HOW TO PREPARE AND ANALYZE PAIR DATA IN THE NATIONAL SURVEY ON DRUG USE AND HEALTH

Substance Abuse and Mental Health Services Administration
Center for Behavioral Health Statistics and Quality
Rockville, Maryland

May 2017

This page intentionally left blank

HOW TO PREPARE AND ANALYZE PAIR DATA IN THE NATIONAL SURVEY ON DRUG USE AND HEALTH

RTI Authors:

Jennifer Cooney
Peter Frechtel
Jeremy Porter
Lauren Klein Warren

SAMHSA Authors:

Jonaki Bose
Sarra Hedden
Satoko J. Kuramoto
Rachel Lipari
Matthew Williams

RTI Project Director:

David Hunter

SAMHSA Project Officer:

Peter Tice

For questions about this manual, please e-mail Peter.Tice@samhsa.hhs.gov.

Acknowledgments

Where to Go for Help

For Data Portal Users:
SAMHDA Support
E-mail: CBHSQRequest@samhsa.hhs.gov
Phone: 888.741.7242 or 734.615.9524
Fax: 734.647.8200

Table of Contents

Table of Contents (continued)

List of Tables

List of Exhibits

This page intentionally left blank

1.   Introduction

Before attempting to use pair data from the National Survey on Drug Use and Health (NSDUH)[1] to answer a research question of interest, it is important to understand what pair data are and whether using pair data is appropriate for the reader's needs. Pair data are data that have been collected from two individuals surveyed from the same dwelling unit. These pairs can come in many forms, such as father-child, sibling-sibling, and spouse-spouse. A complete list of pair types can be seen in Table 5 in Section 3.1 and in Appendix A.

There is an abundance of information to be gained from collecting data of this nature; however, the same analysis questions can sometimes be answered without the use of pair data. Analysts should be aware of which variables are available from all respondents, and not just from the pair interviews, to avoid conducting analyses using pair data instead of data from all respondents. The following sections of the 2013 NSDUH analytic codebook (Center for Behavioral Health Statistics and Quality [CBHSQ], 2014) provide information about other members of the dwelling unit: "Parenting Experiences," "Youth Experiences," "Household Composition (Roster)," "Roster Pair Variables," and "Income." Some of these modules are associated only with pair data; others are available for all respondents. Chapter 4 of this manual describes in more detail the limitations of using pair data.

Once the reader has determined whether pair data should be used to conduct his or her analysis, the information in this manual can be used to aid the reader in conducting the analysis properly. This manual provides guidance on how to create a pair analysis file and on the appropriate weights and design variables needed to analyze pair data, and it provides example code in multiple software packages.

1.1   Audience

The target audience of this manual is those interested in determining whether pair data are needed for their analysis or those needing to use NSDUH pair data for their analysis. Information in this manual is intended for researchers, programmers, and analysts but not all information may pertain to each specific user. The manual is designed so that basic material is easily accessible, and more advanced topics are included for those who are interested.

1.2   NSDUH Sample Selection and Data Collection

An area probability sample design has been used to select the sample dwelling units needed for the NSDUH. Land areas, dwelling units, and people were randomly sampled based on U.S. Census Bureau estimates of population and housing unit counts to ensure the sample represents the entire U.S. population. For NSDUH, all 50 states and the District of Columbia have been divided into state sampling regions (SSRs), and the number of SSRs varies depending on the population of the state. In each SSR, eight segments (with boundaries usually defined by surface features such as streets, railroad tracks, and rivers) were selected: two segments for each

---

[1] The name of the survey was the National Household Survey on Drug Abuse (NHSDA) until 2002. To avoid confusion, however, the term "NSDUH" will be used throughout this manual.

of the four quarterly data collection periods. Specially trained field staff were assigned to make a list of the dwelling units in each of the selected segments. From these listings, specific housing units and group quarters units have been selected for each segment, and these selected units are called sample dwelling units (SDUs). During the data collection period, NSDUH interviewers visit each SDU to determine whether the SDU is eligible to participate in the NSDUH and to develop a roster of individuals who reside at the unit the majority of the calendar quarter. Interviewers enter these rosters into electronic devices that use a randomized selection process to choose zero, one, or two individuals within the screened SDU to participate in the NSDUH. The selection process does not take into account the relationships between members in the household; therefore, this is a random effect. Interviewers then conduct the NSDUH interview with the selected individuals. Households where two individuals are selected form part of the pair data.

1.3   Access to the Pair Data

Because of confidentiality restrictions, NSDUH pair indicators are not available on the public use files. Pair data are available via the confidential restricted-use files (RUFs) for years 2002 onward (or 2004 onward for data portal users). If, after reviewing the material presented in this manual, the user determines he or she is interested in using the NSDUH pair data for analysis, the first step for non-CBHSQ researchers is to gain access to the data by applying for data portal access. Instructions are available at http://datafiles.samhsa.gov/. Once in the data portal, the user can access RUFs with more variables and more records than are included on the NSDUH public use files.[2] The user can also access a detailed codebook describing the variables. The codebook descriptions of all variables on the RUFs that are mentioned in this manual are presented in Appendix B.

1.4   Definitions

The following terms are used throughout this manual.

- **NSDUH:** The National Survey on Drug Use and Health (NSDUH) is sponsored by the Substance Abuse and Mental Health Services Administration and provides national, state, and substate data on substance use and mental health in the civilian, noninstitutionalized population aged 12 or older. NSDUH is a continuous survey, with approximately 67,500 interviews completed annually. NSDUH uses a multistage, state-level design where zero, one, or two individuals are selected from a dwelling unit. More information on the NSDUH can be found at https://www.samhsa.gov/data/.
- **Responding Pair, or Pair:** When two members of the same dwelling unit are selected, and both complete an interview, a "responding pair" is formed. These responding pairs permit detailed analyses involving members of the same dwelling unit. In the 2002-2013 NSDUHs, approximately 58 percent of the respondents were members of responding pairs, and in the 2014-2015 NSDUHs, approximately 53 percent were members of responding pairs. See Appendix A for the distribution of responding pairs for years 2002 to 2015.

---

[2] Pair analyses cannot be done using the NSDUH public use files. There is no way to determine which respondents are members of the same pair.

- **Confidential Restricted-Use Files (RUFs):** These annual person-level data files contain more records and variables than the public use files, and they contain all variables needed for pair analyses.
- **Pair-Level Variables:** These are variables where the two people sampled from the same household (pair members) are guaranteed to have the same value. For example, all geographic variables (e.g., state code) are pair-level variables because both pair members live in the same dwelling unit and therefore live in the same geographic area. All variables are considered either pair-level variables or person-level variables.[3] Pair-level variables are the same for pair members, whereas person-level variables are specific to each individual in the pair and can differ. Appendix F contains a list of all the pair-level variables on the RUFs.
- **Person-Level Variables:** These are variables that can differ between pair members, such as the gender of the respondent (*IRSEX*). Outcome data, such as substance use indicator variables, are also examples of person-level variables. Most NSDUH variables are person-level variables.
- **Pair-Level Dataset:** This type of dataset contains one observation per pair instead of one observation per respondent. This manual describes how to convert a person-level dataset to a pair-level dataset. In the pair-level dataset, each record will contain two person-specific variables for person-level variables (e.g., gender) and only one variable for pair-level variables (e.g., census region).
- **Pair Analysis:** This type of analysis uses responding pairs from the pair-level dataset to test a research hypothesis. For example, a researcher may be interested in whether fathers report using drugs given that a child reported that his or her parent or guardian talked with the child about the dangers of substance use.

1.5   Questionnaire Modules Relevant for Analysis of Pair Data

The computer-assisted interviewing questionnaire used in the NSDUH is divided into sections that are administered to specific modules of respondents. Data from these specific modules are of particular interest to analysts working with pair data because some sections pertain specifically to certain members of a pair (e.g., the parents in parent-child pairs). The questionnaire modules can be divided into five groups based on who they are administered to: (1) questions administered to youths aged 12 to 17, (2) questions administered to adults aged 18 or older, (3) employment questions administered to respondents aged 15 or older, (4) questions administered to parents of youths aged 12 to 17, and (5) questions administered to all respondents, regardless of their age or parental status.

- Sections that pertain only to youths aged 12 to 17:

    - Youth Experiences
    - Youth Mental Health Service Utilization
    - Adolescent Depression

---

[3] Some pair-level variables could be considered county-level variables or state-level variables, for example, but in this manual, all variables that occur at a level higher than the pair are considered pair-level variables.

- Sections that pertain only to adults aged 18 or older:

  - Social Environment
  - Mental Health
  - Adult Mental Health Service Utilization
  - Adult Depression

- Section that pertains only to people aged 15 or older:

  - Employment

- Section that pertains only to parents:

  - Parenting Experiences (questions administered to parents whose child was also chosen to complete an interview)

All other sections are administered to all respondents.

1.6   Examples of Pair Analyses

To help readers determine whether their analyses are suited to pair data, and if so, how best to analyze the data, three examples of analyses involving pair data are listed below. Each focuses on a different type of relationship that exists in the pair data. These three examples, along with the weight adjustments that are needed for each type of analysis, are described further in Chapter 2. Code that creates and analyzes the data for these examples is included in Appendices C through E. Note that it is not recommended to combine father-child and mother-child pairs to create a parent-child pair analysis. More details on this limitation can be found in Appendix G.

1. Father-Child Pairs: Past year drug use by the father among fathers whose child spoke to at least one parent or guardian about the dangers of drug, tobacco, or alcohol use (i.e., the father's behavior is the dependent variable and the child's behavior is the independent variable)
2. Mother-Child Pairs: Past year drug use by the child among children whose mother discussed the dangers or problems associated with drug, tobacco, or alcohol use with this child (i.e., the child's behavior is the dependent variable and the mother's behavior is the independent variable)
3. Spouse-Spouse Pairs: Past year drug use by both spouses (i.e., the behavior of both members of the pair are jointly the dependent variable)

1.7   Conventions Used in this Manual

In this manual, all variables on the RUFs are displayed in all capital letters and in italics (e.g., *AGE*). All Stata® variables used in the code (other than the official NSDUH variables on the RUFs) are displayed in all lowercase letters and in italics (e.g., *suffix*). All Stata code and commands in the text are displayed in Courier font (e.g., the `reshape` command). Analysts can use the code as a starting point and then customize it based on their specific analysis.

2. General Steps for Completing a Pair Analysis

This chapter describes simple research questions for which a user might use pair-level data. The steps to create and analyze the data are included along with example Stata code. Example code for other software packages is included in Appendices C through E.

2.1 Description of Example Research Questions and Mock Results

Section 1.6 introduced three types of analyses that can be performed using pair data. Each of these is shown below with an example of a research question that may be of interest to an analyst related to drug use and/or verbal communication of dangers of drug use of a parent to his or her child. Mock results and conclusions are also shown to further illustrate using the data.

1. Father-child pairs where the focus is on the *father*'s behavior (i.e., the father's behavior is the dependent variable and the child's behavior is the independent variable)

   - Example Research Question: What is the prevalence of paternal drug use given that the child reported talking to at least one parent or guardian about the dangers of drug, tobacco, or alcohol use?
   - Mock Result: "The prevalence of paternal drug use is higher among fathers whose child talked to a parent or guardian about drug use than fathers whose child did not talk to a parent or guardian about drug use." However, the child may be having these discussions because of his or her father's drug use and not the other way around.

2. Mother-child pairs where the focus is on the *child*'s behavior (i.e., the child's behavior is the dependent variable and the mother's behavior is the independent variable)

   - Example Research Question: What is the prevalence of child drug use (reported by the child) given that the mother discussed the dangers or problems associated with drug, tobacco, or alcohol use with this child?
   - Mock Result: "The prevalence of child drug use is lower among mothers who discussed the dangers of drug use with their child than among children of mothers who did not discuss the dangers of drug use." However, there may be other household members influencing the child's drug use.

3. Spouse-spouse pairs where the focus is on the *pair itself* (i.e., the behavior of both members of the pair are jointly the dependent variable)

   - Example Research Question: What is the joint prevalence of drug use among opposite-gender couples compared with same-gender couples?
   - Mock Result: "The percentage of same-gender couples in which at least one member used drugs is higher than for opposite-gender couples." However, it is possible that this prevalence is in fact low for both types of couples, but the prevalence for same-gender couples is statistically higher.

Note that only one parent is included in the analysis for examples 1 and 2. Because NSDUH does not collect data from both parents and the child, it is not recommended that analysts combine father-child and mother-child pairs to examine parent-child relationships. These limitations on analyzing parent-child pairs are described in Section 4.2 and Appendix G. Additionally, with only minor changes to the research questions for examples 1 and 2, person-level data can be used to analyze (1) parent drug use given that the same parent reported discussing the topic of dangers or problems associated with substance use with his or her child aged 12 to 17, and (2) the level of child drug use *as reported by the parent* given the parent had this discussion with his or her child. See Section 4.1 for more information about when using person-level data may be more appropriate than using pair-level data.

This manual primarily uses example 1 to explain the different concepts associated with using the pair data. For this example, the following information is needed for all father-child pairs in the sample:

- whether the father in the pair reported using illicit drugs in the past year, and
- whether the child aged 12 to 17 reported that his or her parent discussed with him or her the dangers or problems associated with substance use.

In planning the analysis, one must also determine which member of the pair is of interest (i.e., the focus). In this example, the interest is on the father's drug use behavior given the child reported that the parent did or did not communicate verbally with him or her about dangers with substance use behaviors. The focus of the analysis (i.e., which variable is considered the dependent variable) is important and will determine which weight is needed for the analysis. See Section 2.4.2 about weight adjustments needed for this example and the others introduced in Section 1.6, and see Section 3.2 for more detailed information about weights for pair data.

2.2   Creation of the Pair-Level Dataset for Analysis

This section explains how to convert a person-level dataset (i.e., where each of the pair members is a separate record) to a pair-level dataset (i.e., where each pair has a single associated record and only the cases that form part of a pair are included). This conversion can be performed using any standard statistical software package like SAS® or Stata. Some terms for this process are going from "long to wide," "skinny to fat," or "stacked to unstacked."

2.2.1 Obtain the Appropriate Data from the Data Source

The following variables are needed for most or all pair-level analyses. Further documentation for these variables can be found in Appendix B.

- Pair-Level Variables

  – *ENCCASE*: Household-level identification number (encrypted)
  – *IRPRREL*: Imputation-revised pair relationship indicator
  – *PRANALWT*: Final pair-level analysis weight
  – *PRVESTR*: Final pair-level analysis stratum
  – *PRVEREP*: Final pair-level analysis replicate

- Person-Level Variables

  - *QUESTID*: Respondent identification
  - *AGE*: Respondent age
  - *IRSEX*: Respondent gender
  - *PRNTIND*: Parent indicator for parent-child pairs
  - any substance use, mental health, or subgroup variables needed for the analysis of interest

For this example, because interest is in whether the child reported that the parent or guardian talked to the child about substance use dangers and whether the father reported using illicit drugs, the additional person-level variables *SUMYR* (past year illicit drug use) and *YEPRTDNG* (indicator for whether parent or guardian talked to child about danger of substance use) are included.

2.2.2 Subset Data to Only Members of Pairs

Once the appropriate data have been read in from the files that contain all respondents (both pair and nonpair), the next step is to discard observations from respondents who are not members of a pair (i.e., nonpairs). *IRPRREL* (pair relationship indicator) can be used for this purpose because it is set to 99 for nonpairs. **All types of pairs should remain on the dataset, including those not relevant to the research question.**[4] This is discussed further in Section 2.4.4.

2.2.3 Sort the Dataset

Regardless of the software package used, the dataset must first be sorted before converting to a pair-level dataset. The variable *ENCCASE* is the identification variable at the dwelling unit level, and *PRNTIND* is the parent indicator for parent-child pairs. In the case of parent-child pairs, the analyst should sort the dataset by (1) *year* (if multiple years of data will be used), then (2) *ENCCASE*, and then (3) descending *PRNTIND*. This sorting order forces the mother or father to appear first in a parent-child pair. For some analyses, it may be desirable to sort by variables other than *PRNTIND*. For example, if an analyst is interested in wives in husband-wife pairs, it would make more sense to sort by *ENCCASE*, followed by *IRSEX*. If an analyst is interested in sibling-sibling pairs, it might make more sense to sort by *ENCCASE*, followed by descending *AGE*, to put the older sibling first.

Once the sorting is complete, the first few observations of the person-level dataset might look like those in Table 1. Only a subset of variables on the file is shown.

---

[4] Resultant data will remain with the full design and/or a design with sampling domains; see the 2014 NSDUH Methodological Resource Book Sample Design Report (Center for Behavioral Health Statistics and Quality, 2015, Section 2.6). Also note that all survey-eligible pairs of individuals had some nonzero chance of being selected (Center for Behavioral Health Statistics and Quality, 2015, Section 1.5).

Table 1.   Person-Level Input Dataset

ENCCASE	QUESTID	PRNTIND	AGE	SUMYR	YEPRTDNG	IRSEX	REGION	IRPRREL
1	1	0	16	Yes	1	2	West	5
1	2	0	13	No	1	2	West	5
2	3	1	40	No	99	1	South	1
2	4	0	12	Yes	2	2	South	1

2.2.4 Convert Dataset to Pair-Level Dataset

The next step is to convert the two rows corresponding to each member of the pair into one row representing the pair. For person-level variables, separate new variables will store the values of the pair members. Exhibit 1 shows an example of this conversion. In this example, QUESTID is dropped because it is a person-level variable that is no longer relevant.

Exhibit 1.   Conversion of Person-Level Records to Pair-Level Records

ENCCASE	QUESTID	SUMYR
1	1	Yes
1	2	No
2	3	No
2	4	Yes

ENCCASE	SUMYR_X	SUMYR_Y
1	Yes	No
2	No	Yes

As shown in the Stata code provided in Section 2.3, the following steps are completed to create this pair-level dataset.[5] These steps are specific to Stata software programming logic, but Appendices C through E show the steps programmed using SAS, SPSS, and R. In the Stata example below, the "_X" and _"Y" are appended to the person-level variables to indicate the variables assigned to each pair member. Analysts working with parent-child or husband-wife pairs may prefer to use "_P" and "_C" or "_H" and "_W" so that the pair members are more easily identified.

1. Read in the person-level sorted dataset from Section 2.2.3.
2. The sorted dataset from Section 2.2.3 has two observations per household. The prior step sorted the data by the father indicator (*PRNTIND*) in descending order. So the father's record would be the first observation within the household and the child's record would be the second observation within the household. Create a variable called *suffix* that is assigned to the string of "_X" for all observations.
3. Use the `replace` command to assign a value of "_Y" to the *suffix* variable for every even-numbered observation.
4. Drop any nonpair-level variables prior to converting the file from a person-level to a pair-level file (e.g., *PRNTIND, AGE*) that are not the same within the pair. This eliminates confusion caused by having person-level variables on the file that have not been converted to pair-level variables.

---

[5] As is typical of programming problems, there are other ways to complete this conversion. For example, in SAS, one reasonable alternative method involves the use of the TRANSPOSE procedure.

5. Reshape the dataset from long to wide using the reshape command with the keyword `wide`. Next, list all the person-level variables of analytic interest by listing the variables with an @ symbol at the end to indicate where the variable *suffix* is to appear in the variable name. Use the `i()` option to indicate the variable whose values denote a subobservation. The variable *ENCCASE* is used as the subobservation to indicate the household. In addition, the `j()` and `string` options are used to append the variable *suffix* to where the @ is located at the person-level variables of analytic interest.

The code below can be easily modified to include other person-level variables that are needed for the analysis. Along with *SUMYR*, the variables *YEPRTDNG* (indicator for whether the parent or guardian talked to the child about danger of substance use) and *IRSEX* should also be converted into two variables for this example.[6] To convert other person-level variables for the pair dataset, add the variables to the variable list in the `reshape` command. As a computational convenience, converting does not need to be done for pair-level variables (e.g., *REGION*) that are guaranteed to have the same value within the pair, because only one copy of these variables will be output to the final dataset, and this value should correspond to both members of the pair.

After this step, the first two observations of the resulting dataset look like those in Table 2. Only a subset of variables on the file is shown. Note that the values of *ENCASE*, *REGION*, and *IRPRREL* are the same for both parent and child. These are pair-level variables and therefore do not require any conversion for the pair file. The variables shown in Table 1, but not shown in Table 2, are dropped in the code (Step 4 in the previous list).

Table 2.   Pair-Level Output Dataset

ENCCASE	REGION	sumyr_x	sumyr_y	yeprtdng_x	yeprtdng_y	irsex_x	irsex_y	IRPRREL
1	West	Yes	No	1	1	2	2	5
2	South	No	Yes	99	2	1	2	1

2.3   Stata® Code for Creating the Pair-Level Dataset

In order for a programmer/analyst to follow along with the text above, Stata code is provided below with comments referring the user to the steps above. SPSS, R, SAS survey, and SAS-callable SUDAAN® code completing the steps discussed in Chapter 2 are provided in Appendices C through E.[7]

```
/*Section 2.2.2: Reduce file to only members of pairs*/
* drop observations if irprrel = 99
drop if irprrel == 99
```

---

[6] For parent-child pairs, YEPRTDNG for the parent's record will be equal to a skip code, and YEPRTDNG_X will not be used in the example.

[7] Results will vary slightly across software packages. The coefficient estimates and their standard errors will agree, but the packages differ with respect to the default degrees of freedom used in hypothesis tests involving subpopulations. By default, Stata uses the degrees of freedom associated with the strata that include at least one member of the subpopulation; whereas R, SPSS, the SAS survey package, and SUDAAN use the degrees of freedom associated with the entire sample design.

```
* keep variables in dataset
keep pranalwt irprrel enccase prntind age prvestr prverep irmpcp17 yeprtdng sumyr ///
irsex

/*Section 2.2.3: Sort dataset*/
*Sort by ascending household-level identification number and descending parent
*indicator
gsort +enccase -prntind

/*Section 2.2.4: Convert dataset to pair-level*/
*Create suffixes to distinguish pair-level variables

* Create a variable called suffix and initialize to "_x"
* The purpose of "_x" is to indicate the parent in the household
generate str2 suffix = "_x"

* Replacing "_x" with "_y" for every even-numbered observation to indicate the child
replace suffix = "_y" if mod(_n,2) == 0

*Drop any non-pair level variables prior to reshaping the dataset or an error will
*occur during the reshaping process

drop prntind age

*Reshape from long to wide, using i(encase) to identify households and j(suffix) to
*append the "_x" or "_y" suffix to the person-level variables of analytic interest.
*The string option indicates that the suffix variable in j(suffix) is a string
*variable. List all person-level variables of analytic interest with the '@' mark

reshape wide sumyr@ yeprtdng@ irsex@, i(enccase) j(suffix) string
```

2.4 Steps for Running the Pair-Level Analysis

This section describes the steps needed to perform the statistical analysis on the pair-level dataset.

2.4.1 Sample Design Variables

Generally, analysis of NSDUH pair-level data should be done using the principles outlined for person-level data, as described in the 2014 NSDUH Methodological Resource Book Statistical Inference Report (Center for Behavioral Health Statistics and Quality, 2016a). The dataset should be analyzed as if it were the result of a single-stage stratified random sample drawn from a population whose size is infinite. In Stata, the command `svyset prverep [pweight=?], strata(prvestr)` describes the design. For the pair analyses, *PRVEREP* is the variance replicate (i.e., the primary sampling unit) and *PRVESTR* is the variance stratum, and the absence of the `fpc` option tells Stata to treat the population size as infinite. The weight variable is described in the next section. See the example code in Appendices C through E for the code that sets up the sample design for SPSS, R, the SAS survey package, and SUDAAN.

2.4.2 Weight Variable

For some analyses using pair data, no weight adjustments are needed and the analyst can simply use the pair analysis weight, *PRANALWT*, directly to weight their data in the `svyset` command described in Section 2.4.1. The creation of the *PRANALWT* variable is described in

detail in Westlake, Chen, & Gordek (2013). However, there are many cases when an adjustment needs to be applied to the weight variable to account for the presence of multiple pairs of a certain type within one household. Whenever the analysis is focused on one member of the pair, an adjustment, using one of the multiplicity variables listed in Table 5 in Section 3.1, is often needed to account for the fact that this member of the pair may be a member of more than one pair of this type (Chromy & Singh, 2001).

The examples introduced in Section 1.6 and expanded upon in Section 2.1 of this manual are shown below with information on the weights appropriate for the analyses.

1. Father-child pairs where the focus is on the father's behavior (i.e., the father's behavior is the dependent variable and the child's behavior is the independent variable)

   – Weight: For this analysis, the father's behavior is the focus. This means the interest is in the father's outcome given certain child characteristics. However, PRANALWT represents the population of pairs, so it needs to be adjusted by the number of children living with the father. For example, in a household where the father has three children younger than 18 residing in the household, he is part of three parent-child pairs. Table 5 shows that this adjustment is done using the variable *IRMPCP17*, which accounts for the number of children aged 12 to 17 living with the father. The weight for this analysis is *PRANALWT* divided by *IRMPCP17*.

2. Mother-child pairs where the focus is on the child's behavior (i.e., the child's behavior is the dependent variable and the mother's behavior is the independent variable)

   – Weight: For this analysis, the child's behavior is the focus. This means the interest is in the child's behavior given certain maternal behaviors. For the purpose of the pair-level analyses, the assumption is that for any given household, there is only one mother per child. Therefore, the specific child is a member of only one mother-child pair, and no weight adjustment is needed. The weight for this analysis is *PRANALWT*.

3. Spouse-spouse pairs where the focus is on both members of the pair (i.e., the behavior of both members of the pair are jointly the dependent variable)

   – Weight: For this analysis, the focus is on the pair itself so *PRANALWT* can be used directly. If the focus was instead on one member of the pair (e.g., the husband's behavior was the dependent variable), *PRANALWT* could still be used because the assumption is that each member of a spouse-spouse pair has only one spouse in the household.

Note that if the analyst is interested in an annual estimate of a population total, then the weight should be divided by the number of years that were pooled. This is described further in Section 3.2.

2.4.3 Dependent Variable

The dependent variable for the first example is *sumyr_x*, which indicates illicit drug use by the father in the father-child pair. *SUMYR* is the NSDUH variable for past year drug use, and the "_X" appended to the variable indicates that this is the variable corresponding to the father. In Stata, this variable is indicated in the `mean` command.

2.4.4 Independent Variable

The independent variable *yeprtdng_y* is coded as 1 for "Yes" and 2 for "No." All other values of this variable are various "missing" data codes and are not used in the analysis and can be set to missing prior to running the analysis. The variable *yeprtdng_m* is created by setting the values 85, 89, 94, 97, 98, and 99 to a missing value. This variable can be created either before or after converting the file to the pair-level file. Respondents with missing values for *yeprtdng_m* are excluded from the analysis automatically when including this variable on the subgroup statement.

2.4.5 Subpopulation Variable(s)

In this example, two proportions are compared: (1) the proportion of fathers using drugs in the past year among children who report their parent *did not* talk to him or her about substance use, and (2) the proportion of fathers using drugs in the past year among children who report their parent *did* talk to him or her about substance use. Many statistical software packages, including Stata, allow the user both to identify a subpopulation of interest and to create multiway tables within that subpopulation allowing easy comparisons between subgroups.

Following this reasoning, the subpopulation for the analysis described in Section 2.1 consists of all father-child pairs where the child is 12 to 17. In Stata, the statement that identifies the subpopulation reads `subpop (if (irprrel == 1 | irprrel == 2) & irsex_X == 1)`. The analyst may also subset the data to include only observations in the subpopulation using the command `if (irprrel == 1 | irprrel == 2) & irsex_X == 1`.

2.5 Stata® Code for Running the Pair-Level Analysis

```
*Section 2.4.2: divide the pair weight by the correct multiplicity
generate newwt = pranalwt/irmpcp17

*Section 2.4.1: svyset command
svyset prverep [pweight=newwt], strata(prvestr)

*Section 2.4.4: create the independent var by setting missing codes to a missing value
recode yeprtdng_y (1 = 1) (2 = 2) (else = .), generate yeprtdng_m

*Run analysis
svy, subpop (if (irprrel == 1 | irprrel == 2) & irsex_x == 1): mean sumyr_x,
over(yeprtdng_m)
```

2.6   Analysis Results

Table 3 is a template that the analyst populates by running the pair analysis.

Table 3.   Example of Pair Analysis

Child Aged 12-17 Reported that Parent/Guardian Talked to Child about Dangers of Substance Use	Father Used an Illicit Drug in the Past Year	
	Yes	**No**
Yes	n_{11}	n_{12}
No	n_{21}	n_{22}

The code included in the previous sections will calculate estimates, standard errors, and (unweighted) sample sizes of the following:

- $\dfrac{n_{11}}{n_{11} + n_{12}}$, the proportion of fathers who used illicit drugs in the past year given that the child aged 12 to 17 reported that the parent or guardian talked to the child about substance use at least once in the past year; and

- $\dfrac{n_{21}}{n_{21} + n_{22}}$, the proportion of fathers who used illicit drugs in the past year given that the child aged 12 to 17 reported that the parent or guardian did not talk to the child about substance use at least once in the past year.

The estimates and their standard errors are shown in Table 4.

Table 4.   Example of Pair Analysis Completed, Row Percentages (Standard Errors), 2012 NSDUH

Child Aged 12-17 Reported that Parent/Guardian Talked to Child about Dangers of Substance Use	Father Used an Illicit Drug in the Past Year		
	Yes	**No**	**Odds Ratio**
Yes	11.9% (2.17%)	88.1% (2.17%)	1.15 (0.33)
No	10.6% (1.88%)	89.4% (1.88%)	1.00

2.7   Regression Analyses

Instead of the contingency table sort of analysis described in the previous sections, an analyst might be interested in fitting a regression model. The steps leading up to a contingency table analysis and the steps leading up to a regression analysis are the same. In either scenario, the analyst must convert the person-level dataset to a pair-level dataset, describe the sample design, and create the weight. The weight that should be used is determined in the same manner as described in Section 2.4.2.

For example, an analyst might be interested in the same problem described in the previous sections but would like to estimate the odds ratio instead of the conditional probabilities. This would allow the analyst to estimate the odds of a father using an illicit drug in

the past year among fathers whose child reported that the father talked to him or her, relative to the odds of a father using an illicit drug in the past year among fathers whose child reported that the father did not talk to him or her. This odds ratio can easily be estimated using logistic regression. The following Stata code could be used:

```
svy, subpop (if (irprrel == 1 | irprrel == 2) & irsex_x == 1): logistic
sumyr_x ib2.yeprtdng_m[8]
```

3.   More Advanced Topics

The examples described in Chapter 2 are meant to provide the user with an outline for a basic pair-level analysis. This chapter provides more detail on topics relevant to the analysis of pair-level data.

3.1   Inferential Population and Multiplicities

As discussed earlier, there are different perspectives through which pair data can be analyzed: (1) with pairs as the focus (e.g., where the focus of the analysis, or dependent variable, is on how the parent and child jointly report behaviors of interest), or (2) with one member of the pair as the focus (e.g., the child's behavior is the dependent variable and his or her parent's behavior is the independent variable). When the focus is on the pair, the *PRANALWT* variable can be used to weight the data directly with no adjustments. However, when the focus is on one member of the pair, an adjustment often needs to be made to the weight to account for the fact that this member of the pair may be a member of more than one pair of this type that is present in the household (Chromy & Singh, 2001).

Section 2.4.2 describes the weights for the three examples introduced in Section 1.6. The first two of these three examples, examples 1 and 2, focus on one member of the pair (i.e., the behavior of one member of the pair is the dependent variable and behavior of the other pair member is the independent variable). The third of these examples focuses on the pair itself (i.e., the behaviors of both members of the pair are jointly the dependent variable). When the focus is on the pair itself, as is the case in example 3, no multiplicity is needed and the *PRANALWT* variable can be used directly.

Because examples 1 and 2 focus on one member of the pair, there is potentially a multiple counting problem. The *PRANALWT* variable weights each pair member by the number of pairs of that type that the respondent represents. In example 1, the father with three children is part of three father-child pairs; whereas, the child in the father-child pair is part of only one possible father-child pair. *PRANALWT* represents the number of father-child pairs, so the father is represented three times and the child is represented once. When the father's behavior is the focus, *PRANALWT* is more heavily weighted toward fathers with multiple children versus fathers with only one child. This introduces a multiple counting problem.

---

[8] In Stata, "ib2" is an indication that level 2 of the variable *yeprtdng_m* should be used as the reference level. In Stata, svy is a modifier to standard functions; refer to help logistic for further syntax options.

Adjustments for this multiple counting problem are done using "multiplicities" (Chromy & Singh, 2001). Many multiplicities have already been computed and are on the data file; in analyses, these can be used to adjust the weight variable. Analysts should exercise care or seek assistance through a data request (CBHSQRequest@samhsa.hhs.gov) in computing multiplicities for any pair types not listed. The process by which the multiplicities are created is described in detail in Chapter 11 of the 2012 NSDUH Methodological Resource Book Editing and Imputation Report (Laufenberg et al., 2014). Table 5 lists the pair domains for which multiplicities can be used to perform person-level analyses. These multiplicity variables are used in conjunction with the *PRANALWT* variable when adjusting the weight. The new weight is assigned the value of *PRANALWT* divided by the multiplicity variable.

Table 5.   Pair Domains and Multiplicities

IRPRREL Levels	Pair Domain	Description	Focus[1]	Name	Multiplicity Variable — Description
1	Children living with mothers or fathers	Parent-child, child aged 12-14	Parent	*IRMPCP14*	Number of children aged 12-14 living with responding parent
1			Child	*IRMPCC14*	Number of parents living with responding child aged 12-14
1,2		Parent-child, child aged 12-17	Parent	*IRMPCP17*	Number of children aged 12-17 living with responding parent
1,2			Child	*IRMPCC17*	Number of parents living with responding child aged 12-17
1,2,3		Parent-child, child aged 12-20	Parent	*IRMPCP20*	Number of children aged 12-20 living with responding parent
1,2,3			Child	*IRMPCC20*	Number of parents living with responding child aged 12-20
2		Parent-child, child aged 15-17	Parent	*IRMPCP57*	Number of children aged 15-17 living with responding parent
2			Child	*IRMPCC57*	Number of parents living with responding child aged 15-17
5	Siblings living with siblings	Sibling-sibling, older sibling aged 15-17, younger sibling aged 12-14	Older sibling	*IRMS1417*	Number of siblings aged 12-14 living with responding sibling aged 15-17
5			Younger sibling	*IRMS1714*	Number of siblings aged 15-17 living with responding sibling aged 12-14
6		Sibling-sibling, older sibling aged 18-25, younger sibling aged 12-17	Older sibling	*IRMS1725*	Number of siblings aged 12-17 living with responding sibling aged 18-25
6			Younger sibling	*IRMS2517*	Number of siblings aged 18-25 living with responding sibling aged 12-17
8,9	Spouses/ partners living together	Spouse-spouse and partner-partner			No multiplicity necessary: assume only one spouse per person
8		Spouse-spouse and partner-partner, with children aged 0-17			No multiplicity necessary: assume only one spouse per person

[1] No weight adjustment is needed when the inferential focus is on the pair.

3.2 Pooling of Data across Years

An analyst may be interested in pooling NSDUH data across several years in order to boost the sample size. The extra sample may support analyses that cannot be done with the desired precision when only a single year of data is used. To do this, the single-year pair-level datasets need to be concatenated. The analyst can (1) convert the person-level data to pair-level data for each year, and then concatenate the converted files, or (2) concatenate the person-level datasets first, and then convert the concatenated dataset to the pair-level dataset. Either way, it is good practice for the analyst to include a variable indicating the survey year because ENCCASE is not guaranteed to be unique across years. That way, the analyst is guaranteed to have variables that uniquely identify each record (ENCCASE plus the year indicator). If the analyst is interested in an annual estimate of a population total, the weight should be divided by the number of years that were pooled. For example, if you want to average 3 years of NSDUH data, you would adjust the weight *PRANALWT* by dividing by 3.

3.3 Stata® Code for Pooling Data across Years

```
/*Section 3.3: Pooling Data Across Years*/
use "PairLevel_2009.dta", clear

gen year = 2009
gen pradjwt3 = pranalwt/3
save PairLevel_2009_new.dta

use "PairLevel_2010.dta", clear

gen year = 2010
gen pradjwt3 = pranalwt/3
save PairLevel_2010_new.dta

use "PairLevel_2011.dta", clear

gen year = 2011
gen pradjwt3 = pranalwt/3
save PairLevel_2011_new.dta

append using PairLevel_2009_new.dta PairLevel_2010_new.dta
save PairLevel_Total.dta
```

4. Limitations of Pair-Level Analyses

4.1 Efficiency Gains through Using Person-Level Analyses

As mentioned in the introduction of this manual and in Section 2.1, because some characteristics of the dwelling unit might be known from a single interview, some analyses look like pair analyses but in reality can be done more efficiently at the person level. For example, one might be interested in whether 12- to 17-year-old children who live with their father tend to use drugs more often than those who do not live with their father (Table 6).

Table 6.   Example of Analysis that Could Be Done at the Pair Level or Person Level

Lives with Father	Child Used Drugs	
	Yes	**No**
Yes	n_{11}	n_{12}
No	n_{21}	n_{22}

The variable *IFATHER* is an indicator of whether there is a father in the dwelling unit. If completed using pair data, the sample size for this analysis is limited to the number of father-child pairs in the sample where the child is aged 12 to 17 (see Appendix A for a yearly estimate of the pair sample sizes). However, if *IFATHER* is used and the analysis is performed at the person level, the sample would include all 12- to 17-year-old respondents, which is a much larger number. A larger sample would likely lead to estimates with lower standard errors.[9]

As mentioned in the introduction to this manual, analysts should be aware of which variables are available with information on members of the dwelling unit from a single interview to avoid conducting analyses using pair data instead of person-level data when pair data are not necessary. The following sections of the 2013 NSDUH analytic codebook (Center for Behavioral Health Statistics and Quality, 2014) provide information about other members of the dwelling unit: "Parenting Experiences," "Youth Experiences," "Household Composition (Roster)," "Roster Pair Variables," and "Income." Note that only the variables in the "Roster Pair Variables" section are required to be consistent across members of a pair (i.e., they are pair-level variables).

4.2   Absence of Triads

No more than two respondents are surveyed from a single dwelling unit, regardless of the number of individuals living in the dwelling unit. Therefore, there are limitations on the relationships that can be analyzed. For example, weights can be created for an analysis on father-child or mother-child pairs where the target population is the child and the child lives with at least one parent. However, it would be impossible to perform an analysis that investigates attributes of both parents as well as the child. For that reason, the pair data do not support analyses incorporating the behavior of both parents or the behavior of either parent (i.e., pooling mother-child and father-child pairs for the analysis of "paren*ts*-child" relationships). More information on this limitation can be found in Appendix G. Some of these analyses can be done by combining analyses that focus on each of the three pairs that comprise the triad, but these are complex. Table 7 shows an example of an analysis that cannot be done in a straightforward manner because of the absence of triads in the sample. Note that the analysis cannot be performed even by separately analyzing the three pairs that comprise the triad (i.e., mother-child, father-child, and mother-father), because the "Parenting Experiences" module is not given unless the respondent is the parent in a parent-child pair; therefore, parents in a mother-father pair did not receive these questions.

---

[9] Another benefit of using person-level data where appropriate is that IMOTHER and IFATHER may include parents in the military who are not eligible for NSDUH and would not be captured using pair analyses.

Table 7. **Example of Analysis that Would Be Straightforward Only if the Sample Included Triads**

Both Parents Report Talking to Child about Drugs	Child Used Drugs	
	Yes	No
Yes	n_{11}	n_{12}
No	n_{21}	n_{22}

References

Ashley, O. S., Penne, M. A., Loomis, K. M., Kan, M., Bauman, K. E., Aldridge, M., Gfroerer, J. C., & Novak, S. P. (2008). Moderation of the association between parent and adolescent cigarette smoking by selected sociodemographic variables. *Addictive Behaviors, 33*(9), 1227-1230.

Center for Behavioral Health Statistics and Quality. (2012). *2011 National Survey on Drug Use and Health: Final analytic file codebook*. Rockville, MD: Substance Abuse and Mental Health Services Administration.

Center for Behavioral Health Statistics and Quality. (2014). *2013 National Survey on Drug Use and Health: Final analytic file codebook*. Rockville, MD: Substance Abuse and Mental Health Services Administration.

Center for Behavioral Health Statistics and Quality. (2015). *2014 National Survey on Drug Use and Health: Methodological resource book (Section 2, Sample design report)*. Rockville, MD: Substance Abuse and Mental Health Services Administration. Retrieved from https://www.samhsa.gov/data/

Center for Behavioral Health Statistics and Quality. (2016a). *2014 National Survey on Drug Use and Health: Methodological resource book (Section 13, Statistical inference report)*. Rockville, MD: Substance Abuse and Mental Health Services Administration. Retrieved from https://www.samhsa.gov/data/

Center for Behavioral Health Statistics and Quality. (2016b). *Evaluation of imputation methods for the National Survey on Drug Use and Health* (unpublished internal documentation). Rockville, MD: Substance Abuse and Mental Health Services Administration.

Chromy, J. R., & Singh, A. C. (2001). Estimation for person-pair drug-related characteristics in the presence of pair multiplicities and extreme sampling weights. In *Proceedings of the 2001 Joint Statistical Meetings*, *American Statistical Association, Survey Research Methods Section, Atlanta, GA* [CD-ROM]. Alexandria, VA: American Statistical Association. [Available as a PDF at http://www.amstat.org/ASA/Membership/Sections-and-Interest-Groups.aspx]

Laufenberg, J., Kroutil, L., Frechtel, P., Carpenter, L., Edwards, S., Ewing, G., Gulledge, K., Handley, W., Martin, P., Moore, A., & Scott, V. (2014). Editing and imputation report. In *2012 National Survey on Drug Use and Health: Methodological resource book* (Sections 10/11, prepared for the Substance Abuse and Mental Health Services Administration, Contract No. HHSS283201000003C, Deliverable No. 41, RTI/0212800.001.107.006.007). Research Triangle Park, NC: RTI International. Retrieved from https://www.samhsa.gov/data/

Westlake, M., Chen, P., & Gordek, H. (2013). Questionnaire dwelling unit-level and person pair-level sampling weight calibration. In *2011 National Survey on Drug Use and Health: Methodological resource book* (Section 13, prepared for the Substance Abuse and Mental Health Services Administration, Contract No. HHSS283200800004C, Deliverable No. 39, RTI/0211838.207.005). Research Triangle Park, NC: RTI International. Retrieved from https://www.samhsa.gov/data/

Appendix A: Distribution of Responding Pairs in 2002-2015 NSDUHs

Table 8.   **Distribution of Responding Pairs, 2002-2013 and 2014-2015 NSDUHs**

Imputation-Revised Pair Type (IRPRREL)	2002-2013			2014-2015		
	Number of Pairs[1]	Unweighted Percentage among All Pairs	Unweighted Percentage among All Respondents	Number of Pairs[1]	Unweighted Percentage among All Pairs	Unweighted Percentage among All Respondents
1 = Parent-child, child aged 12-14	25,300	10.6	6.2	5,600	15.4	8.2
2 = Parent-child, child aged 15-17	21,500	9.1	5.3	4,800	13.1	7.0
3 = Parent-child, child aged 18-20	8,700	3.7	2.1	1,700	4.8	2.5
4 = Parent-child, child aged 21+	10,000	4.2	2.4	2,100	5.8	3.1
5 = Sibling-sibling, younger sibling aged 12-14, older sibling aged 15-17	26,800	11.3	6.5	2,700	7.5	4.0
6 = Sibling-sibling, younger sibling aged 12-17, older sibling aged 18-25	28,900	12.2	7.1	3,300	9.1	4.8
7 = Other sibling-sibling pair	33,700	14.2	8.2	4,100	11.2	6.0
8 = Spouse-spouse, with children aged 0-17	22,300	9.4	5.5	3,600	10.0	5.3
9 = Spouse-spouse, without children aged 0-17	24,900	10.5	6.1	3,400	9.3	5.0
11 = Grandparent-grandchild	2,400	1.0	0.6	500	1.5	0.8
12 = Other clearly identifiable pair (e.g., roommate)	32,000	13.5	7.8	4,200	11.6	6.2
13 = Pair type not clear, definitely not pair codes 1-11	1,300	0.5	0.3	200	0.7	0.4
Total = All pairs	237,800	100.0	58.2	36,200	100.0	53.2

NOTE: The distribution in 2014 onward will differ because of changes in sampling rates. Data portal users will have access to 2004 onward.

[1] Rounded to the nearest 100.

Appendix B: NSDUH Codebook Descriptions of Variables Described in the Manual

QUESTID Len: 7   RESPONDENT IDENTIFICATION
ENCCASE   Len: 10 HOUSEHOLD LEVEL IDENTIFICATION NUMBER (ENCRYPTED)

SUMYR   Len: 1   ANY ILLICIT DRUG - PAST YEAR USE
        0 = Did not use in past year (ALL SOURCE VARS = 0)
        1 = Used in past year (ANY SOURCE VARIABLE = 1)

YEPRTDNG   Len: 2   HAVE YOU TALKED W/PARENT(S) ABT DANGER OF TOB/ALC
        1 = Yes
        2 = No
        85 = BAD DATA Logically assigned
        89 = LEGITIMATE SKIP Logically assigned
        94 = DON'T KNOW
        97 = REFUSED
        98 = BLANK (NO ANSWER)
        99 = LEGITIMATE SKIP

PXKIDYR Len: 2 # TIMES TALK W/CHLD ABT PROB TOB/ALC/DRUG 12 MOS
        1 = 0 times
        2 = 1 to 2 times
        3 = A few times
        4 = Many times
        89 = LEGITIMATE SKIP Logically assigned
        94 = DON'T KNOW
        97 = REFUSED
        98 = BLANK (NO ANSWER)
        99 = LEGITIMATE SKIP

PRNTIND   Len: 2 1/0: IS THIS RESPONDENT PRNT IN CHLD-PRNT PAIR?
        0 = Parent-child pair, not a parent
        1 = Parent-child pair, a parent
        9 = In another pair, not child-parent
        99 = Not in a pair

IRPRREL Len: 2 IMP REV FAMILY PAIR RELATIONSHIP INDICATOR
        1 = Parent-child, child is 12-14
        2 = Parent-child, child is 15-17
        3 = Parent-child, child is 18-20
        4 = Parent-child, child is 21 or over
        5 = Sibling-sibling, younger is 12-14, older is 15-17
        6 = Sibling-sibling, younger is 12-17, older is 18-25
        7 = Other sibling-sibling pair
        8 = Spouse-spouse with children aged 0-17
        9 = Spouse-spouse without children aged 0-17
        11 = Grandparent-grandchild
        12 = Other clearly identifiable pair (e.g. roommate)
        13 = Pair type not clr, defntly not pr codes 1 thru 11
        99 = Not in a pair

IRMPCC14 Len: 2 IMP REV MLTPLCTY: CHLD-PRNT, CH FOCUS, CH IS 12-14
   1 = Child has one parent
   2 = Child has two parents
   99 = Not member of imp rev chld-prnt pr w/12-14 yr old

IRMPCC17 Len: 2 IMP REV MLTPLCTY: CHLD-PRNT, CH FOCUS, CH IS 12-17
   1 = Child has one parent
   2 = Child has two parents
   99 = Not member of imp rev chld-prnt pr w/12-17 yr old

IRMPCC20 Len: 2 IMP REV MLTPLCTY: CHLD-PRNT, CH FOCUS, CH IS 12-20
   1 = Child has one parent
   2 = Child has two parents
   99 = Not member of imp rev chld-prnt pr w/12-20 yr old

IRMPCC57 Len: 2 IMP REV MLTPLCTY: CHLD-PRNT, CH FOCUS, CH IS 15-17
   1 = Child has one parent
   2 = Child has two parents
   99 = Not member of imp rev chld-prnt pr w/15-17 yr old

IRMPCP14 Len: 2 IMP REV MLTPLCTY: CHLD-PRNT,PRNT FOCUS,CH IS 12-14
   RANGE = 1 - 3
   99 = Not member of imp rev chld-prnt pr w/12-14 yr old

IRMPCP17 Len: 2   IMP REV MLTPLCTY: CHLD-PRNT,PRNT FOCUS,CH IS 12-17
   RANGE = 1 - 5
   99 = Not member of imp rev chld-prnt pr w/12-17 yr old

IRMPCP20 Len: 2   IMP REV MLTPLCTY: CHLD-PRNT,PRNT FOCUS,CH IS 12-20
   RANGE = 1 - 6
   99 = Not member of imp rev chld-prnt pr w/12-20 yr old

IRMPCP57 Len: 2   IMP REV MLTPLCTY: CHLD-PRNT,PRNT FOCUS,CH IS 15-17
   RANGE = 1 - 4
   99 = Not member of imp rev chld-prnt pr w/15-17 yr old

IRMS1417 Len: 2   IMP REV MLTPLCTY: SIB-SIB(12-14/15-17),15-17 FOCUS
   RANGE = 1 - 3
   99 = Not member of imp rev sib-sib pr in these age grps

IRMS1714 Len: 2   IMP REV MLTPLCTY: SIB-SIB(12-14/15-17),12-14 FOCUS
   RANGE = 1 – 3
   99 = Not member of imp rev sib-sib pr in these age grps

IRMS1725 Len: 2   IMP REV MLTPLCTY: SIB-SIB(12-17/18-25),18-25 FOCUS
   RANGE = 1 - 5
   99 = Not member of imp rev sib-sib pr in these age grps

IRMS2517 Len: 2   IMP REV MLTPLCTY: SIB-SIB(12-17/18-25),12-17 FOCUS
   RANGE = 1 - 4
   99 = Not member of imp rev sib-sib pr in these age grps

IRHPCC14 Len: 1   IMP REV HHLD CNT: CHLD-PRNT, CH FOCUS, CH IS 12-14
   RANGE = 0 - 4

IRHPCC17   Len: 1   IMP REV HHLD CNT: CHLD-PRNT, CH FOCUS, CH IS 12-17
                RANGE = 0 - 6

IRHPCC20   Len: 1   IMP REV HHLD CNT: CHLD-PRNT, CH FOCUS, CH IS 12-20
                RANGE = 0 - 7

IRHPCP14   Len: 1   IMP REV HHLD CNT: CHLD-PRNT,PRNT FOCUS,CH IS 12-14
                RANGE = 0 - 4

IRHPCP17   Len: 1   IMP REV HHLD CNT: CHLD-PRNT,PRNT FOCUS,CH IS 12-17
                RANGE = 0 - 4

IRHPCP20   Len: 1   IMP REV HHLD CNT: CHLD-PRNT,PRNT FOCUS,CH IS 12-20
                RANGE = 0 - 4

IRHS1417   Len: 1   IMP REV HHLD CNT: SIB-SIB(12-14/15-17),15-17 FOCUS
                RANGE = 0 - 4

IRHS1725   Len: 1   IMP REV HHLD CNT: SIB-SIB(12-17/18-25),18-25 FOCUS
                RANGE = 0 – 4

IRHCSPSP   Len: 1   IMP REV HHLD CNT: NUMBER OF SPOUSE-SPOUSE PRS HHLD
                RANGE = 0 - 3

IRHCSPWC   Len: 1   IMP REV HHLD CNT: NMBR SPOUSE-SPOUSE PRS W/ CHLDRN
                RANGE = 0 – 2

AGE   Len: 3   FINAL EDITED AGE
                RANGE = 12 - 99

IRSEX   Len: 1   IMPUTATION REVISED GENDER
                1 = Male
                2 = Female

WT1   Len: 12   INVERSE PROBABILITY OF SELECTING CENSUS TRACT

WT2   Len: 12   INVERSE PROBABILITY OF SELECTING SEGMENT

WT3   Len: 12   QUARTER SEGMENT WEIGHT ADJUSTMENT

WT4   Len: 12   SUBSEGMENTATION INFLATION ADJUSTMENT

WT5   Len: 12   INVERSE PROBABILITY OF SELECTING DWELLING UNIT

WT6   Len: 12   SUBSAMPLING OF ADDED DWELLING UNIT ADJUSTMENT

WT7   Len: 12   DWELLING UNIT PERCENT RELEASE ADJUSTMENT

WT8   Len: 12   DU NON-RESPONSE ADJUSTMENT

WT9   Len: 12   DU POSTSTRATIFICATION ADJUSTMENT

WT10   Len: 12   DU EXTREME WEIGHT ADJUSTMENT

DUANALWT   Len: 12   FINAL QUESTIONNAIRE DU LEVEL WEIGHTS

PRANALWT   Len: 12   2012 FINAL PERSON PAIR LEVEL ANALYSIS WEIGHTS

PRVEREP   Len: 12   PERSON PAIR LEVEL VERSION OF ANALYSIS REPLICATE

PRVESTR   Len: 12   PERSON PAIR LEVEL VERSION OF ANALYSIS STRATUM

Appendix C: SPSS Code for Chapter 2 Examples

```
*Retrieve data from SAS.
GET
 SAS DATA='<location of data set>.sas7bdat'.
DATASET NAME DataSet12 WINDOW=FRONT.

*sort dataset by ascending ENCCASE, descending PRNTIND, ascending AGE.
DATASET ACTIVATE DataSet12.
MATCH FILES FILE=DataSet12 /KEEP = IRPRREL PRANALWT ENCCASE PRNTIND AGE PRVESTR
PRVEREP YEPRTDNG SUMYR IRSEX.
SORT CASES BY encCASE(A) prntind(D) .

*Create lastcase variable and save as data12_withlast.sav.
DATASET ACTIVATE DataSet12.
MATCH FILES FILE=DataSet12 /BY enccase / LAST=lastcase.
DATASET NAME DataSet12_withlast.

DATASET ACTIVATE DataSet12_withlast.

* Create lag variables.
COMPUTE lagSUMYR=LAG(sumyr).
COMPUTE lagYEPRTDNG=LAG(yeprtdng).
COMPUTE lagIRSEX=LAG(irsex).

* Assign pair-level indicators.
DATASET ACTIVATE DataSet12_withlast.
DO IF (lastcase EQ 1).
COMPUTE sumyr_x=lagSUMYR.
COMPUTE sumyr_y=sumyr.
COMPUTE YEPRTDNG_X=lagYEPRTDNG.
COMPUTE YEPRTDNG_Y=YEPRTDNG.
COMPUTE irsex_X=lagIRSEX.
COMPUTE irsex_Y=irsex.
END IF.

DATASET ACTIVATE DataSet12_withlast.
RECODE yeprtdng_Y (1=1) (2=2) INTO yeprtdng_m.
VARIABLE LABELS yeprtdng_m 'Child reports parents talked to child about substance use'.
EXECUTE.

SELECT IF lastcase eq 1.
SELECT IF (IRPRREL NE 99).
EXECUTE.

DATASET ACTIVATE DataSet12_withlast.
SORT CASES BY PRVESTR(A) PRVEREP(A).

* Create variable needed for analysis.
DO IF (irsex_X = 1 & (irprrel=1 OR irprrel=2)).
COMPUTE IRPRREL2=1.
ELSE.
COMPUTE IRPRREL2=0.
END IF.
```

```
EXECUTE.
VARIABLE LABELS IRPRREL2 'Indicator for father-child pair'.
EXECUTE.

* Recode missing level for yeprtdng_m so none are missing.
DATASET ACTIVATE DataSet12_withlast.
RECODE yeprtdng_m (SYSMIS=3).
EXECUTE.

* Create revised weight to account for multiplicity.
DATASET ACTIVATE DataSet12_withlast.

COMPUTE NEWWT= PRANALWT/IRMPCP17.
EXECUTE.


CSPLAN ANALYSIS
 /PLAN FILE= '<filepath>.csaplan'
 /PLANVARS ANALYSISWEIGHT=NEWWT
 /PRINT PLAN
 /DESIGN STRATA=prvestr CLUSTER=prverep
 /ESTIMATOR TYPE=WR.

* Perform analysis.
CSDESCRIPTIVES
 /PLAN FILE='<filepath>.csaplan'
 /SUMMARY VARIABLES=sumyr_x
 /SUBPOP TABLE =yeprtdng_m by IRPRREL2
 /MEAN
 /STATISTICS SE COUNT
 /MISSING scope = analysis .
```

Appendix D: R Code for Chapter 2 Examples

```
# Install the pylr library, if necessary.
install.packages("plyr")
library(plyr)

setwd("//location of data set ")
PersLevel_2012 <- read.csv(file="Perslevel_2012.csv")

#Section 2.2.3: Sort dataset
PersLevel_2012_sorted <- PersLevel_2012[with(PersLevel_2012, order(encCASE, -prntind )),]

#Section 2.2.4: Convert dataset to pair-level
# creating an identifier (personid) for people within household
PersLevel_2012_sorted$personid=rep(c("x","y"),nrow(PersLevel_2012_sorted)/2)
PairLevel_2012=reshape(PersLevel_2012_sorted,idvar="encCASE",direction="wide",
        v.names=c("SUMYR","YEPRTDNG","irsex"),timevar="personid",
        drop=c("questid","age","prntind"))


# Section 2.4: Run Pair-Level Analysis
# First sort by the NEST variables.
sorted <- PairLevel_2012[with(PairLevel_2012, order(prvestr, prverep)), ]

colnames(sorted)=c(names(sorted)[1:5],gsub(".y","_Y",gsub(".x","_X",tolower(names(sorted)[6:11]),fixed=
T), fixed=T))

# checking names
names(sorted)

# Create independent variable YEPRTDNG_M.
talked <- function(vname){
 if(is.na(vname)) return(NA)
 if(vname == 1) return(1) # Parent talked
 if(vname == 2) return(2) # Parent did not talk
 return(NA) # This variable is not imputed.
}

# Create revised weight to account for multiplicity.
sorted$NEWWT <- sorted$pranalwt/sorted$irmpcp17

sorted$yeprtdng_M <- sapply(sorted$yeprtdng_Y, talked)
#write.csv(sorted, file = "sorted_r.csv",row.names=FALSE)

# Estimate means and SEs.
# two-stage with replacement
install.packages("survey")
library(survey)

dclus2wr<-svydesign(strat=~prvestr, id=~prverep, weights=~NEWWT, data=sorted, nest=TRUE)
svyby(~sumyr_X, ~yeprtdng_M, subset(dclus2wr,(irprrel==1 | irprrel==2) & irsex_X==1), svymean)
```

Appendix E: SAS®/SAS Survey/SAS-Callable SUDAAN® Code for Chapter 2 Examples

```sas
LIBNAME in12 '<location of data set>';

/*Section 2.2.2: Reduce file to only members of pairs*/
data perslevel;
    SET in12.mast12 (KEEP = ENCCASE QUESTID IRPRREL PRANALWT IRMPCP17 PRVESTR
    PRVEREP AGE IRSEX PRNTIND YEPRTDNG SUMYR);
    WHERE IRPRREL ^= 99;
run;

/*Section 2.2.3: Sort dataset*/
proc sort DATA = perslevel;
    BY ENCCASE descending PRNTIND;
run;

/*Section 2.2.4: Convert dataset to pair-level. The TRANSPOSE procedure can also be used to do this. */
data pairLevel;
 SET persLevel;
 BY enccase;
 lagSUMYR=lag(sumyr);
 lagYEPRTDNG =lag(yeprtdng);
 lagIRSEX=lag(irsex);
 if last.enccase then do;
  sumyr_X=lagSUMYR;
  sumyr_Y=sumyr;
  yeprtdng_X=lagYEPRTDNG;
  yeprtdng_Y=yeprtdng;
  irsex_X=lagIRSEX;
  irsex_Y=IRSEX;
  output;
 end;
 drop prntInd age sumyr lagSUMYR yeprtdng lagYEPRTDNG irsex lagIRSEX questid;
run;

data pairlevel2;
    SET pairlevel;
    *create the independent variable (Section 2.4.4);
    if yeprtdng_y in (1,2) then yeprtdng_m = yeprtdng_y=;
     else yeprtdng_m = .;
    *Create revised weight to account for multiplicity;
    NEWWT=PRANALWT/IRMPCP17;

run;

ods listing;
proc freq data=pairlevel2;
    tables yeprtdng_m* yeprtdng_y/list missing;
run;

/*Section 2.4: Sort by NEST variables*/
proc sort data=pairLevel2; by PRVESTR PRVEREP; run;
```

```
data pairlevel3;
   set pairlevel2;
   if irprrel in (1,2) and irsex_x = 1 and yeprtdng_m =1 then subpop1_1 = 1;
    else subpop1_1 =0;
   if irprrel in (1,2) and irsex_x = 1 and yeprtdng_m =2 then subpop1_2 = 1;
    else subpop1_2 =0;
run;

/*Section 2.4: Run Pair Analysis using SAS survey procedure */
proc surveymeans data = pairlevel3 varmethod = Taylor;
   strata prvestr;
   cluster prverep;
   domain subpop1_1;
   var sumyr_X;
   weight NEWWT;
run;

proc surveymeans data = pairlevel3 varmethod = Taylor;
   strata prvestr;
   cluster prverep;
   domain subpop1_2;
   var sumyr_X;
   weight NEWWT;
run;

/*Section 2.4 alternate: run pair analysis using SAS-callable SUDAAN */
proc descript data= pairLevel2 filetype=sas design=wr;
      nest PRVESTR PRVEREP;
      weight NEWWT; /*NEWWT=PRANALWT/IRMPCP17*/
      var SUMYR_X;
      subgroup YEPRTDNG_M;
      levels 2;
      tables YEPRTDNG_M ;
      subpopx IRPRREL in (1,2) and IRSEX_X=1;
      print nsum mean semean/ meanfmt=f8.5 semeanfmt=f8.5;
run;
```

Appendix F: List of Pair-Level Variables on RUFs

There are two sets of pair-level variables on the RUFs. The first set stores information related to the dwelling unit, which is known before the questionnaires are filled out. These are mainly geographic variables, variables involved in the sampling of dwelling units, and variables related to the screening process. The second set includes variables that are reconciled across pair members during post-survey data processing. The variables in these two sets from the 2011 NSDUH analytic codebook (Center for Behavioral Health Statistics and Quality, 2012) are listed below. These sets are similar in NSDUH years other than 2011. The variables in the first set are described below.

- "Identification" module: *ENCSEG, ENCCASE*
- "Interview Information" module: *GQTYPE, SDUMIL, DUCHAR1–DUCHAR10, CATYPE1–CATYPE5*
- "Core Demographics" module: *SKPX2PER, SKPX1217, SKPXPRNT, STATELOC*
- "Geographic" module: all variables in this module
- "Sample Weighting and Estimation Vars" module: all variables are unique to the pair, except for *ANALWT* and *WT11–WT15*

The second set of pair-level variables includes all variables in the "Roster Pair Variables" and "Imputed Roster Pair Variables" modules. The reconciliation process between members of the pair is described in detail in Chapter 11 of the 2012 NSDUH Methodological Resource Book Editing and Imputation Report (Laufenberg et al., 2014).

There are many variables where the pair members usually agree, but it is not guaranteed they will agree. Variables that usually, but not always, agree include those in the "Household Composition (Roster)" and "Income" modules. Disagreement could occur because of respondent error or changes that occurred to the dwelling unit between the times the pair members responded to the questionnaire. For example, the first pair member may report that there are four people living in the dwelling unit, but one resident moves out the next day, which affects the other pair member's response when he or she fills out the questionnaire a few days later. A detailed analysis of variables of this type can be found in Chapter 8 of the *Evaluation of Imputation Methods for the National Survey on Drug Use and Health* (Center for Behavioral Health Statistics and Quality, 2016b). Analysts should use caution if they choose to convert these variables to the pair level.

Appendix G: Combining Mother-Child and Father-Child Pairs for Comparison Purposes

An analyst may be interested in evaluating parent-child pairs by combining the mother-child and father-child pairs into one analysis for two purposes: (1) to examine the association of parents' behaviors and the child's behavior (versus just the father's or mother's behavior), and (2) to compare the relative association between mothers and fathers on the children's outcome. As discussed in Section 4.2, because triads are not captured, the effect of "both parents" or "either parent" cannot be determined, but rather a mother's effect and a father's effect can be determined. However, when a child lives with more than one parent, even these mother-child and father-child relationships are influenced by the nonsampled parents. This adversely affects "parent-child" conclusions if there is a disparate rate of behaviors or characteristics between mothers and fathers.

After discussing the absence of triads, it is relatively straightforward that there is not enough information to conclude that both parents discussed dangers of substance use with their child, but consider the following question: "Did at least one parent discuss dangers of substance use with his or her child?" This is the type of question that might be of interest in a parent-child analysis in which the mother-child and father-child pairs are (mistakenly) combined. Table 9 shows the two scenarios, in italics, in which the incorrect conclusions are drawn if the father-child or mother-child pair results are used to answer a parent-child pair question. As outlined in the previous paragraph, the data will incorrectly answer this question with "No" when the sampled parent is not the parent who discussed dangers with his or her child. This leads to an artificially low estimate for the number of parents who discussed dangers with their child.

Table 9.   **Information Obtained from Pair Data versus Reality for Parent-Child Pairs**

	Reality			
Sampled Parent	**Neither Parent Discussed Dangers**	**Only Mother Discussed Dangers**	**Only Father Discussed Dangers**	**Both Mother and Father Discussed Dangers**
Mother Sampled	Data: No Answer: No	Data: Yes Answer: Yes	*Data: No* *Answer: Yes*	Data: Yes Answer: Yes
Father Sampled	Data: No Answer: No	*Data: No* *Answer: Yes*	Data: Yes Answer: Yes	Data: Yes Answer: Yes

NOTE: Italics represent incorrect conclusions.

It is recommended that analysts do not combine mother-child and father-child pairs to evaluate a general "parent" association. There have been several manuscripts published that analyze both mother-child and father-child pair-level NSDUH data. In a manuscript by Ashley et al. (2008), which examined the association between parent and adolescent smoking behaviors, mother-child pairs and father-child pairs were analyzed, but they were always analyzed separately and were not combined into one analysis to measure the association between parent and child.

Although these pairs should not be combined overall, analysts have used both mother-child and father-child pairs to compare the effect of the mother versus the father on a specific behavior by using an interaction effect in a regression model. This is programmatically feasible using a statistical software package, but the caveats should be seriously considered when performing this type of analysis. In this example, a model is fit including the genders of both parents, the covariate of interest (e.g., whether the parent discussed substance use dangers with the child), and the interaction between these two terms. The weight to use is PRANALWT. The output for the interaction term can help the user determine whether the effect for talking to the child differs by parent gender. The analysis should then be run stratified by the mother and the father to obtain mother-child or father-child associations by parent gender.

In this example, although this test can be performed using the code in this appendix, the results observed compare the marginal effect of mothers versus the marginal effect of fathers. What is possibly of greater interest is the joint or conditional effects, which are not available with the pair data. If an analyst wants to get a clearer idea of the effect of the individual parents, it is suggested that the analyst consider reducing the analysis to pairs in a single-parent household.

Stata code is provided below for comparing the association of the mother's behavior and the child's behavior using all sampled pairs where the mother and child live in the same household with the association of the father's behavior and the child's behavior using all sampled pairs where the father and child live in the same household. The variable for gender of the parent is *IRSEX_X*, and the variable for the substance use behavior of the child is *SUMYR_Y*. The covariate of interest, whether the father or mother discussed the dangers with their child, is *PTALKDICHOT*.

```
/*Appendix G: testing interaction effect*/
svyset prverep [pweight=pranalwt], strata(prvestr) vce(linearized) dof(900)

svy linearized, subpop(if irprrel==1 | irprrel==2) dof(900) : logistic
sumyr_Y ib2.pTalkDICHOT ib2.irsex_X irsex_X#pTalkDICHOT
```

NOTE: pTalkDICHOT in this example is a dichotomous version of PXKIDYR found in Appendix B.