight.0) ;
  class &classvar;
  var &weight.0;
  output out=_sum(rename=(_freq_=N)) sum=b_sumwgt ;
  &valfmt;

*-- Count # columns and # rows and save into macro variables: nrow ncol--*;
*proc print;
* title '_sum wgt0';
data _null_;
  set _sum end=last;
  retain c r 0;
  if &k=2 then do;
    if _type_=1 then c+1;
    else if _type_=2 then r+1;
  end;

if last then do;
  if &k ne 2 then do; * no row% & col% for 1 way crosstab - assign a dummy value;
    c=1;
    r=1;
    if &k > 2 then do;
      *-- Print ERROR message when users enter more than 2 ways crosstab --*;
      file print;
      put " ***** ERROR MESSAGE *****";
      put " ***** YOU ENTERED &K CLASS VARIABLES *****";
      put " ***** THIS PROCEDURE ALLOWS UP TO 2 CLASS VARIABLES *****";
      put " ***** SPECIFY "WHERE" STATEMENT USING ONE OF YOUR CLASS *****";
      put " ***** VARIABLES ON THE "SUBSET= " PARAMETER AND *****";
      put " ***** RERUN YOUR PROGRAM *****";
    end;
  end;
  call symput ('ncol',trim(left(c)));
  call symput ('nrow',trim(left(r)));
  title1 "TABLE : &crsstab &subset";
  title2 " ";
end;
run;

*-- Select total counts and save in a file for computing percentages later--*;

%if &k=1 or &k=2 %then %do;
data _basetot(keep=sumall colsum1-colsum&ncol rowsum1-rowsum&nrow
               colval1-colval&ncol rowval1-rowval&nrow );
  set _sum end=last;
  retain col row sumall colsum1-colsum&ncol rowsum1-rowsum&nrow 0
         colval1-colval&ncol rowval1-rowval&nrow ' ';
  array colsum (*) colsum1 - colsum&ncol;
  array colval (*) colval1 - colval&ncol;
  array rowsum (*) rowsum1 - rowsum&nrow;
  array rowval (*) rowval1 - rowval&nrow;

if _type_ = 0 then sumall = b_sumwgt;

if &k=2 then do;
  if _type_=1 then do ;
    col+1;

```

```

        colsum(col) = b_sumwgt;
        colval(col) = trim(left(put(&var2,3.)));
    end;
    else if _type_=2 then do;
        row+1;
        rowsum(row) = b_sumwgt;
        rowval(row) =trim(left(put(&var1,3.)));
    end;
end;
if last then output;
*proc print;
*   title '_basetot';
proc sort data=_sum; by &classvar _type_;
*proc print;
*   title '_sum sorted by classvar type';

proc transpose data=_sum(drop=N _type_) out=_new(rename=(col1=sumwgt));
  by &classvar;
*proc print;
*   title '_sum after transpose';

*-- Compute percent, column percent and row percent --*;

data _base(drop=XXX _name_ r c sumall colsum1-colsum&ncol rowsum1-rowsum&nrow);
  set _new;
  LENGTH XXX 3;
  if _n_=1 then set _basetot;

  array colsum (*) colsum1-colsum&ncol;
  array colval (*) colval1-colval&ncol;
  array rowsum (*) rowsum1-rowsum&nrow;
  array rowval (*) rowval1-rowval&nrow;

  percent=sumwgt/sumall*100;

  if &k=2 then do;
    do c=1 to &ncol;
      if trim(left(put(&var2,3.))) = colval(c) then colpct=sumwgt/colsum(c)*100;
    end;
    if colpct=. then colpct=sumwgt/sumall*100;
    do r=1 to &nrow;
      if trim(left(put(&var1,3.))) = rowval(r) then rowpct=sumwgt/rowsum(r)*100;
    end;
    if rowpct=. then rowpct=sumwgt/sumall*100;
  end;
*proc print;
*   title '_base';

*-- Get sum weight and compute percent, column percent and row percent --*;
*-- for each of the replicate weights          --*;

data _matrix;
  set _null_;

%do i=1 %to &n;
proc means data=&indata(keep=&classvar &weight.&i) noprint ;
  class &classvar;
  var &weight&i;
  output out=_tmp&i(drop=_freq_)   sum=i_sumwgt;
  &valfmt;

```

```

data _tot(keep=sumall colsum1-colsum&ncol rowsum1-rowsum&nrow
          colval1-colval&ncol rowval1-rowval&nrow);
set _tmp&i end=last;
retain col row sumall colsum1-colsum&ncol rowsum1-rowsum&nrow 0
          colval1-colval&ncol rowval1-rowval&nrow ' ' ;
array colsum (*) colsum1 - colsum&ncol;
array colval (*) colval1 - colval&ncol;
array rowsum (*) rowsum1 - rowsum&nrow;
array rowval (*) rowval1 - rowval&nrow;

if _type_ = 0 then sumall = i_sumwgt;

if &k=2 then do;
  if _type_=1 then do ;
    col+1;
    colsum(col) = i_sumwgt;
    colval(col) = trim(left(put(&var2,3.)));
  end;
  else if _type_=2 then do;
    row+1;
    rowsum(row) = i_sumwgt;
    rowval(row) = trim(left(put(&var1,3.)));
  end;
end;
if last then output;

proc sort data=_tmp&i; by &classvar _type_;
proc transpose data=_tmp&i(drop= _type_) out=_new(rename=(col1=i_sumwgt));
  by &classvar;

data _tmp&i(drop=XXX _name_ r c sumall colsum1-colsum&ncol rowsum1-rowsum&nrow);
set _new;
LENGTH XXX 3;
if _n_=1 then set _tot;

i_pct=i_sumwgt/sumall*100;

array colsum (*) colsum1-colsum&ncol;
array colval (*) colval1-colval&ncol;
array rowsum (*) rowsum1-rowsum&nrow;
array rowval (*) rowval1-rowval&nrow;

if &k=2 then do;
  do c=1 to &ncol;
    if &var2 = colval(c) then i_colpct=i_sumwgt/colsum(c)*100;
  end;
  if i_colpct=. then i_colpct=i_sumwgt/sumall*100;
  do r=1 to &nrow;
    if &var1 = rowval(r) then i_rowpct=i_sumwgt/rowsum(r)*100;
  end;
  if i_rowpct=. then i_rowpct=i_sumwgt/sumall*100;
end;

data _matrix;
set _matrix _tmp&i;
%end;

proc sort data=_matrix; by &classvar ;

*Calculate deviation matrix;
*-----;

```

```

proc sort data=_sum; by &classvar;

data _deviat ;
  merge _base _matrix;
  by &classvar ;

  variance=(sumwgt - i_sumwgt)**2;
  pctvari =(percent - i_pct)**2;
  colvari =(colpct - i_colpct)**2;
  rowvari =(rowpct - i_rowpct)**2;

*Calculate sum of deviations;
*-----;
proc summary data=_deviat missing nway;
  class &classvar ;
  var i_sumwgt variance pctvari colvari rowvari;
  output out=_deviat(drop= _type_ _freq_ ) sum=;

*Create output dataset;
*-----;

data _temp;
  merge _sum(keep=&classvar N) _base _deviat;
  by &classvar ;
  dof=60;

  stderr=variance**(1/2);
  t = sumwgt /stderr;
  p=(1-cdf('T',abs(t),dof))*2;

  std_pct=pctvari**(1/2);
  t_pct =percent/std_pct;
  p_pct=(1-cdf('T',abs(t_pct),dof))*2;

  std_col=colvari**(1/2);
  t_col =colpct/std_col;
  p_col=(1-cdf('T',abs(t_col),dof))*2;

  std_row=rowvari**(1/2);
  t_row =rowpct/std_row;
  p_row=(1-cdf('T',abs(t_row),dof))*2;

data &outdata(keep=&classvar type value N count std t_stat prob);
  set _temp;

  array _cnt (i) sumwgt percent colpct rowpct;
  array _std (i) stderr std_pct std_col std_row ;
  array _t (i) t t_pct t_col t_row;
  array _p (i) p p_pct p_col p_row;
  do i=1 to 4;
    count = _cnt;
    std = _std;
    t_stat = _t;
    prob = _p;
    if i=1 then do; value='SUMWGT '; type='1'; end;
    else if i=2 then do; value='PERCENT'; type='2'; end;
    else if i=3 then do; value='COL % '; type='3'; end;
    else if i=4 then do; value='ROW % '; type='4'; end;
    output;
  end;

```

```

format N comma10. count std comma15.2 t_stat prob 10.2;
proc sort data=&outdata; by type &classvar ;

*-- Print output --*;

%if &k = 1 %then %do;
proc print data=&outdata noobs label split='*';
  where value in ('SUMWGT','PERCENT') ;
  id &classvar value;
  var N count std t_stat prob;
  &valfmt ;
  label count ="Estimate"
        std  ="Standard* Error"
        t_stat='t-Stat'
        prob  ='Prob>|T|'
        value ='Estimate* Type'
        N     ='Number of*Observation'
  ;
  title1 "CORRECTED STANDARD ERRORS";
  title2 "DISTRIBUTION USING REPLICATE WEIGHTS: &weight.0-&weight&n";
  title3 "          ";
  title4 "TABLE : &crssth &subset";
  title5 "&nmiss observations were omitted due to missing data";
  title6 "&nused observations used";
  footnote "***** Class rows with blanks or . are TOTAL rows *****";

%end;

%if &k > 1 %then %do;
proc print data=&outdata noobs label split='*';
  id &classvar value;
  var N count std t_stat prob;
  &valfmt ;
  label count ="Estimate"
        std  ="Standard* Error"
        t_stat='t-Stat'
        prob  ='Prob>|T|'
        value ='Estimate* Type'
        N     ='Number of*Observations'
  ;
  title1 "CORRECTED STANDARD ERRORS";
  title2 "DISTRIBUTION USING REPLICATE WEIGHTS: &weight.0-&weight&n";
  title3 "          ";
  title4 "TABLE : &crssth &subset";
  title5 "&nmiss observations were omitted due to missing data";
  title6 "&nused observations used";
  footnote "***** Class rows with blanks or . are TOTAL rows *****";

%end;

%end;

*Cleanup;
*-----;
proc datasets nolist;
  delete _temp _base _deviat _matrix _tmp0- _tmp60 _sum _basetot _tot;

%mend;
*=====;

```

## JRROLS.SAS

```
*=====;
%macro jrrols(indata, depvar, indvar, weight, outdata);

  %local n;      * Number of replicateweights;
  %local k;      * Number of independent variables;
  %local i;      * Simple counter variable;
  %local tmp;    * Temp variable;

  *-- Number of replicate weights. Should be 60 --*;
  %let n=60;

  *-- Remove trailing and leading spaces from list of independent variables --*;
  *-- Then count spaces+1 to determine the number of variables passed --*;;

  %do %while (%substr(&indvar,1,1)=" ");
    %let &indvar=%substr(&indvar,2,%length(&indvar)-1);
  %end;
  %do %while (%substr(&indvar,%length(&indvar),1)=" ");
    %let &indvar=%substr(&indvar,1,%length(&indvar)-1);
  %end;
  %let k=1;
  %let tmp=%qscan(&indvar,&k,%str( ));
  %do %while (&tmp ne);
    %let k=%eval(&k+1);
    %let tmp=%qscan(&indvar,&k,%str( ));
  %end;
  %let k1=%eval(&k);
  %let k=%eval(&k-1);

  *-- Get stats for missing variables and zero weights --*;

  data _null_;
    set &indata(keep=&depvar &indvar) end=last;
    length count cntmiss miss 3;

    count+1;
    miss=0;

    array lookmiss{&k1} &depvar &indvar;

    do i=1 to &k1;
      if lookmiss{i}=. then miss=1;
    end;

    cntmiss+miss;

    if last then do;
      call symput('nmiss',cntmiss);
      call symput('ncount',count);
      output;
    end;
  run;

  %let nused=%trim(%eval(&ncount-&nmiss));
  %let nmiss=%trim(&nmiss);
```



```

%let ncount=%trim(&ncount);

*-- Normalize the base weight --*;

proc summary data=&indata nway;
  var &weight.0;
  output out=_mwgt0(drop=_type_ _freq_) mean=mwgt0;
*proc print;
* title 'mean weight 0';
data _temp(keep=&depvar &indvar normwgt0);
  set &indata;
  if _n_ =1 then set _mwgt0;
  normwgt0=&weight.0/mwgt0;
*proc means data=_temp;
* var normwgt0;
* title 'weight 0 after normalized';

*-- Get betas using base weights(0) --*;

proc reg data=_temp(keep=&depvar &indvar normwgt0
                    where=(normwgt0 gt 0)) outest=_beta0;
  model &depvar = &indvar;
  weight normwgt0;
  title1 "UNCORRECTED STANDARD ERRORS";
  title2 "Simple OLS procedure using normalized base weight: NORMWGT0 ONLY";

data tmp0(keep=match intercep base1-base&k rename=(intercep=base0))
  _beta0(keep=intercep &indvar); *Store base weight betas for later;

set _beta0;

array indvar{&k} &indvar;
array base{&k} base1-base&k;

do i=1 to &k;
  base{i}=indvar{i};
end;
match=1;

proc transpose data=_beta0 out=_beta0;
data _beta0;
  set _beta0(rename=( _name_=variable _label_=label coll=betahat));

  if variable="INTERCEP" then variable=' INTERCP';

*-- Get betas for each of the replicate weights --*;

data tmpmatix;
  set _null_;

%do i=1 %to &n;
  proc reg data=&indata(keep=&depvar &indvar &weight.&i) outest=tmp&i noprint;
    model &depvar = &indvar;
    weight &weight.&i;
    where &weight.&i gt 0;
  data tmpmatix;
    set tmpmatix tmp&i;
  %end;

```

```

*-- Clean up matrix for merge --*;

data tmpmatix;
  set tmpmatix;
  match=1;
  keep match intercep &indvar;

data tmpdiff;
  merge tmp0 tmpmatix;
  by match;

  array _beta0{&k1} intercep &indvar;
  array tmpbase{&k1} base0-base&k;
  do i=1 to &k1;
    _beta0{i}=(_beta0{i}-tmpbase{i})*2;
  end;

proc summary data=tmpdiff nway;
  class match;
  var intercep &indvar;
  output out=tmpdiff(drop=_type_ _freq_ match) sum=intercep &indvar;

proc transpose data=tmpdiff out=tmpdiff;

data tmpdiff;
  set tmpdiff(drop= _label_ rename=( _name_=variable
                                   coll=variance));
  *stderr=coll**(.1/2);
  if variable="INTERCEP" then variable=' INTERCP';

proc sort data=_beta0; by variable;
proc sort data=tmpdiff; by variable;
data &outdata;
  merge _beta0 tmpdiff;
  by variable;
  stderr=variance**(.1/2);
  tstat=betahat/stderr;
  dof=60 - &k;
  p=(1-cdf('T',abs(tstat),dof))*2;
  select;
    when (p lt .01) signif="*** p<.01";
    when (p lt .05) signif="** p<.05";
    when (p lt .1 ) signif="*  p<.1";
    otherwise;
  end;

  if variable=" INTERCP" then variable='INTERCEP';

label variable="Original Variable Name"
  label="Original Variable Label"
  betahat="Beta Coefficient"
  variance="Variance of Coefficient"
  stderr="Standard Deviation of Coefficient"
  tstat="t-Statistic"
  signif="Significance";

*-- Clean up datasets used --*;

proc datasets nolist;

```

```

delete tmp0-tmp&n tmpmatix _beta0 tmpdiff;

proc print data=&outdata noobs uniform split="|";
  var variable label betahat variance stderr tstat p signif;
  label variable="Variable"
        label="Label"
        betahat="Parameter|Estimate"
        variance="Estimate|Variance"
        stderr="Standard|Error"
        tstat="t-Stat"
        signif="Signif.";
  title1 "CORRECTED STANDARD ERRORS";
  title2 "Simple OLS using replicate weights: &weight.0 - &weight.&n";
  title3 "&ncount total observations";
  title4 "&nmiss observations were omitted due to missing data";
  title5 "Approximately &nused observations used in regressions (not adjusting for 0 weights)";
  title6 "DEPENDENT VARIABLE: &depvar";
run;
%mend;
*=====;

```

## JRRTEST.SAS

```

*****
* program:jrrtest.sas
* This program computes sample means, standard errors for each of
* two groups of observations identified by levels of a CLASS variables
* and tests the hypothesis that the population means are the same.
*****

*=====;
%macro jrrtest(indata=, classvar=, varlist=, weight=, outdata=,subset=);

  *Datasets used;
  *-----;
  *_mean0: Means using base weights;
  *_tmp(i): Means using replicate weight i;
  *_matrix: Holds the matrix of means from the replicate weights;

  *Define local variables;
  *-----;
  %local n; *Number of replicate weights;
  %local k; *Number of variables passed;
  %local i; *Simple counter variable;
  %local tmp; *Temp variable;

  *Number of replicate weights. Should be 60;
  %let n=60;

  *Remove trailing and leading spaces from list of mean variables;
  *Then count spaces+1 to determine the number of variables passed;
  *-----;

  %do %while (%substr(&varlist,1,1)=" ");
    %let &varlist=%substr(&varlist,2,%length(&varlist)-1);
  %end;

```

```

%do %while (%substr(&varlist,%length(&varlist),1)=" ");
  %let &varlist=%substr(&varlist,1,%length(&varlist)-1);
%end;

%let k=1;
%let tmp=%qscan(&varlist,&k,%str( ));
%do %while (&tmp ne);
  %let k=%eval(&k+1);
  %let tmp=%qscan(&varlist,&k,%str( ));
%end;
%let k=%eval(&k-1);

*-- Check the "SUBSET" parameter for subsetting --*;
*-- the sas file provided on the "INDATA" parameter --*;

%let chkwhere=%eval(%length(%left(&subset)));

%if &chkwhere GT 0 %then %do;
  data _temp;
    set &INDATA;
    &subset;
    %let indata=_temp;
%end;

*Get means using base weights(0);
*-----;
proc summary data=&indata(keep=&classvar &varlist &weight.0) nway ;
  class &classvar;
  var &varlist;
  weight &weight.0;
  output out=_mean00(drop=_type_ rename=( _freq_=N)) mean=;
  *title1 'UNCORRECTED STANDARD ERRORS';
  *title2 "MEANS USING BASE WEIGHT: &weight.0 ";

*Create dataset of base weight means for later merging;
*-----;
proc transpose data=_mean00
  out=_base(rename=( _name_=variable col1=mean ));
  var &varlist;
  by &classvar;
proc sort; by variable &classvar;

*Transpose base weight means dataset;
*-----;

proc transpose data=_mean00
  out=_mean0(rename=( _name_=variable col1=mean_01 col2=mean_02 ));
  var &varlist;
  idl &classvar;
proc sort ; by variable;

*Get means for each of the replicate weights;
*-----;
data _matrix;
  set _null_;

```

```

%do i=1 %to &n;
proc summary data=&indata(keep=&classvar &varlist &weight.&i) nway;
  class &classvar;
  var &varlist;
  weight &weight&i;
  output out=_tmp&i(drop=_type_ _freq_) mean=;
proc transpose data=_tmp&i
  out=_tmp&i(rename=( _name_=variable col1=mean_i1 col2=mean_i2 ));
  var &varlist;
  idl &classvar;
data _matrix;
  set _matrix _tmp&i;

%end;
proc sort data=_matrix;    by variable;

*Calculate deviation matrix;
*-----;
data _deviat;
  merge _mean0 _matrix;
  by variable;

  variance=((mean_i2-mean_i1) - (mean_02 - mean_01))**2;
  varianc1=(mean_i1-mean_01)**2;
  varianc2=(mean_i2-mean_02)**2;

*Calculate sum of deviations;
*-----;
proc summary data=_deviat nway;
  class variable;
  var variance varianc1 varianc2 ;
  output out=_deviat(drop=_type_ _freq_ ) sum= ;

data _deviat;
  merge _mean0 _deviat;
  by variable;

*-- compute T statistic --*;
label='diff';
stderr=variance**(1/2);
stderr1=varianc1**(1/2);
stderr2=varianc2**(1/2);
estimate=mean_02 - mean_01;
t = estimate/stderr;
dof=60;          *set degree of freedom;

p=(1 - cdf('T',abs(t),dof))*2;

keep label t p varianc1 varianc2 stderr stderr1 stderr2 variable estimate;

*Create output dataset;
*-----;

data _temp;
  merge _base _deviat(drop=label estimate stderr t p);
  by variable;
  length label $ 40;
  label=_label_;

```

```

proc sort; by &classvar variable;
data _temp;
  merge _temp _mean00(keep=&classvar N);
  by &classvar;

data &outdata(keep=&classvar N variable label mean variance stderr t_val );
  set _temp;
  by &classvar variable;
  if first.&classvar then group+1;

  if group=1 then do;
    variance=varianc1;
    stderr =stderr1;
    t_val = mean/stderr;
    output;
  end;
  else do ;
    variance=varianc2;
    stderr =stderr2;
    t_val = mean/stderr;
    output;
  end;
proc sort; by variable &classvar;

*Cleanup;
*-----;
proc datasets nolist;
  delete _base _matrix _mean0 _tmp0-_tmp&k _temp;

proc print data=&outdata noobs uniform split="*";
  by variable;
  var &classvar label N mean variance stderr t_val ;
  format mean variance stderr t_val 12.3;
  label variable="Variable"
    label="Label"
    variance="Variance"
    stderr="Standard*Error"
    t_val='t-Stat';
  title2 "CORRECTED STANDARD ERRORS";
  title3 "TTEST using replicate weights &weight.0-&weight&n";

proc print data=_deviat noobs uniform split="*";
  var label variable estimate stderr t p ;
  format estimate stderr t p 12.3;
  label label='Label'
    variable='Variable'
    stderr="Standard*Error"
    t = "T*Statistic"
    p = "Prob>|T|" ;
%mend;
*=====;

```

## Appendix E. Computing JRR standard errors for means and regressions in STATA

Bowen Garrett  
The Urban Institute  
May 2000

To facilitate analyses of the NSAF for STATA users, I have written a series of STATA programs that produce survey design-adjusted standard errors using the jackknife repeated replicate (JRR) implemented for NSAF. This brief report describes the commands that are currently developed and their syntax, discusses their limitations, and illustrates their structure using OLS with JRR standard errors as an example. With this example, users of the NSAF with experience programming in STATA should be able to readily extend the method to other regression commands or customize the routines described here. These commands were written in STATA Version 5 for Windows 95. STATA Version 6 has an improved matrix language which should make the creation of user-written programs more straightforward.

The commands are implemented as STATA .ado files. The commands developed to date are:

- JRMEAN - Produces a table of means with standard errors
- JRTAB - Produces a table of means and standard errors by one categorical variable, with significant differences to a reference group indicated by asterisks
- JRREG - Linear regression
- JRLOGIT - Logit regression (optionally computes odds-ratios)
- JRMLOGIT - Multinomial logit (optionally computes relative risk ratios)

### Preliminaries

Before running the JRR commands, the user must specify which set of NSAF replicate weights to use. For example, if the random adult weights are being used, at the command line (or in a do file) type: *global jrwt wgrn* where *wgrn* is the prefix of the set of random adult weights. The user will receive a warning if the weights are not specified.

### Commands and descriptions

1. *jrmean* [*varlist*] if [*exp*] in [*range*]

Produces a table of means with standard errors. The N is the number of observations that contain no missing values for any of the variables in *varlist*.

2. *jrtab* [*catvar*] [*varlist*] if [*exp*] in [*range*]

Produces a table of means and standard errors of varlist by the categorical variable catvar. Catvar must be a whole (positive) number, starting with 0, where the 0 category is set to be the comparison category. The N is the number of observations that contain no missing values for any of the variables in varlist.

3. `jrreg [depvar] [varlist] if [exp] in [range]`

Estimates OLS regression. Supports post-estimation testing and prediction. The unadjusted standard errors are estimated with pweights.

4. `jrlogit [depvar] [varlist] if [exp] in [range], or tab`

Estimates logit regression. Supports post-estimation testing and prediction. The unadjusted standard errors are estimated with pweights.

5. `jrmlogit [depvar] [varlist] if [exp] in [range], or tab`

Estimates multinomial logit regression. Supports post-estimation testing and prediction. Uses aweights because the STATA mlogit command does not support pweights. The interested user could produce these in the usual way using svymlog in STATA. Unlike the usual mlogit command, this command forces 0 to be the comparison group.

### Options

The tab option creates additional output that can be parsed with MS Word (using Table: Convert text to columns) to create regression tables in a conventional format.

The or option produces odds ratios instead of logit coefficients.

The rrr option produces relative risk ratios instead of multinomial logit coefficients.

### Matsize

All of the JRR commands require that STATA's matsize be set to something greater than 61. This is done automatically. Remove the matsize command if you are running STATA version for Win 3.1, Dos, or Mac because STATA for Windows 95 or higher is needed to set matsize on the fly. Some users may also need to increase matsize for certain applications, beyond the default setting of 100. Simply edit the matsize command in jrwarn.ado accordingly. Unless one is doing analyses on more than 99 variables, this can be ignored.



### Example

Code for the JRREG program (line numbers are added for presentation only):

```
1. program define jrreg
2. version 5.0
3. set trace off
4. local varlist "req ex"
5. local if "opt"
6. local in "opt"
7. parse "`*' "
8. tempvar touse e
9. tempname bb
10.mark `touse' `if' `in'
11.jrwarn
12.if $jrtest==1 {exit}

13.local i=0
14.regress `varlist' [pweight=$jrwt`i'] if `touse'
15.matrix bb0=get(_b)

16.local i=1
17.while `i'<=60 {
18.quietly regress `varlist' [pweight=$jrwt`i'] if `touse'
19.matrix bb=get(_b)
20.if `i'==1 {matrix cc = bb}
21.else {matrix cc=cc\bb}
22.local i=`i'+1
23.}

24.*Make mm, a matrix where each row contains the coefficients from the
   regression using the 0 weights
25.mat ones=J(60,1,1)
26.matrix mm=ones*bb0
27.*Make ee, a matrix of deviations
28.matrix ee=cc-mm
29.*Make ff, the squared deviations = the covariance matrix of m0
30.matrix ff=ee'*ee

31.matrix post bb0 ff
32.matrix mlout

33.end
```

Lines 1-10 parse the command and define the analysis subsample.

Line 11 calls subprogram *jrwarn* that sets the matsize and verifies the user has specified weights. It displays the weights being used if they were specified and gives a reminder if the user did not. Line 12 exits if *jrwarn* identified a problem.

Line 14 performs the initial regression using the base weights (with suffix 0) for the full estimation sample. Line 15 puts the resulting coefficients in vector *bb0*. These are the final coefficients.

Lines 16-23 loop 60 times, reestimating the regression model using each set of replicate weights. These are accumulated in matrix *cc*.

Lines 28-31 create *ee*, a matrix of deviations of the replicate coefficients from the base weight coefficients. Then *ff*, which is the matrix product  $ee'ee$ . The matrix *ff* is the covariance matrix of coefficient vector *bb0*.

The remaining lines post these results to STATA's internal memory areas and display the regression results. Post-estimation commands can then proceed as usual.

#### Comparison of STATA JRR results to WESVAR results

We tested the proper functioning of these new STATA commands against output produced by WESVAR. The variables included an indicator of Medicaid/state insurance coverage, age, and poverty categories for a sample of children who recently left welfare. Comparisons using the JRTAB command (Exhibit 1) and the JRREG command (Exhibit 2) are reported below, with the result that the STATA commands produced estimates identical to those produced by WESVAR.

#### Conclusion

Producing JRR standard errors in STATA is a mechanical process and the commands presented here add to the set of tools available for analyzing the NSAF. The method here extends readily to other estimation commands. Users who have needs that are not addressed by these commands can use these commands as a starting point for writing their own. The command files, additional documentation, and the associated STATA help file will be made available over the web at <http://www.ui.urban.org>. Users of these commands should contact [nsaf@ui.urban.org](mailto:nsaf@ui.urban.org) if they have further questions or to report any remaining bugs in the programs provided. Users are also encouraged to submit any programs they write for possible inclusion among the programs we distribute.

Exhibit 1. Comparison of WesVar and STATA JRTAB means by category output

**WesVar Table Output:**

		MEDSTATE	
		0	1
age	Estimate	30.9	28.7
	Std Error	0.632	0.816
pov_1		25.0	30.0
		4.5	5.6
pov_2		26.6	39.8
		4.1	5.2
pov_3		20.6	13.7
		3.3	3.6
pov_4		13.6	8.2
		2.2	2.6
pov_5		7.8	5.6
		1.9	1.7
pov_6		6.5	2.7
		1.6	1.3

**STATA JRR Table Output:**

medstate	0	1
-----		
N	594	410
age	30.9	28.7***
	(0.632)	(0.816)
pov_1	.250	.300
	(0.045)	(0.056)
pov_2	.266	.398*
	(0.041)	(0.052)
pov_3	.206	.137
	(0.033)	(0.036)
pov_4	.136	.082
	(0.022)	(0.026)
pov_5	.078	.056
	(0.019)	(0.017)
pov_6	.065	.027*
	(0.016)	(0.013)

Note: Tests of significance are in comp

## Exhibit 2. Comparison of WesVar OLS regression and STATA JRREG output

### WesVar Regression Output:

PARAMETER	ESTIMATE	STANDARD ERROR OF ESTIMATE	TEST FOR H0: PARAMETER=0	PROB> T
INTERCEPT	0.701171	0.1293824	5.4193701	0
AGE	-0.0070143	0.0033038	-2.1231235	0.0379
POV	-0.0491332	0.0207497	-2.3678977	0.0211

R\_SQUARE VALUE = 0.0370333

### STATA JRR Regression Output:

```
. jrreg medstate age pov;
(sum of wgt is 1.3773e+006)
```

Regression with robust standard errors	Number of obs =	1004
	F( 2, 1001) =	4.85
	Prob > F =	0.0080
	R-squared =	0.0370
	Root MSE =	.47291

	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval]
medstate					
age	-.0070143	.003745	-1.873	0.061	-.0143632 .0003346
pov	-.0491332	.0196947	-2.495	0.013	-.0877809 -.0104856
_cons	.701171	.1403251	4.997	0.000	.425806 .9765359

Using weights with prefix: wgrn

	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
age	-.0070143	.0033038	-2.123	0.034	-.0134896 -.000539
pov	-.0491332	.0207497	-2.368	0.018	-.089802 -.0084645
_cons	.701171	.1293824	5.419	0.000	.4475862 .9547557

<sup>1</sup>The replicate weight variable assumes that you have 61 weight variables with identical names and a numeric suffix. For example, passing a weight variable with the name WGFCAD, will assume that you have weight variables with the names WGFCAD0, WGFCAD1, WGFCAD2.... WGFCAD60 in the input dataset.