National Longitudinal Study of Adolescent Health

Strategies to Perform a Design-Based Analysis Using the Add Health Data

❖ ❖ ❖ ❖ ❖

Kim Chantala
Joyce Tabor

Carolina Population Center
University of North Carolina at Chapel Hill

June 1999
Revised August 2010

This research was supported by grant P01-HD31921 from the Eunice Kennedy Shriver National Institute of Child Health and Human Development with cooperative funding from 23 other federal agencies and foundations. Further information may be obtained by contacting addhealth@unc.edu.

Abstract

The Add Health Study is a nationally representative, probability-based survey of adolescents in grades 7 through 12 conducted between 1994 and 1996. The sample design used to collect the data has introduced a complexity to analysis. Failing to account for this complexity may result in biased parameter estimates and incorrect variance estimates. Hence, you must correct for design effects and unequal probability of selection to ensure that your results are nationally representative with unbiased estimates. Specialized, "user-friendly" statistical software is now available for analyzing data from complex surveys. SUDAAN and STATA are two examples of this type of software. Using both SUDAAN and STATA, we show you how to incorporate characteristics of the sample design into an analysis so that your estimates and standard errors are unbiased. We will first present a simplified description of the Add Health sampling process including a description of the sample attributes and data elements needed for correctly analyzing the data when the unit of analysis is either the school or the adolescent. A brief description of statistical software available to analyze survey data is presented followed by "code templates" you can use as a guide in doing your own analysis using SUDAAN or STATA. Next, we present a 7-step process for performing analysis of Add Health data using any software package designed to handle complex surveys. We then conclude with an example using this process with both STATA and SUDAAN. Results are compared with an analysis from SAS to show how ignoring the design effects can lead to misleading conclusions.

Overview

The Add Health data collection was designed as a cluster sample in which the clusters were sampled with unequal probability. While reducing the cost of data collection, this design complicates the statistical analysis because the observations are no longer independent and identically distributed. To analyze the data correctly, you must use special survey software packages specifically designed to handle observations that are not independent and identically distributed. The purpose of this document is to provide a strategy to correctly analyze the Add Health data. To do this, we describe the characteristics and data elements needed by the survey software packages. We conclude by providing examples using two of the survey software packages, SUDAAN and STATA. All tables, figures, and examples were created using the contractual dataset.

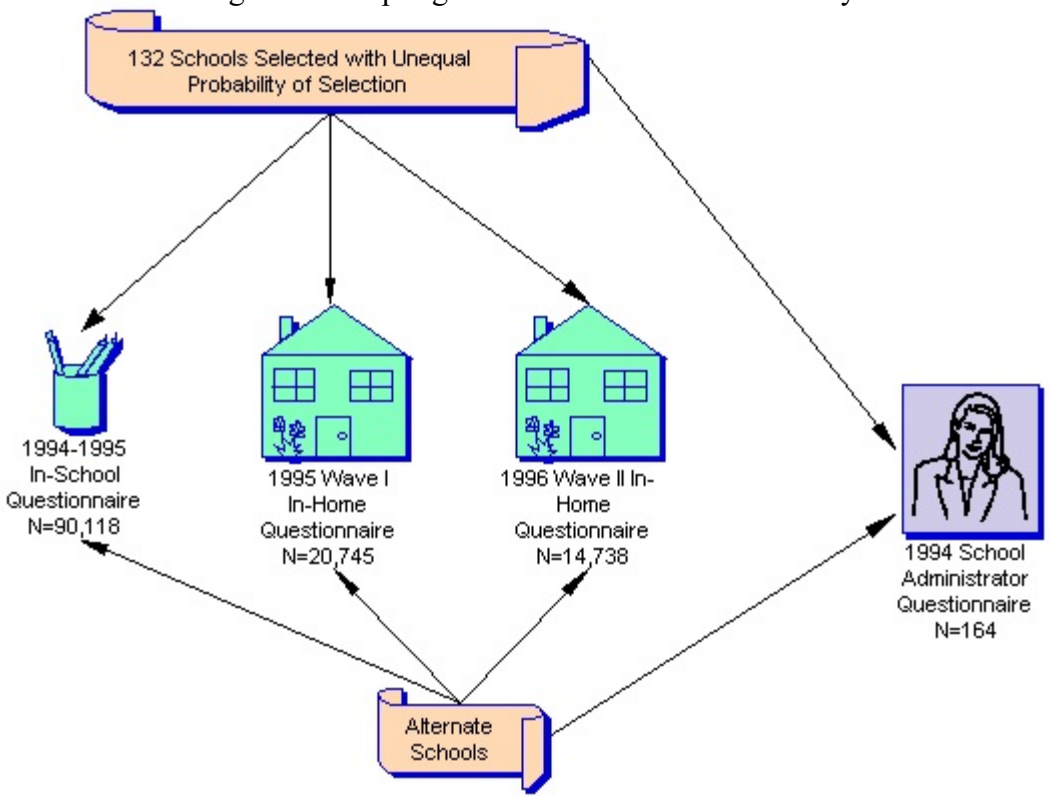
Design Characteristics of the Add Health Data

This section describes how the sampling strategy has influenced the structure of the data. We will focus on why some of the adolescents in our dataset do not have sample weights. The details of the sampling strategy are beyond the scope of this paper, but can be found in the document "Grand Sample Weight" by Roger Tourangeau and Hee-Choon Shin.

Overview of Sample Selection

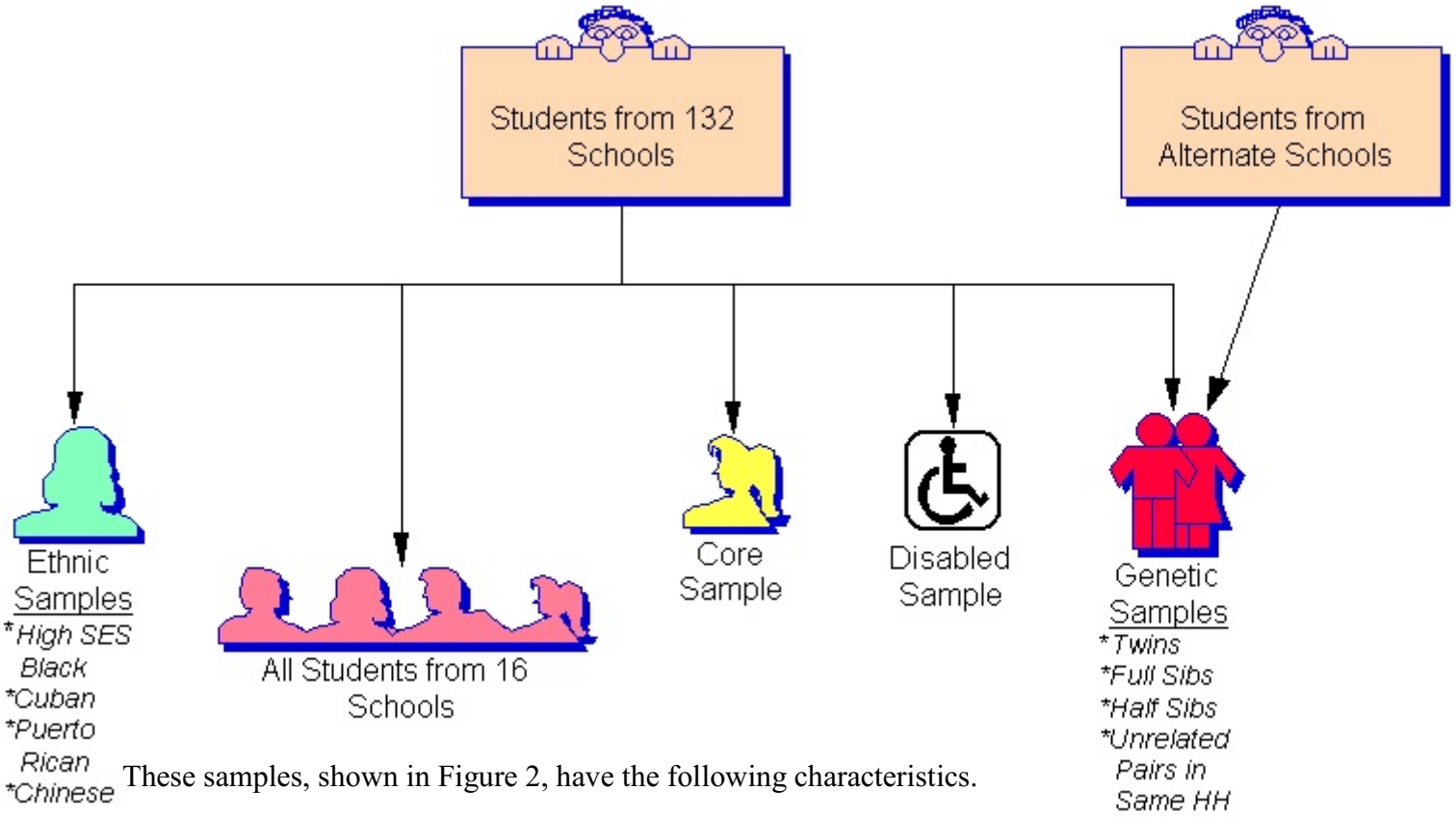
An overview of the Add Health sampling method is shown below.

Figure 1. Sampling Structure for Add Health Study



A sample of 80 high schools and 52 middle (feeder) schools from the U.S. was selected with unequal probability of selection. Thus, school became the cluster identifier or primary sampling unit (PSU). Administrators of these 132 schools were asked to fill out a questionnaire describing the characteristics of these schools. Adolescents attending these schools were eligible for selection into any of the three panels of data: the In-School Questionnaire (1994-1995), the Wave I In-Home Questionnaire (1995), and the Wave II In-Home Questionnaire (1996). Students attending participating schools filled out the In-School Questionnaire. Samples of students from the school rosters and those filling out the In-School Questionnaire were then selected to participate in the in-home data collection phase.

Figure 2. In-Home Questionnaire Target Populations



These samples, shown in Figure 2, have the following characteristics.

- *Core—a nearly self-weighting sample*. Schools were chosen with probability proportional to size and a fixed number of students (~200) were selected from each school. Because the non-response rate varied from school to school, some of this self-weighting property is lost and we consider the core to be a *nearly* self-weighting sample. This is why we needed to develop core weights. Even with a self-weighting property, you must still account for the clustering of the sample. Because of this there is no advantage to analyzing the core instead of the grand sample.

- *Saturation Sample— all students from 16 schools.* Two large schools for adolescent network analysis were chosen; 14 small schools included all students because of the small enrollment size of school.

- *Disabled Sample.* Eligibility for this sample was determined by responses to several

questions on the In-School Questionnaire.

- *Ethnic Samples—High Education Black, Cuban, Puerto Rican, Chinese.* Eligibility was determined by race/ethnicity listed on the In-School Questionnaire.

- *Genetic Samples—identical and fraternal twins, full siblings, half siblings, unrelated adolescent pairs in the same home.* Eligibility was based on responses to the household grid in the In-School Questionnaire.

Availability of Sample Weights

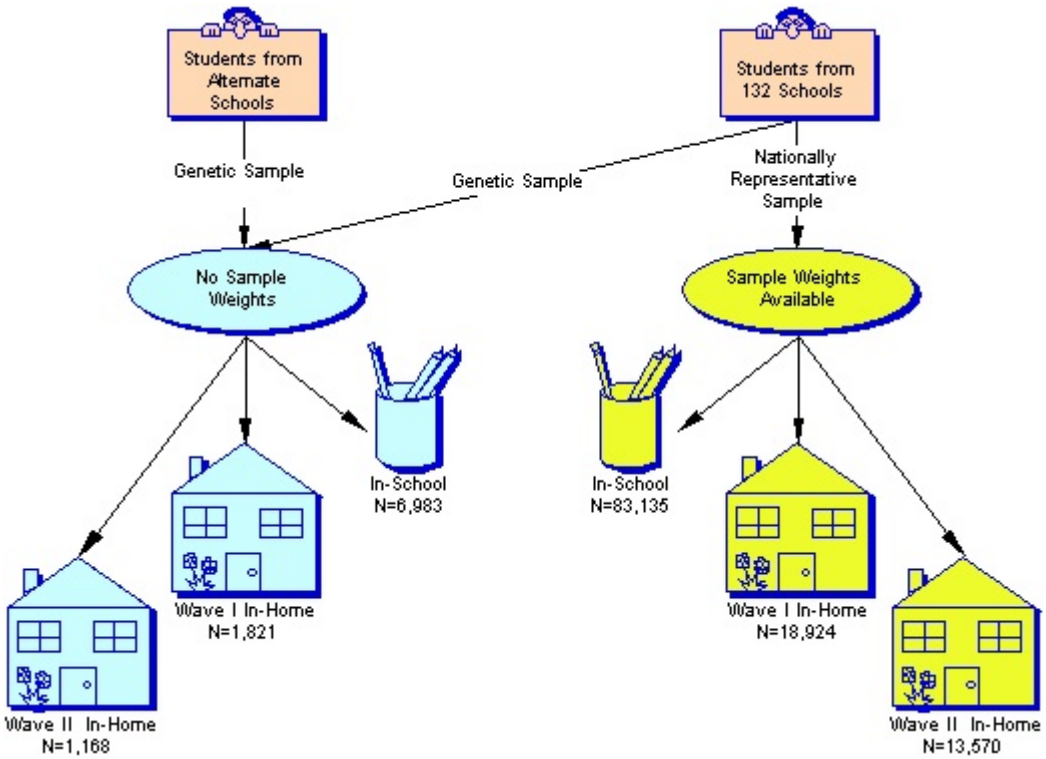
It is important to note that the adolescents in the Add Health Study were selected for two different analytical purposes:

- Analyses to provide nationally representative estimates

- Specialized genetic analyses

Figure 3 illustrates these groups for the three panels of data.

Figure 3. Genetic Sample and the Nationally Representative Sample



The most striking feature of this data schematic is to note that only the adolescents selected to be in the group that can be used to make nationally representative estimates have sample weights. Because the sample size was too small for genetic analyses using only this group, we had to augment the genetic sample with students who were not part of the sampling plan. Thus, weights

could not always be computed for adolescents that were selected for the specialized genetic analysis.

A Note on Weights Needed for Analyzing Pairs of Respondents

Some of the analyses you might be interested in involve serendipitous pairs of respondents. This might include friends as well as twins or siblings. For example, the Add Health data includes respondents and their friends who both filled out the survey. You might be predicting an outcome using information from both the respondent's and their friend's surveys. There is no simple answer to the proper weight to use when your analysis includes observations that are based on data from two different respondents. To correctly compute the weight for each pair, we need to compute the joint inclusion probability of each pair and then the weight for the pair is the inverse of that joint inclusion probability. To compute this weight, we must go back to the details of the sample selection process for both of the individuals and their schools. This can vary for each type of pair (pairs of friends, siblings, twins, and romantic partners) so the method of computing the weight for the pair might be different for each type of pair. We are currently working on this problem and will make the pairs weights available as soon as we are confident of the proper method needed to compute them.

Specifying the Design Structure of the Add Health Data

Next we will discuss the information that must be known about the design to use the survey software. This information is listed in Table 1.

Design Type: Specify With Replacement as the Design Type

The information needed to make finite population corrections for analyzing the dataset as a "without replacement design" is not available. However, we can assume that the schools were selected with replacement. The variance estimation technique is derived using large sample theory and will justify our assumption of with replacement sampling even though schools were not placed back on the list before the next school was selected.

Stratum Variable: Use REGION

The Add Health sampling plan did not include a stratification variable. However, a post-stratification adjustment was made to the sample weights so that region of country (variable REGION) could be used as a post-stratification variable. This involved using the total number of schools on the sampling frame for each region (Northeast, Midwest, South, and West) of the country. For each region, an adjustment was made to the initial school weights so that the sum of the school weights was equal to the total number of schools on the sampling frame.

Cluster Variable or Primary Sampling Unit (PSU): Use the School Identifier

This is the variable named PSUSCID for the In-School, Wave I, and Wave II data. The sampling units in the Add Health Study are middle and high schools from the United States, hence the School Identifier is the appropriate variable to use as the cluster or PSU variable.

Table 1. Variables for Correcting for Design Effects in Contractural Dataset

	Design Type = With Replacement			
	Unit = School	Unit = Adolescent		
	School Admin N = 164	In-School N = 90,118*	Wave I N = 20,745*	Wave II N = 14,738*
Strata variable	REGION	REGION	REGION	REGION
Cluster variable	PSUSCID	PSUSCID	PSUSCID	PSUSCID
Weight variable	SCHADMWT	SCHWGTPS	GSWGT1	GSWGT2
# with weights	130	83,135	18,924	13,570
# missing weights	34	6,983	1,821	1,168
Mean of weights	250.4650	269.9790	1173.8392	1387.4386
Sum of weights	32560.4510	22444705.540	22213733.941	18827541.830
Minimum weight value	37.4121	78.3522	16.3183	18.1956
Maximum weight value	4346.8648	6660.9359	6649.3618	8246.0876

* These numbers are based on individual datasets, not combined datasets.

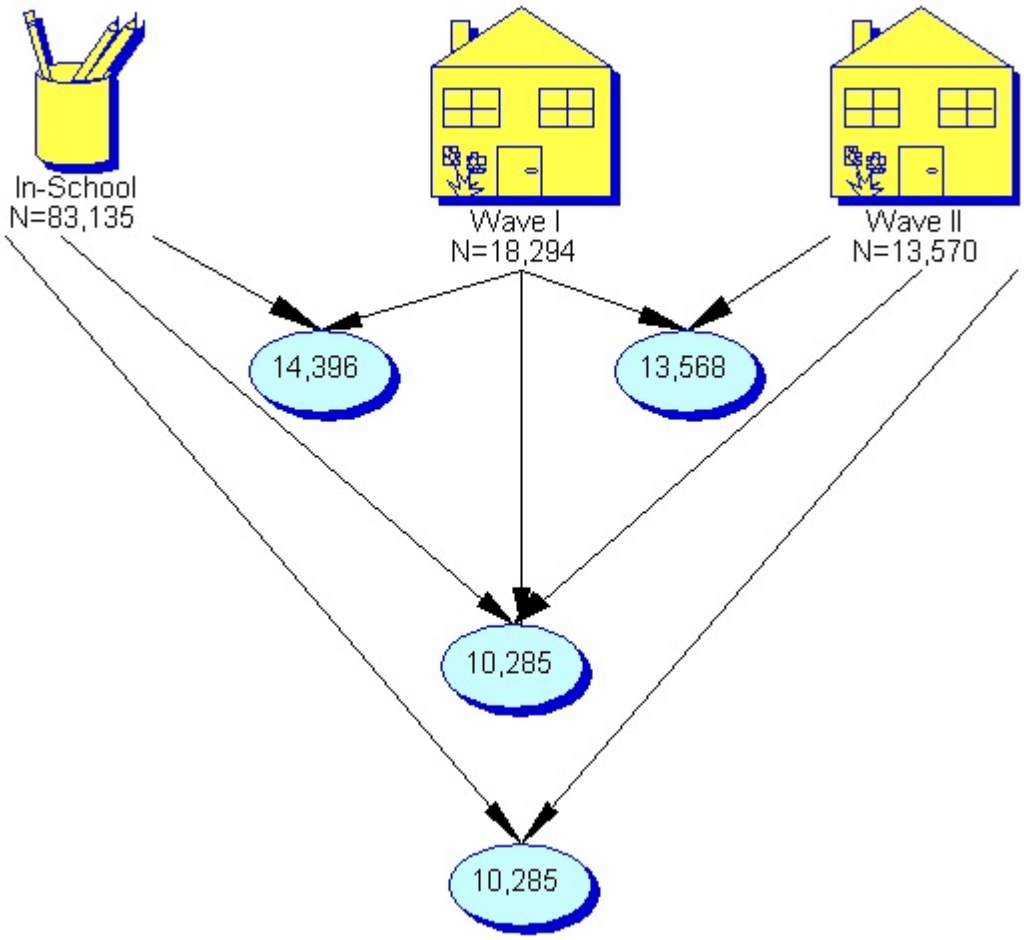
Weight Variables: Use Grand Sample Weight Variables

These variables are SCHWGTPS for the in-school data, GSWGT1 for the Wave I data, and GSWGT2 for the Wave II data. *It is important that you eliminate adolescents with missing values of weight from your analysis.* This is so that any tabulation ignoring weights you do will be based on the same observations you will use in your weighted analysis. The analysis ignoring weights might be part of a preliminary examination to make sure there are a sufficient number of observations for the proposed analysis to be valid.

Samples Using Data from Multiple Panels

Your analysis might involve variables from more than one panel of data. *You should pick the sample design characteristics from the panel that was most recently collected.* For example, you might want to combine data from the Wave II In-Home Questionnaire and the In-School Questionnaire. You would use the sample weight (GSWGT2) from Wave II. Figure 4 shows the sample size of sub-populations with sample weights defined by adolescents in different combinations of the panels.

Figure 4. Number of Adolescents in Multiple Datasets



For this example, your dataset created from combining the in-school data with the Wave II data would have N=13,570 adolescents, and your sub-population of interest would have N=10,285 adolescents. See the important note in the next section for special instructions on analyzing sub-populations with the survey-software.

Current Recommendations on Correcting for Design Effects

We recommend that you use a survey software package to analyze the Add Health data. These software packages have been developed to implement the correct formulas for estimating variances when analyzing complex survey data. SUDAAN and STATA are two examples of these software packages. As a motivation to use survey software, we would like to discuss the pitfalls of using the more traditional techniques to analyze Add Health data.

The design characteristics we just discussed influence your results in very specific ways:

- Point estimates (means, regression parameters, proportions, etc.) are affected by only the weights
- Variance estimates are affected by the clustering, stratification, weight and design type

SUDAAN and STATA incorporate the survey design characteristics into their computational formulas and are considered to give the correct results. Other sample survey software packages you may choose to use would give identical or very close results.

Two methods have been used to correct for design effects: model-based and design-based. The basic model fit by these methods will look something like:

```
OUTCOME = COVARIATES + DESIGN VARIABLES + ERROR TERMS
```

COVARIATES are the characteristics of the adolescent that the analyst believes affect the outcome. The DESIGN VARIABLES and ERROR TERMS describe the clustering, stratification, and weight variables as well as the correlation structure of the data. When using model-based methods, the analyst must determine if and how to account for the DESIGN VARIABLES and ERROR TERMS. In the design-based approach, the DESIGN VARIABLES and ERROR TERMS (including the correlation structure of the data) are dictated by the sampling design and are automatically incorporated by the survey software packages.

Table 2. Comparison of Techniques Used to Analyze Survey Data

	Ignore Design Structure			Incorporate Design Structure	
	Model-Based Analysis			Model-Based Analysis	Design-Based Analysis
Effect on	Ignore Weights	Use Weights	Use Normalized Weights	Use Weights, Strata, Cluster	Use Weights, Strata, Cluster
Estimates of totals	Incorrect	Correct	Incorrect	Correct	Correct
Estimates of ratios, such as proportions, means, regression parameters	Incorrect	Correct	Correct	Correct	Correct
Estimates of variances, standard errors, confidence intervals	Incorrect	Incorrect	Incorrect	Close to Correct	Correct

In Table 2 we have classified analysis techniques into five different approaches. Ignoring both weights and the design structure produces incorrect point estimates and variances. But including the weights in an analysis in which the design structure is ignored gives correct point estimates (totals and ratios). This means that if you only need point estimates and your standard software package allows you to use weights, there is no need to use the survey software packages. Notice that using normalized weights produces incorrect estimates of the totals such as the total number of adolescents or total consumption of adolescents in the population.

The last two techniques incorporate the design structure. The *model-based analysis* can be implemented in a traditional analysis package like SAS while the *design-based analysis* is implemented with a survey package like SUDAAN. It can be very difficult and time consuming to produce acceptable results with the model-based methods. You must decide if and how to incorporate detailed characteristics of the sampling plan, weighting scheme, and intracluster correlation, as well as understand the formulas used by the standard package and the adjustments that might need to be made to these formulas. The survey software package frees you from this and let you easily compute correct results.

Software for Design-Based Survey Analysis

While there are many survey software packages available, we will limit our discussion to SUDAAN and STATA. A list of websites and review articles for other packages is included in Appendix A and the Reference section of this paper. We encourage you to investigate these other packages as well as STATA and SUDAAN.

SUDAAN vs. STATA

SUDAAN does not have any data management capabilities and requires you to do all of your data manipulation in another package before doing your analysis. There is a version of SUDAAN that is SAS-callable. STATA, on the other hand, is an integrated package that offers data management capabilities, traditional model-based and design-based analysis capabilities. There is a rich source of design-based analytical techniques available from SUDAAN and STATA. Many of these are listed in Table 3.

Table 3. Partial Listing of Capabilities of STATA and SUDAAN

Analytical Technique	SUDAAN Procedure	STATA Command
Means, totals, proportions, standard errors	DESCRIPT	symean
Ratios	RATIO	svyratio
Quantiles	DESCRIPT	
Contingency table analysis	CROSSTAB	svytab
Linear regression	REGRESS	svyreg
Logistic regression	LOGISTIC	svylogit
Polytomous logistic regression	MULTILOG	svymlog
Proportional hazards models	SURVIVAL	
Log-linear models of contingency tables	CATAN	
Probit modeling		svyprob
Tobit & interval regression		svytobit
Subpopulation analysis	SUBPOPN statement	subpop option

Contact SUDAAN or STATA to determine the availability of techniques not listed in table 3 . If your technique is not available, then we recommend you analyze the grand sample by incorporating the design structure and weights into your chosen methods.

IMPORTANT NOTE: Analyzing Subpopulations

You may want to analyze just a subpopulation of the Add Health data. For example, you might be interested in restricting your analysis to adolescents from rural areas. *If you use only data for this subpopulation, the correct point estimates will be produced, but the standard errors may be computed incorrectly because the survey design structure has been compromised.* This is because the software needs to be able to identify all PSUs to correctly compute a variance estimate. For example, if a stratum (from the REGION stratification variable) has 20 PSUs and 10 are lost because of restricting the sample to a subset, than the analysis software used to correct for design effects will use an incorrect formula to compute contributions to the variance. Only by running the full dataset can SUDAAN or STATA correctly handle the 10 empty PSUs. SUDAAN provides a SUBPOPN statement for all procedures to define the subpopulation of interest; STATA provides a SUBPOP option or BY statement for you to handle this.

The size of difference in the two variance estimates from analyzing the full dataset with the subpopulation option (SUBPOPN, SUBPOP) and the subset of the data is hard to predict. If only a few PSUs are missing in each level of the stratification variable (REGION), then your results will probably be nearly the same. Defining subpopulations by aggregates of the stratification variable in general should not need to have the subpopulation options used. For example, if you want to analyze all adolescents from REGION=1 level of the stratification variable, you will not need to use the subpopulation option. However, we recommend that you always use the subpopulation options to specify your population of interest. Otherwise, you will have to carefully examine the data to make sure that all PSUs are represented in each level of the stratification variable.

Often some of the respondents did not answer the questions that you need to use in your analysis. This means that the parameters will not be estimated from the full sample, so that you are actually analyzing a subset of the data. We recommend you define the adolescents with complete data as your subpopulation. This will be particularly useful when you want to compare results from models that contain different subsets of covariates since you will want the results from all models to be based on the same observations.

Code Templates for SUDAAN and STATA

This section shows you the basic commands you will use in STATA (versions 10 and 11) and SUDAAN to analyze the Add Health data. You will need to replace the part of the command in *italics* with information from Table 1.

Using STATA for Your Analysis

The following STATA commands will associate variables with options or weights for your analysis:

```
.svyset psuscid [pw = wt_var], strata(region)
```

The stratum (REGION), sampling weight (*wt_var*), and primary sampling unit (PSUSCID) information have now been specified to STATA. STATA defaults to a with replacement design

type, so this information is not specified. You would then proceed with one of the STATA commands for survey analysis. For example, to compute mean PVT scores you would then use:

```
.svy: mean ah_pvt
*Include specifications for subpopulation analysis if applicable;
```

Recall that the survey software must be able to account for all primary sampling units in your analysis. This means that you must tell STATA your population of interest. The survey analysis commands in STATA implement this differently. Use the subpop option if it is available for your routine, otherwise use the over statement. For example, to do the above analysis for boys you could create an indicator variable called sex with the value 1 for boys and 0 for girls and then use:

```
.svy: mean ah_pvt, subpop(sex)
```

or use

```
.svy, subpop(sex): mean ah_pvt
```

or use the over statement to calculate the mean PVT scores separately for boys and girls

```
.svy: mean ah_pvt, over(sex)
```

Using SUDAAN for Your Analysis

Your SUDAAN template has the form:

```
PROC whatever data="AH_data" FILETYPE=SAS DESIGN=WR;
NEST REGION PSUSCID;
WEIGHT wt_var;
SUBPOPN  mydata=1;

Add other modeling statements, printing options here;
```

The first statement specifies the appropriate SUDAAN procedure for your analysis, the name (*AH_data*) and type (*SAS* transport) of the data file, and indicates that the appropriate design type is with replacement (WR). You will need to replace *whatever* with a procedure name of your choosing. The second statement (NEST command) specifies the strata variable (REGION) and the primary sampling unit or cluster variable (PSUSCID). Unless otherwise specified, SUDAAN assumes the first variable in this statement is the stratification variable and the second is the primary sampling unit. The fourth statement is used to specify the population of interest for your analysis. The variable *mydata* is an indicator variable with the value 1 for all observations that need to be included in the parameter estimates and 0 for the observations you want omitted.

Steps to Perform a Design-Based Analysis Where the Unit of Analysis is an Adolescent

This section outlines the steps to consider when correcting for design effects in your analysis.

1. Determine the panel(s) of data you need for your analysis.

2. Identify the attributes and elements of the sample design (WR design, strata variable, cluster variable, weight variable) for the most recent panel of data identified step 1.

3. Make sure that the above variables identified in step 2 are identified on each sample record.

4. Delete any of the cases from the Add Health dataset that have missing weights from your analysis dataset. All of the other design information (strata variable and cluster variable) should be non-missing. Make sure you are analyzing the full sample by checking that the number of observations matches the value given in the "# with weights" row of Table 1 for the most recent panel of data you are analyzing.

5. Identify the population you are interested in analyzing and create an appropriate indicator variable to use for specifying the sub-population.

6. Determine the procedure and the set of commands you need to run an appropriate design-based analysis.

7. Run the analysis and interpret your results.

Example Using SUDAAN and STATA Compared to SAS

This section first shows an example of how to perform a valid and reliable analysis using SUDAAN and STATA to correct for design effects. The results are then compared with different ways of analyzing the data in SAS. We conclude with an example showing how subpopulation variance estimates will be incorrect if you omit the SUBPOPN statement in SUDAAN or SUBPOP statement in STATA.

Example Showing Recommended Technique to Correct for Design Effects

The following example shows how to use SUDAAN and STATA to compute the average number of hours of TV watched during a week for adolescents in RURAL areas. For this analysis we will use data from the Wave II In-Home Questionnaire (step 1). From Table 1 we see that the design type is with replacement, REGION is our stratification variable, PSUSCID is the primary sampling unit, and GSWGT2 is the weight variable (step 2). We included these variables on our analysis dataset (step 3) and deleted all observations with missing weights (step 4), leaving a total population of 13,570. There are no missing values for REGION, PSUSCID, or GSWGT2. Next create an indicator variable called RURAL with a value of 1 for kids attending schools in rural areas, and 0 otherwise (step 5). The set of commands (step 6) needed using SUDAAN is:

```
proc descript data="ALLKIDS" filetype=ASCII design=WR;
nest region psuscid;
weight gswgt2;
subpopn rural=1;
```

```
var hr_tv ;
setenv pagesize=40 linesize=60;
title “USE ALLKIDS with SUBPOPN statement”;
```

These commands will produce the output shown in Figure 5.

Figure 5. SUDAAN Output

```
1   PROC DESCRIPT DATA="ALLKIDS" FILETYPE=ASCII DESIGN=WR;
2   NEST REGION PSUSCID;
3   WEIGHT GSWGT2;
4   SUBPOPN RURAL=1;
5   VAR HR_TV ;
6   SETENV PAGESIZE=40 LINESIZE=60;
7   TITLE "USE ALLKIDS with SUBPOPN statement";


Number of observations read    :  13570    Weighted count : 18827542
Number of observations skipped :      0
(WEIGHT variable nonpositive)
Observations in subpopulation  :   2366    Weighted count:  3009476
Denominator degrees of freedom :    128

For Subpopulation: RURAL = 1
USE ALLKIDS with SUBPOPN statement
by: Variable, One.

-------------------------------------------------
|                 |                  |               |
| Variable        |                  | One           |
|                 |                  | 1             |
------------------------------------------------- |
|                 |                  |               |
| Hours watch TV | Sample Size      |          2347 |
|                 | Weighted Size    |   2983089.40 |
|                 | Total            |  42303049.75 |
|                 | Mean             |         14.18 |                |
                  | SE Mean          |          1.21 |
-------------------------------------------------
```

In STATA you will need to use:

```
.svyset psuscid [pw = gswgt2], strata(region)
.svy: mean hr_tv, subpop(rural)
```

These commands will produce the output shown in Figure 6.

Figure 6. STATA Output

```
. svyset psuscid [pw = gswgt2], strata(region)
. svy: mean hr_tv, subpop(rural)

Survey mean estimation

pweight:  gswgt2                                  Number of obs    =     13519
Strata:   region                                  Number of strata =         4
PSU:      psuscid                                 Number of PSUs   =       132
Subpop.:  rural==1                                Population size  =  18746928
------------------------------------------------------------------------------
    Mean |   Estimate    Std. Err.   [95% Conf. Interval]        Deff
---------+--------------------------------------------------------------------
   hr_tv |   14.18095      1.2106    11.78557    16.57633    14.61483
------------------------------------------------------------------------------
```

SAS Compared to SUDAAN and STATA Analysis

PROC MEANS in SAS was used to compute estimates of totals, means and standard errors for hours watched television. Two different statements are available to specify a variable whose values are used to weight each observation. The variable specified by a WEIGHT statement is used to compute a weighted mean, a weighted variance, and a weighted sum. This can be used to give a variance or standard error that is computed based on the actual sample size rather than a size inflated by the value of the weight variable. The value of the variable specified by a FREQ statement is assumed to represent the number of adolescents each observation in the dataset represents, where n is the truncated value of the FREQ variable. This means that someone with a weight of 100.78 will receive a weight of 100. Values less than one will not be used. This can lead to under representing totals and means and make the FREQ statement a bad choice to use to implement the sample weights.

Two different weights were used. The grand sample weight, GSWGT2, and a normalized weight. The normalized weight was computed by:

```
NORMWT2=GSWGT2*13570/18827542;
```

This is just the grand sample weight (GSWGT2) multiplied by the number of people in the grand sample (13,570) and then divided by the sum of the weights (18827542, the weighted count from the SUDAAN example). The sum of NORMWT2 will be the actual sample size of 13,570.

Table 4 compares the correct results from SUDAAN and STATA to results obtained by using PROC MEANS in SAS. Ignoring both the weights and clustering effects (column 1) gives incorrect totals, means, and standard errors. Using the WEIGHT statement gives very different results from the FREQ statement, so care must be taken when choosing which statement will

provide correct totals and means. The FREQ statement (column 4) truncates the sample weights, causing the total number of Hours Watched TV and weighted sample size to be underestimated. Since the WEIGHT statement computes a variance that incorporates the sample weights to use the actual sample size in the computation, we have the same variance estimate using the GSWGT2 as the NORMWT2. Using GSWST2 with the FREQ statement told SAS we had 2981937 adolescents in our population of interest rather than 2347. The resulting standard error of the mean (0.0085) is severely underestimated. Standard errors computed by other methods are one-forth to one-third the size of the correct value (1.21). Note that the total number of hours watched TV is correctly estimated only when using the WEIGHT statement with GSWGT2.

Table 4. Comparison of SAS to SUDAAN for Different Ways of Estimating the Mean Number of Hours Watched Television an Adolescent Attending a School in Rural Areas

	Incorrect SAS PROC MEANS				
Variable: Hours Watch TV	IGNORE WEIGHTS	Use WEIGHT Statement Specifying GSWGT2	Use WEIGHT Statement Specifying NORMWT2	Use FREQ Statement Specifying GSWGT2	Use FREQ Statement Specifying NORMWT2
Sample Size		2347			
Weighted	2347	2983089.40	2347	2347	2347
Size	N/A	42303049.7	2150.07	2981937	1140
Total	33301.00	5	30490.03	42286535.00	16049.00
Mean	14.19	14.18	14.18	14.18	14.08
SE Mean	0.31	0.30	0.30	0.0085	0.43

	Correct SUDAAN or STATA
Variable: Hours Watch TV	Full Dataset with SUBPOPN Statement or SUBPOP Option
Sample Size	
Weighted	2347
Size	2983089.40
Total	42303049.75
Mean	14.18
SE Mean	1.21

Effect of Omitting the SUBPOPN Statement in SUDAAN

If we rerun our analysis without the SUBPOPN statement in SUDAAN or SUBPOP option in STATA, but limit the analysis dataset to the 2,366 adolescents in RURAL areas, our standard error changes as shown in Table 5. This is because not all of the primary sampling units were present in the analysis and the variance formulas were used incorrectly. Note all other computed

values are identical.

Table 5. Effect of Omitting SUBPOPN Statement in SUDAAN or SUBPOP Option in STATA

Variable: Hours Watch TV	INCORRECT Subset Dataset, Omit SUDAAN SUBPOPN Statement or STATA SUBPOP Option	CORRECT Full Dataset, Use SUDAAN SUBPOPN Statement or STATA SUBPOP Option
Dataset size	2366	13570
SUBPOP size	2366	2366
Sample Size	2347	2347
Weighted size	2983089.40	2983089.40
Total	42303049.75	42303049.75
Mean	14.18	14.18
SE Mean	1.09	1.21

Summary

The two main goals of any analysis using data from a complex survey are to produce

- unbiased estimates of parameters for the entire population as well as subpopulations, and
- unbiased estimates of variance and standard errors.

We have shown that the easiest, quickest, and most reliable way to achieve these two goals when analyzing the Add Health data is to use survey software. If you are only interested in the first goal of obtaining unbiased estimates, then you can investigate using your standard statistical analysis package with an appropriate statement to incorporate the sample weights. To obtain unbiased estimates of variance and standard errors, you must account for clustering and correlation of your data. Our main recommendations are:

- Use a survey software package
- Use the special statements in the survey software package to define the subpopulation you are interested in analyzing
- Do not normalize the weights. If you normalize the weights, estimates of population totals will be incorrect even if you use the survey software
- Specify:
  Design type = With Replacement
  Stratification Variable = Region
  Primary Sampling Unit = School Identifier

Appendix A. Where to Go for Additional Information

Websites

- Add Health: http://www.cpc.unc.edu/addhealth
- SUDAAN: http://www.rti.org/sudaan/
- STATA: http://www.stata.com/
- WESTAT: Information about WesVarPC, a survey package that has been distributed without cost to the user: http://www.westat.com/westat/statistical_software/WesVar/index.cfm

List Servers

- Add Health to interact with other analysts: Send email to listserv@unc.edu and in the body of the message type:
  `subscribe addhealth2 firstname lastname`

References

Brogan, D., D. Daniels, D. Rolka, F. Marsteller, and M. Chattopadhay. 1998. "Software for Sample Survey Data: Misuse of Standard Packages." Pg. 4167-4184 in P. Armitage and T. Colton (Eds.) *Encyclopedia of Biostatistics*, Vol. 5. New York: John Wiley.

Carlson, B.L. 1998. "Software for Sample Survey Data." Pp. 4160-4167 in P. Armitage and T. Colton (Eds.) *Encyclopedia of Biostatistics*, Vol. 5. New York: John Wiley.

Cohen, S.B. 1997. "An Evaluation of Alternative PC-Based Software Packages Developed for the Analysis of Complex Survey Data." *The American Statistician* 51(3):285-292.

Lee, E.S., R.N. Forthofer, and R.J. Lorimer. 1989. *Analyzing Complex Survey Data*. Newbury Park, CA**:** Sage Publications.

Lepkowski, J. and J. Bowles. *Sampling Error Software for Personal Computers*. http://www.fas.harvard.edu/~stats/survey-soft/iass.html.

Levy, P.S. and S. Lemeshow. 1999. *Sampling of Populations Methods and Applications*. New York: John Wiley & Sons.

SAS Institute, Inc. 1990. *SAS Procedures Guide*, Version 6, Third Edition. Cary, NC: SAS Institute.

Shah, B.V., B.G. Barnwell, and G.S. Bieler. 1995. *SUDAAN User's Manual: Release 6.4*. Research Triangle Park, NC: Research Triangle Institute.

STATA Corporation. 1999. *STATA Reference Manual, Release 6*. College Station, TX: STATA Press.

Tourangeau, R. and S. Hee-Choon. 1998. *National Longitudinal Study of Adolescent Health Grand Sample Weight*. Chapel Hill, NC: Carolina Population Center, University of North Carolina at Chapel Hill.

Centers for Disease Control and Prevention. 1995. *Variance Estimation for Person Data Using the NHIS Public Use Person Data Tape, 1995*. http://www.cdc.gov/nchswww/data/pvar.pdf.