

## WELFARE, CHILDREN, AND FAMILIES: A THREE-CITY STUDY, WAVE 3 USER'S GUIDE

This volume documents the contents of survey data from *Welfare, Children and Families: A Three-City Study, Wave 3*. The data come from interviews conducted between February 2005 and January 2006 with 2,056 caregivers and 1,944 children who resided in Boston, Chicago, or San Antonio at wave 1 in 1999.

### **Purpose of the study**

The Welfare, Children and Families Study is a longitudinal study of children and their caregivers in low-income families that were living in low-income neighborhoods in three cities in 1999. The purpose of the study is to investigate the well-being of low-income children and their families after passage of the Personal Responsibility and Work Opportunity Reconciliation Act of 1996 (PRWORA)<sup>1</sup>. The survey was designed to provide information on the health and cognitive, behavioral, and emotional development of children and on their primary caregivers' labor force behavior, welfare experiences, family lives, use of social service, health, and well-being. A detailed description of the research design can be found in *Welfare, Children and Families: A Three City Study, Overview and Design*, available at [www.threecitystudy.jhu.edu](http://www.threecitystudy.jhu.edu) or in hardcopy upon request.

### **Wave 2 Sample**

We assume that the reader is familiar with the documentation from the first and second waves of this study, and we will not review all of the material presented there. See the design report described above or the Wave 1 User's Guide for a summary of how the sample was drawn and a description of each of the cities studied. This section describes the sample at wave 3 only.

At wave 1, the data file contained one record for each household, with data from the caregiver and focal child data on the same record. In contrast, at 3 (as in wave 2) we have *three* data files, two for caregivers and one for focal children. This is because not all caregiver/child pairs observed at wave 1 remained together at wave 3. Therefore, we followed to their new homes both caregivers and children who separated. This design enables users to study both caregivers and focal children longitudinally, without losing from the sample any dyad that separated between waves. In addition to tracking focal children, we interviewed children's new caregivers, using a modified version of the interview administered to continuing caregivers. This structure results in four types of respondents: continuing, separated, or new caregivers; and focal children. Continuing and new caregivers have been combined into a single file (N=1,835); separated caregivers are in another file (N=221); and data from focal children are in the third file (N=1,944). It is possible for one household from wave 1 to be represented on each of the three

---

<sup>1</sup>The survey is one component of a multidisciplinary project that also includes an embedded developmental study of about 600 children age 2 to 4 in 1999 and an ethnographic study of about 250 families residing in the same neighborhoods as the survey families.

files: in the case that a caregiver and focal child have separated, that original household may be represented at wave 3 by a new caregiver interview, a separated caregiver interview, and a focal child interview.

The following table shows the pairings for focal children and caregivers at wave 3:

Table 1. Focal child/caregiver interview pairings

<u>Interview Pair</u>	<u>N</u>
Focal child/wave 1 continuing caregiver	1698
Focal child/wave 2 new caregiver	10
Focal child/new caregiver/separated wave 1 caregiver	49
Focal child/wave 2 new caregiver/separated wave 1 caregiver	11
Focal child/new caregiver only	31
Focal child/separated caregiver only	102
Focal child only (no caregiver interview)	43
Total	1944

Among caregivers, there are 34 continuing caregiver interviews, 59 separated caregiver interviews, and 2 new caregiver interviews that are not matched to a focal child interview.

In 21 cases in the table above, focal children continue to reside with the caregiver who was interviewed as a new caregiver at wave 2. At wave 3, that respondent is interviewed as a continuing caregiver.

The set of possible interview scenarios is described in a variable called SCENARIO that is included on each data file. The variable values are described in Table 2:

Table 2. Set of possible interview scenarios, described in the variable SCENARIO

<u>Value</u>	<u>Interview scenario</u>
A	Focal child lived with same caregiver at each wave
B	Focal child was with new caregiver at wave 2, with same caregiver at wave 3 (wave 1 caregiver is separated at wave 3)
C	Focal child was with wave 1 caregiver at wave 2 and with a new caregiver at wave 3
D	Focal child was with a new caregiver at wave 2, reunited with wave 1 caregiver at wave 3 (wave 2 caregiver not followed)
E	Focal child is living independently at wave 3 (wave 1 caregiver interviewed as a separated caregiver)

Note that new caregivers at wave 2 who were no longer caring for the focal child at wave 3 are not followed.

Because households at wave 3 may have multiple associated caregivers, we have created a new unique person-level identifier by which to identify individual caregivers in cross-wave analyses.

See the section on unique identifiers later in this user's guide for more information on the new variable, called NEWID.

## Caregivers

### *Continuing and new caregivers*

The wave 3 sample includes interviews with 1,753 continuing caregivers. These are women who were caring for the focal child at wave 1, and who continued to care for the child at the time of the wave 3 interview. These may include women who were separated from the focal child at wave 2 and reunited at wave 3. In addition, the wave 2 sample includes 82 new caregiver interviews.

Because of errors in the field, 5 continuing caregivers were erroneously interviewed as new caregivers. These cases *were not* converted to continuing caregiver interviews during data cleaning, but they may be identified by the flag variable CGTFL (CGTFL=1 if the respondent was mistakenly interviewed as a new caregiver). The erroneous interview type will only be a concern for researchers conducting longitudinal analyses, but using NEWID, the cross-wave unique person-level identifier, should obviate potential problems.

We have combined data from continuing and new caregiver interviews into a single file for two reasons. First, this approach enables a data user to follow the child's living situation across waves. For example, from the perspective of a child who has changed households, the new caregiver's household income at wave 3 is more pertinent to the child's current situation than is the household income for the caregiver from whom the child is separated. Second, the continuing and new caregiver interview instruments each contain extensive data on the child's well-being and the caregiver-child relationship. The continuing/new caregiver data set includes observations on caregivers from 1,835 households.

### *Features of the new caregiver interview*

The new caregiver interview includes data on the caregiver's demographic characteristics. These data were not obtained again for continuing caregivers. The new caregiver also provides the date the focal child came to live with her/him.

### *Separated caregiver interview*

The separated caregiver interview stands alone, and excludes nearly all information about the child's well-being and the caregiver-child relationship. The exception is that in cases where a separated caregiver has been in contact with the focal child in the last six months, she completes the Child Behavior Checklist, an assessment of behavior problems. (If the focal child is age 19 or over and has been in contact with the separated caregiver, the separated caregiver completes the Adult Behavior Checklist.) These cases may be identified by using the variable SEPCGSEEN on the Separated Caregiver data file.

The separated caregiver provides the date that she and the focal child stopped living together. Combined with information from the new caregiver interview, a data user may assemble a month-by-month record of the child's place of residence between interviews.

The data set includes observations on 221 caregivers. The number of observations is higher compared to wave 2 (N=63) because many of the children in the study have entered young adulthood and are living independently. Therefore, many focal children who are separated from their wave 1 caregivers are not residing with a new caregiver at wave 2. (See Scenario E above.)

### Child interview

Data from the child interview are included on a separate file from the caregiver interview, in contrast to the wave 1 file structure. The child interview includes four new modules at wave 3: Middle Childhood, Romantic Relationships, Adolescent Demographics, and Youth Work and Welfare. It also includes an expanded module on Schooling.

At wave 1, sampled children were between the ages of 0 and 4 or 10 and 14. At wave 3, children are between the ages of 5 and 11 or 15 and 21.

### Phone interview

Telephone interviews were administered to caregivers and/or children who had moved more than 100 miles away from the home where they were interviewed at wave 1. The telephone instrument was designed as a 45-minute interview, and is not as detailed as the in-person interviews. The variable CAPIMODE indicates whether the respondent participated in a telephone or in-person interview:

Table 2. Distribution of CAPIMODE

CAPIMODE	Continuing/new caregivers	Separated caregivers	Focal children
1=In-person interview	1773	207	1880
2=Telephone interview	62	14	63
TOTAL	1835	221	1943 (missing for 1 case)

The codebooks note those modules that were not administered to respondents who participated in telephone interviews. Within modules, some individual items were not administered. These skips have not been systematically documented in the codebooks, and users should use CAPIMODE to determine whether telephone respondents were excluded from the universe for a particular item.

### Omitted cases

The public release version of the wave 1 data included 2,402 focal children and their caregivers. This file was edited to exclude 56 cases for whom the principal investigators determined data had been falsified. At wave 2, 45 of those households were successfully recontacted. Those 45 cases in our sample were considered eligible for inclusion at wave 3. The 11 cases not recontacted at wave 2 were dropped from the eligible sample because the study would not have included only cross-sectional data on those households at wave 3, and they would have had no utility in a longitudinal analysis.

One note about the cases that were re-introduced at wave 2: At both wave 2 and wave 3, time-invariant demographic data were not collected from continuing caregivers. Instead, data were carried forward to the wave 2 and wave 3 files from wave 1. As a result, most data are missing on the wave 3 file for all cases omitted cases from wave 1 where the respondent is interviewed as a continuing caregiver. However, data on race and ethnicity from the household screener were used to impute values on the race and ethnicity variables for continuing caregivers and focal children at wave 3. Flag variables indicate whether race/ethnicity data were imputed from the household screener file. See the recode variables at the end of the Demographics module in the caregiver codebooks for information on the flag variables and imputations.

In the course of contacting families to re-interview at wave 3, it was discovered that 9 cases had been duplicated at earlier waves; that is, in 9 cases, the same household was interviewed twice and recorded under two different household identifiers. A system was developed for deciding which of the duplicate cases to drop from the sample in each instance. The HHID's shown in Table 3 were removed at wave 3:

Table 3: HHID for dropped duplicate cases

1020060
0950370
1180350
1220300
1540090
1690550
1740110
1920060
1920090

In sum, 2,402 focal children were included in the wave 1 public release. An additional 45 omitted focal children were successfully recontacted at wave 2. Nine focal children were identified as duplicates. Summing those numbers (2402+45-9), the eligible sample at wave 3 included 2,438 focal children and their caregivers.

## Retention rate

We report retention rates for focal children. The *overall wave 3 retention rate* is calculated as the percentage of eligible focal children from wave 1 who provided partial or complete interviews at wave 3:

$$\text{Wave 1 to Wave 3 retention rate} = \frac{\# \text{ of wave 3 focal children}}{\# \text{ of wave 1 focal children}}$$

We use N=2,438 as the denominator.

The wave 2 to wave 3 retention rate is calculated as the percentage of eligible respondents interviewed at wave 2 who were interviewed at wave 3.

$$\text{Wave 2 to Wave 3 retention rate} = \frac{\# \text{ of wave 3 respondents}}{\# \text{ of wave 2 respondents}}$$

For children, the wave 1 to wave 3 retention rate is 80%. The Three-City Study treats this as the official wave 3 response rate.

The wave 2 to wave 3 retention rate for focal children is 84%, meaning that 84% of children interviewed at wave 2 also participated at wave 3. Fifty percent of focal children who were interviewed at wave 1 and who did not participate at wave 2 did participate at wave 3.

For children, the retention rate from wave 1 to wave 2 was 87.8 percent. The response rate at wave 1 (the number of eligible respondents who gave a partial or complete interview) was 74.7 percent.

## **Contents of the wave 2 instrument**

### *Adult Portion*

From adults, conventional measures of income, poverty, and family and labor force behavior were gathered B data that are generally useful for studies of disadvantaged populations. Specific questions address household structure, marriage, fertility, cohabitation, education, job history and characteristics, hours of work, earnings and wage rates, and sources of income. We also collect information on current and past welfare program participation, as well as participation in other programs such as Food Stamps, SSI, and so forth. In addition, we focus particular attention on the time respondents spend in welfare-related activities and information on actions related to job search. Wave 3 also includes a complete marriage and cohabitation history, a record of live births, and a new module to collect data on female caregivers' views of men, marriage, and nonmarital childbearing.

### *Child Portion*

The survey instrument focuses on four main areas of child well-being: behavioral, cognitive, socio-emotional, and physical development. To assess these domains, the instrument combines

measures used in large national studies with more detailed, process-oriented information on family functioning and child development using comprehensive measures. The questionnaire is designed to address environments and situations that pertain specifically to children in certain age groups while at the same time attempting to use similar measures across age groups in order to increase the longitudinal and cross-sectional comparability of findings. Throughout the child portion of the instrument, we have chosen measures that have proven validity and reliability in low-income and minority populations.

The survey instrument is composed of two computer-assisted personal interviews (CAPI). The first is a 100-minute interview conducted with the primary caregiver of the focal child. The second consists of standardized assessments of the child and a 30-minute interview, conducted if he or she is in the 10-16-year-old age group. Interview questions are organized into modules, each focusing on a different topic related to the lives of the children and caregivers in our study. The modules are listed in Table 3.

Items included in the available survey data fall into three categories: original items, recoded items, and constructed items. *Original items* are those asked of the respondent at the time of interview. *Recoded variables* are those that include changes to values in an original item. These changes are usually based on information from other items in the data set. One example from the welfare module is variable RDE17A31, a recoded version of survey item RDE17A, in which original values were changed based on additional information from survey item RDE17O. *Constructed variables* combine information from several variables into a single item. In the Three-City Study data set, welfare receipt status (WELFST31) is one such item.

Table 4. Modules in the survey instrument

**Caregiver interview**

Demographics	Time Use*
Education and Training	Schooling*
Labor Force, Employment, and Work History	Father Involvement* <sup>+</sup>
Self-Esteem/Self-Concept <sup>+</sup>	Child Support* <sup>+</sup>
Networks <sup>+</sup>	Financial Strain Index <sup>+</sup>
Housing	Welfare Participation and Experiences
Neighborhoods <sup>+</sup>	Income
Family Routines Inventory* <sup>+</sup>	Health and Disability
Home Environment* <sup>+</sup>	Illegal Activities** <sup>+</sup>
Child Behavior Checklist* <sup>+</sup>	Domestic Violence** <sup>+</sup>
Challenges to Parenting* <sup>+</sup>	Brief Symptom Inventory
Parenting Style* <sup>+</sup>	Marriage and Relationships <sup>+</sup>

\* Module excluded from separated caregiver interviews

\*\* Administered by audio computer-assisted self interview (ACASI)

+ Module excluded from telephone interviews

**Focal child interview**

Physical measurements<sup>+</sup>

Woodcock-Johnson<sup>+</sup> (ages 4-21)

Letter-Word Identification

Applied Problems

Middle Childhood (ages 5-12)

Adolescent Demographics (children living independently)

Schooling (ages 14-21)

Peer Association (ages 14-21)

Child-Mother Relationship Scale\*  
(ages 14-21)

Mother-Child Activities\* (ages 14-21)

Parental Monitoring\* (ages 14-21)

Father Involvement\* (ages 14-21)

Father-Child Relationship Scale\*  
(ages 14-21)

Romantic Relationships\* (ages 14-21)

Delinquency Scale\* (ages 14-21)

Sex and Pregnancy\* (ages 14-21)

Brief Symptom Inventory\* (ages 14-21)

Youth Work and Welfare (ages 14-21)

\*Administered by audio computer-assisted self-administered interview (ACASI).

+Modules excluded from telephone interviews

## Weighting

Because this survey is based on a stratified sample, we recommend that users employ weights in their statistical analyses. For the Wave 3 sample, adjusted, trimmed and equalized weights are constructed. The weights are *trimmed* at the top and bottom 5<sup>th</sup> percentiles to limit the influence of outliers, and the weights are “equalized” so that each city contributes equally to an analysis (see below). (Previous user’s guides have referred to the equalized weights as normalized weights. The two terms describe the same procedure of scaling the weights to give each city equal weight and to give the equalized weight a mean of 1.)

Weights are included on the wave 3 data files for the 1,944 children who were interviewed (a *focal child weight*), and similar weights are constructed for (a) the longitudinal sample of 1,763 children who were successfully interviewed in all three waves (a *longitudinal focal child weight*), (b) the sample of 2,038 cases where anyone associated with the Wave 1 household was interviewed in Wave 3 (a *dwelling unit weight*), and (c) the longitudinal sample of 1,873 cases where anyone associated with the Wave 1 household was interviewed in all three waves (a *longitudinal dwelling unit weight*). The descriptions for the weights appear below.

The dwelling unit weights adjust household responses to account for the following factors:

**Stratification:** There are two levels of stratification in the sample: at the neighborhood-level, and at the individual-level. As noted above, we did not carry out a simple random sample, in which every household in a city would have had an equal probability of being selected. Rather, block groups were ranked according to percent poor for each city-specific racial/ethnic group; then a subset of block groups with more than a minimum percentage of poor residents was randomly selected. At the household-level, a screening interview was conducted to see whether a household fell within one of the cells of the design matrix. Then households in specific cells were sampled at different rates in order to obtain a diverse sample. The weights adjust for the probability that a household in a given cell was selected.

**Non-response:** The wave 3 weights were adjusted for the probability that a wave 1 respondent participated in the wave 3 interview. Initial analyses indicate that attrition bias is modest. No statistically significant differences were found between the two groups on the following variables (all measured as of wave 1): age, race/ethnicity, educational attainment, currently on welfare, ever on welfare, and marital status. Significant but substantively modest differences were found on focal child’s sex (the percentage female increased from 49.5 to 50.8 percent), child’s age (the percentage from the older half decreased from 48.2 percent to 47.8 percent), and city of residence (the percentage in Boston declined from 38.6 percent to 36.4 percent).

**Child selection:** Because one child was selected per household, children in large families were less likely to be chosen than children in small families. The child weight adjusts for the number of age-eligible children in the household. Use of the weights allows the investigator to generalize to all children in the households selected. The child selection factor is incorporated into the focal child weights only; before trimming, the focal child weight is equal to the dwelling unit weight multiplied by the number of age-eligible in the household (see the variable NELGCHLD).

Other things being equal, the weighting procedure assigns larger values to households in cities with higher populations because these households had a smaller likelihood of being selected than did households in cities with lower populations. In this data set, respondents from Chicago are weighted higher, all else equal, than respondents in San Antonio, who are in turn weighted higher than respondents in Boston because Chicago is the largest city and Boston is the smallest. So analyses that use the non-equalized weights will reflect information from Chicago more than information from the other cities, and information from San Antonio more than from Boston. If the investigator wants to report results that are proportional to population size, the non-equalized weights should be used.

Equalized weights: The principal investigators of the Three-City Study felt that the difference in population size between cities was rather arbitrary and that it might be preferable to report results that weight each city equally and which therefore present the average experiences of households in the three cities. They modified the RTI weights to create equalized weights in which the total weights for households in one city equals the total weights in the other cities.

As in wave 1, the equalized weights are applicable only if an analysis includes the entire sample. If a subset is used, that subset could be clustered in some of the cities and not others. And if so, a normalization performed for the whole sample will no longer weight each city's selected households equally. Instead, the non-equalized weights would have to be equalized anew to preserve the equal-cities property. See Appendix A in the wave 1 User's Guide for a description of this procedure and sample syntax for the SAS system. The equalizing procedure can be summarized in general terms as follows:

- Determine the number of cities retained in the subsample and create a variable with a constant value equal to that number of cities.
- Divide the total N in the subsample by the number of cities in the subsample (e.g., if the subsample has 900 people and uses all 3 cities,  $900/3=300$ ). This is the average number of respondents per city. Create another variable that takes a constant equal to this value.
- For each city, sum the weights for all respondents in that city. Divide by average city size. This is the average weight for that city, controlling for sample size. Create a new variable that takes three values (if all three cities are included in the analysis): the average weight in Boston, the average weight in Chicago, and the average weight in San Antonio.
- In a new variable, divide the respondent's individual weight by his/her city-specific average weight. (new weight = respondent's weight / city-specific average weight). The mean of this new variable should be 1. This is the re-equalized weight.

Note that at wave 1, both dwelling unit weights and focal child weights were included on the data file. The child weights were similar to the household weights, but took into account the probability that a child was selected for the sample from among all age-eligible children in the household. At wave 2, only child-specific weights were included. At wave 3, we return to providing both dwelling unit weights and focal child weights. The decision about the appropriate

weight to use depends on the unit of analysis and the outcome under consideration. Where the unit of analysis is the focal child and outcome data are drawn from the child data set, the focal child weights are recommended. Where the unit of analysis is a caregiver or a household over time, the dwelling unit weights are recommended.

If an analysis is cross-sectional, i.e., uses only wave 3 data, it is appropriate to use the cross-sectional weight. It is also appropriate to use the cross-sectional weight in an analysis using data from waves 1 and 3, but not from wave 2. If an analysis uses data from all 3 waves, it is recommended that the researcher use the appropriate longitudinal weight.

Table 4 summarizes the names, descriptions, and locations of the weights included on the wave 3 data files.

Table 4. Wave 3 weights

<u>Variable name</u>	<u>Variable description</u>	<u>Variable location</u>
R3DUT5WT	Wave 3 cross-sectional dwelling unit weight, trimmed at the top and bottom 5 <sup>th</sup> percentiles	Caregiver files
R3DUE5WT	Wave 3 equalized cross-sectional dwelling unit weight, non-equalized weight trimmed at the top and bottom 5 <sup>th</sup> percentiles	Caregiver files
R3LDT5WT	Wave 3 longitudinal dwelling unit weight, trimmed at the top and bottom 5 <sup>th</sup> percentiles	Caregiver files
R3LDE5WT	Wave 3 equalized longitudinal dwelling unit weight, non-equalized weight trimmed at the top and bottom 5 <sup>th</sup> percentiles	Caregiver files
R3CHT5WT	Wave 3 cross-sectional focal child weight, trimmed at the top and bottom 5 <sup>th</sup> percentiles	Caregiver files and focal child file
R3CHE5WT	Wave 3 equalized cross-sectional focal child weight, non-equalized weight trimmed at the top and bottom 5 <sup>th</sup> percentiles	Caregiver files and focal child file
R3LCT5WT	Wave 3 longitudinal focal child weight, trimmed at the top and bottom 5 <sup>th</sup> percentiles	Caregiver files and focal child file
R3LCE5WT	Wave 3 equalized longitudinal focal child weight, non-	Caregiver files and focal child file

	equalized weight trimmed at the top and bottom 5 <sup>th</sup> percentiles	
--	--	--

Some survey analysts may wish to correct standard errors for weighting and clustering using statistical packages such as SUDAAN or STATA. After considerable analyses, the principal investigators concluded that adjustments for clustering had little effect on the size of the estimated standard errors. However, should you wish to adjust for clustering, you may use the variables and procedure described in the appendix.

### Unique Identifiers

**HHID:** Each household is assigned a household number that remains constant across waves and across interviews. This number appears as the household identifier HHID, a character variable. Use this variable to match households at wave 1 and wave 2, or to match respondents from the same household who appear on different data files (for example, caregiver and focal child at wave 2).

**ZRID:** Within each dataset, each observation is assigned a unique identifier that indicates from which household the interview is drawn, and in which interview the respondent is participating. ZRID is nearly identical to HHID, but it is one character longer, and the leading character indicates the interview type. Interview types are categorized as:

R: Continuing or new caregiver interview

S: Separated caregiver interview

A: Focal child interview

An example:

Table 5. Unique wave-and interview-specific identifiers

HHID	ZRID - wave 1 interview	ZRID - wave 3 continuing caregiver interview	ZRID - focal child interview
0180010	10180010	R0180010	A0180010

**NEWID:** A household may have multiple associated caregivers over time. For example, a child may reside with her biological mother at wave 1, with her grandmother at wave 2, and again with her biological mother at wave 3. In order to facilitate tracking caregivers across waves, NEWID was created. Each caregiver at each wave is assigned a value on NEWID. NEWID takes the value HHID+01 for the wave 1 caregiver. Where a new caregiver is introduced, his/her values of NEWID is HHID+02. (After data cleaning, no cases were found where a child resided with one new caregiver at wave 2 and a different new caregiver at wave 3. Had that been the case, NEWID would equal HHID+03 for the second new caregiver.) Data users may match

caregivers across waves by merging caregiver data files by NEWID. (If NEWID does not appear on the wave 1 or wave 2 caregiver file, please contact the data archivist.)

The example below shows how NEWID works for a case where a child had a new caregiver at wave 2 and was reunited with her wave 1 caregiver at wave 3:

Table 6. An example of NEWID

HHID=0180010	Wave 1	Wave 2	Wave 3
Child's primary caregiver	Child's mother: 018001001	Child's grandmother: 018001002 (new CG)	Child's mother: 018001001 (continuing CG, reunited with child)
Separated caregiver		Child's mother: 018001001 (separated CG)	

## APPENDIX TO USER'S GUIDE

Some survey analysts may wish to correct standard errors for weighting and clustering using statistical packages such as SUDAAN or Stata. This appendix proposes one method to prepare the data for such an analysis.

We have included the following variables on each data file (continuing/new caregiver, separated caregiver, and focal child):

SCRID: Screener ID

PU: Primary frame unit

SEGID: Segment identification number

SITE: City

1=Boston

2=Chicago

3=San Antonio

STR: Race/ethnicity stratum

B=Non-Hispanic Black/African-American

W=Non-Hispanic White

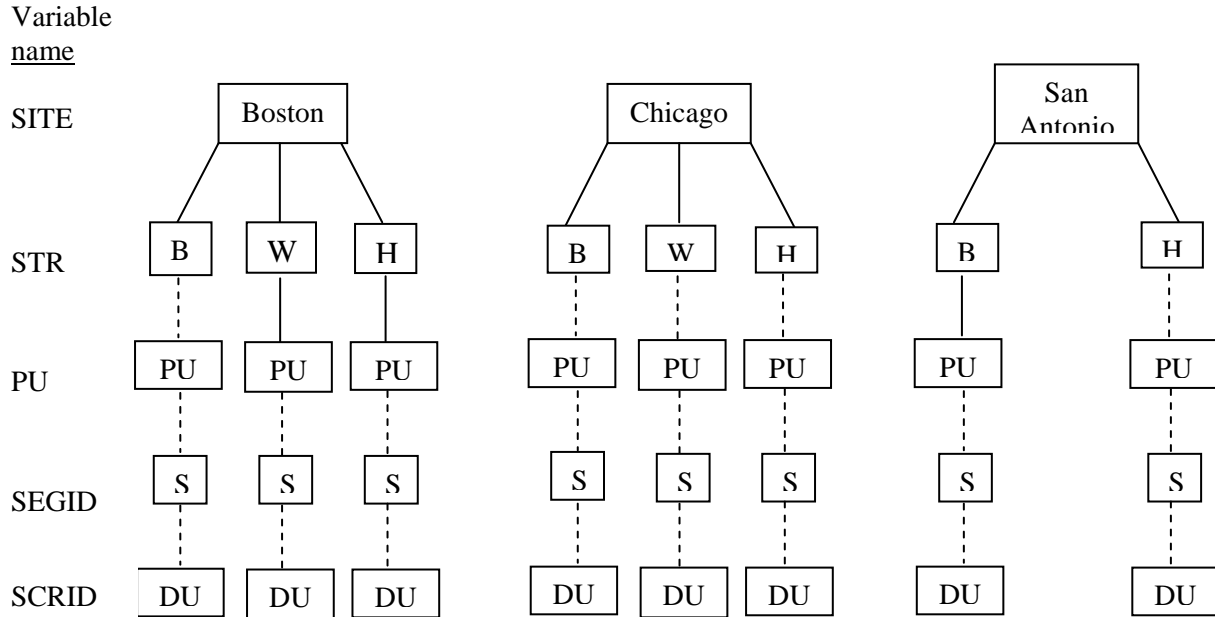
H=Hispanic/Latino

The firm conducting the survey, Research Triangle Institute (RTI), constructed eight sets of block groups from all of the blocks in the three cities in the 1990 Census. Each set ranked all the block groups in a city in descending order of the poverty rate of children in a particular race-ethnic group. In Boston, three such sets were compiled and ranked --one each for Non-Hispanic Whites, Non-Hispanic Blacks, and Hispanics. Three analogous sets were compiled for Chicago, and two sets for San Antonio--Non-Hispanic Blacks and Hispanics. Only block groups falling below a specified poverty level were retained in the sampling frame. The variable SITE refers to the city from which the block groups were drawn. The variable STR refers to the racial/ethnic composition of the block group.

The block groups in the sampling frame are referred to as primary frame units and are represented in the variable PU. In five of the eight sets described above, a random set of the PUs was selected with probability proportional to size. These five sets are referred to as "non-certainty strata," meaning that not all of the eligible PUs were selected into the sample. In the other three sets, all of the PUs were selected. These are certainty strata. Within these strata, all values of the variable PU are blank.

From each stratum, the selected PUs were divided into segments, which are areas of a size typically regarded as convenient for surveying and generally consist of 90-120 dwelling units. A set of segments was chosen randomly from the selected PUs. All selected segments were then counted and listed (i.e., interviewers visited the segments, counted the housing units, and wrote down the addresses of all occupied dwelling units). Segments are represented in the variable SEGID. A random sample of dwelling units (identified by street addresses) was then selected from within each segment. These dwelling units are the households that appear in our sample. The unique identifier SCRID represents each household.

The schematic below shows how households were selected into the sample. The solid lines represent points at which all eligible units were included. The dashed lines represent points where only a subset of units (whether PU's, segments, or dwelling units) was selected.



The method proposed below to account for clustering was developed by the survey contractor. Different methods are used for households in certainty and non-certainty strata. Data users may wish to use other techniques with which they are familiar. For a general discussion of data preparation for complex survey data analysis, see Eltinge and Sribney (1996), Chapter 16 in Levy and Lemeshow (1999), and StataCorp (2003).

For certainty strata (White and Hispanic strata in Boston (site=1, str= "W" or "H") and the Black stratum in San Antonio (site=3, str= "B")):

Sort all households by SITE, STR, SEGID, SCRID. This sorts the data into a geographically ordered list, with households ordered by segment within a site/race-ethnic group. The variable PU takes no value for these cases. Go down the list and form pairs of dwelling units in a new variable called STRATA. Number these pairs from 1 to N/2. Then call each of the dwelling units within a stratum a CLUSTER. The clusters will take on the values of 1 or 2.

For example, assume you are working with a sample that includes 1000 cases from the certainty strata. The first two observations will have the value 1 on the new variable STRATA, the next two observations will have the value 2, and so on. The last two cases will have the value 500. Within each pair, the first observation will have the value 1 on the new variable CLUSTER. The second observation will have the value 2. A list of these data would have the following appearance:

SCRID	SITE	STR	PU	SEGID	STRATA	CLUSTER
1201010S	1	H	.	1201	1	1
1201020S	1	H	.	1201	1	2
1201030S	1	H	.	1201	2	1
1201040S	1	H	.	1201	2	2
1202010S	1	H	.	1202	3	1
1202020S	1	H	.	1202	3	2
.						
.						
2561010S	3	B	.	2561	500	1
2561020S	3	B	.	2561	500	2

No pair should cross between two site/race-ethnic groups. If there is an odd number of dwelling units in an area, the last dwelling unit should be assigned a cluster value of 2 and attached to the last pair in its site/race-ethnic group, so that the last “pair” would actually include three observations.

For non-certainty strata:

Sort all households by SITE, STR, SEGID, and PU. Here, definitions of clusters and strata are not based on the dwelling units and dwelling unit pairs within the non-certainty strata. Rather, PUs and PU pairs are used for cluster and strata definition. Again, the pairing of PUs should be done within a common area, so that pairs do not cross area type.

For example, assume you are working with a sample that includes 1000 observations from the non-certainty strata. Among those 1000 observations, 200 PUs are represented. The number of PU pairs that would emerge from this sample would be (# of PUs)/2, or 200/2=100. All of the observations within the first PU in a given pair would carry a value of 1 on the variable CLUSTER. The observations within the second PU would carry a value of 2. A list of these data would have the following appearance:

SCRID	SITE	STR	PU	SEGID	STRATA	CLUSTER
1401010S	1	B	1	1401	1	1
1401020S	1	B	1	1401	1	1
1401030S	1	B	1	1401	1	1
1401040S	1	B	1	1401	1	1
1402010S	1	B	5	1402	1	2
1402020S	1	B	5	1402	1	2
1403010S	1	B	7	1403	2	1
1403020S	1	B	7	1403	2	1
1403030S	1	B	7	1403	2	1
1404010S	1	B	9	1404	2	1
1404020S	1	B	9	1404	2	2
1404030S	1	B	9	1404	2	2

```

.
.
2261010S    3      H    140    2261        100        2
2261020S    3      H    140    2261        100        2

```

In Stata's svyset command, the analyst may set the variable STRATA as the strata identifier variable, and the variable CLUSTER as the PSU (cluster) identifier variable.

### References

Eltinge, J.L. and W.M. Sibney (1996). svy3: Describing survey data: sampling design and missing data. *Stata Technical Bulletin* 31: 23-26. Reprinted in *Stata Technical Bulletin Reprints*, vol. 6, pp. 235-239.

Levy, Paul S. and Stanley Lemeshow (1999). *Sampling of Populations: Methods and Applications*. 3<sup>rd</sup> ed. New York: John Wiley & Sons, Inc.

StataCorp (2003). *Stata Survey Data Reference Manual, Release 8*. College Station, TX: Stata Press.