2019

Add Health Documentation

Parents (2015 -2017) Parent Study Weighting Specifications

Report prepared by

Paul Biemer

Dan Liao

UNC CAROLINA POPULATION CENTER

CAROLINA POPULATION CENTER | CAROLINA SQUARE - SUITE 210 | 123 WEST FRANKLIN STREET | CHAPEL HILL, NC 27516

Add He alth is 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.

Add Health is a program project directed by Kathleen Mullan Harris and designed by J. Richard Udry, Peter S. Bearman, and Kathleen Mullan Harris at the University of North Carolina at Chapel Hill, and funded 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. Information on how to obtain the Add Health data files is available on the Add Health website (http://www.cpc.unc.edu/addhealth).

Add Health Parent Study Weighting Specifications

1. Introduction

The Parent Study (PS) sample is a probability sample drawn from the Add Health Wave 1 (AHW1) sample. As described in the next section, not all cases in the Add Health sample are eligible for the PS and a set of inclusion criteria was applied to produce the sampling frame for the study. The following lists some important features of the PS sample design.

- Sampling of parents for the study was proportional to size (PPS) using a measure of size (MOS) that attempted to minimize the effect of unequal probability sampling on the variances.
- The sampling unit for the PS was the eligible parent as opposed to the Add Health which use the child as a sampling unit. All eligible children on the frame having the same parent were automatically included in the PS sample if the parent was selected. Thus, the selection probability of an eligible parent was derived based upon the AHW1 selection probabilities of the parent's eligible children.
- Some cases in the Add Health sample have zero (or missing) weights. Because MOSs for the PPS sample are functions of the child weights, zero weights were converted to positive (unity) weights so that these cases could be included in the PS sample with positive probabilities of selection.
- There are three potential units of analysis for the study – the parent, the partner of the parent (if there is one) and the child. However, as shown below, the same set of weights can be used for all three analytic units.
- The sample was divided into two replicates. Replicate 1 consisted of 2,691 cases (parents) and Replicate 2 consisted of 1,114 cases for a total of 3,805 cases (parents). Ultimately all 1,114 cases in Replicate 2 were released to the field.
- A total of 363 eligible cases were not pursued in the data collection either because they were not expected to yield an interview or to reduce costs. These cases will be treated as nonrespondents in the weighting.

2. Target Population, Sampling Unit, and Interviewees

The PS target population consists of parents who satisfy the following criteria:

- Is a biological, adoptive or step parent of an Add Health child
- Is not deceased or incarcerated at the time of sampling.
- Have at least one Add Health child who is also not deceased at the time of sampling.

Using the first criterion only, an estimated 13,585 parents were eligible for the PS having a total of 15,562 eligible children. These parents and their children constituted the sampling frame for PS. Appendix A lists some additional restrictions and steps taken to determine the sampling frame. The criteria involving the deceased or incarcerated could not be applied at sampling but were applied during data collection.
As previously noted, the sampling unit for the survey was the Add Health parent. The parent was interviewed about their own characteristics as well as some characteristics of the bio/step/adoptive child (or children) of the parent that responded in the Add Health, Wave 1 survey. In addition, the current spouse or partner of the parent was also interviewed if he or she currently resided at the Add Health parent's address. Data on the PS sample children was collected by proxy from the Add Health parent. Thus, the PS collected data on multiple persons – the parent, the parent that was present at the Add Health interview (if different from the interview parent), the spouse partner (if available) and the child or children of the parent that are in the Add Health sample.

3. Sample Design

Although the parent is the selection unit, the child is considered as the primary analysis unit for the PS. We will discuss the implications of other analysis units such as the parent or the parent's partner subsequently. Weights for the PS will be derived based upon the selection probabilities of the children (i.e., the Add Health sample members or AHSMs).

An AHSM's selection probability for the PS is the product of two probabilities: (a) the probability that the child is in Add Health sample and (b) the probability that this AHSM is then selected for the PS sample. Probability (a) is well-defined because the Add Health survey selected children with known selection probabilities. Probability (b) is the Add Health parent's (conditional) selection probability because once the parent is selected for the PS, all children that are bio/step/adoptive of the parent are automatically included in the PS sample with certainty. The design of the Add Health sample is such that as many as four children with the same Add Health parent can be included the sample. Note that a parent having multiple children in the Add Health sample has multiple chances on being included in the Add Health sample because each child of that parent would have led to the parent being in the sample. That means that the parent's selection probability is a combination of selection probabilities of that parent's children.

The product of probabilities (a) and (b) for the i^{th} child in the PS target population can be expressed as follows:

$$\begin{aligned}\Pr(i^{th}\text{child in PS sample}) = {} & \Pr(i^{th}\text{child in AH sample}) \\ & \times \Pr(i^{th}\text{child in PS} | i^{th}\text{child in AH sample})\end{aligned} \quad (1)$$

where Pr(i^{th} child in PS sample) denotes the probability that the i^{th} child in the population is in the PS sample, Pr(i^{th} child in AH sample) denotes the probability that the i^{th} child is in the Add Health sample, and on the PS frame, and Pr(i^{th} child in PS|i^{th} child in AH sample) is the conditional probability that the i^{th} child is in the PS sample given that the he/she was selected for the Add Health sample.

We can estimate Pr(i^{th} child in AH sample) by $\omega_{AH,i}^{-1}$ where $\omega_{AH,i}$ is the Add Health grand sample weight for the i^{th} child in the AH sample. This probability is dictated by the Wave 1 sample design and is immutable. The probability Pr(i^{th} child in PS|i^{th} child in AH sample) is equal to Pr(i^{th} child's parent in PS|i^{th} child in AH sample) because the i^{th} child is selected only if the i^{th} child's parent is selected. The probability of including the parent of the i^{th} child in the PS has the following form:

$$\Pr(\text{parent of the } i^{\text{th}} \text{child in PS} | i^{\text{th}} \text{child in AH sample}) = n \times \frac{\text{MOS}_i}{\text{TMOS}} \qquad (2)$$

where n is the target number of parents to be selected for the study, MOS_i is the measure of size assigned to the parent of the i^{th} child and TMOS is the sum of MOS_i over all eligible parents on the PS frame.

It was shown in Biemer (2015), that the MOS_i that minimizes the unequal weighting effects for both children and parent analyses is given by

$$\text{MOS}_i = \sum_{j=1}^{m_i} \alpha_j \omega_{\text{AH},j} \qquad (3)$$

where m_i is the number of eligible children linked to the parent of the i^{th} child, $\alpha_1 = 1$ (if $m_i = 1$), $\alpha_j = \sum_{k \neq j}^{m_i} \omega_{\text{AH},k} / [(m_i - 1) \sum_{k=1}^{m_i} \omega_{\text{AH},k}]$ (if $m_i > 1$), and $\omega_{\text{AH},k}$ is the Add Health Wave 1 Grand Sample Weight for k^{th} child linked to the i^{th} child's parent for $k \neq i$.

3.1 Imputing Weights for Zero-Weight Cases

As previously noted, some of the cases in the Add Health sample have zero weights. In fact, three types of parents are included on the PS frame:

A. Parents for whom *all* the children have a positive weight in the Add Health sample,

B. Parents for whom at least one child has a positive weight and at least one has a zero weight, and

C. Parents for whom *all* their children have a missing (or zero) weight.

For Type A parents, all children have weights and the parent selection probability is well-defined. No imputation is necessary. The total number of parents in this category is 12,965. Thus, only the weights for cases Types B and C parents need to be imputed. For Type B parents, which number 436 parents, the available documentation (see, for example, Tourangeau, et al, 1999) suggests that these children were added with certainty in situations where at least one other child with the same parent was randomly selected for the Add Health. Thus, for the purposes of assigning a weight to these zero weight children, to compute a conditional parent selection probability, the following approach was applied.

Consider a Type B parent and let $m > 1$ denote the number of eligible children associated with this parent. One may interpret the sum of the weights of these m children, denoted by w_{tot}, as the total number of children in the PS target population represented by all m children. Thus, each child in this family represents, on average, w_{tot}/m children. By this logic, and for the purposes of computing the parent conditional selection probability, a weight of w_{tot}/m was assigned to each of the m children.

Finally, for Type C parents (which number 184), none of the child in the family has a positive weight and, further, no random selection mechanism can be associated with these children. In this case, it is reasonable to assume that these children were selected with probability 1 and thus assign a weight of 1 to each child in the family. This is essentially equivalent to assuming that each of these children is self-representing. As a result, the parent will also have a MOS of 1 and thus, the self-representing assumption extends to the parent.

Once the zero weight cases have been imputed in this manner, the process of computing the MOS for each parent in the frame proceeded as described above in equation (3).

3.2 Response Rate Summary

Table 1 provides an overview of the weighted and unweighted response rates for the PS treating the parent as the key respondent. Using these results, the parent response rate is approximately 59%.

Table 1. Parent Study Response Rates

Type	Unweighted Count	Unweighted Percent	Weighted Count[1]	Weighted Percent[1]
All Sample Cases (Parents)	3,805	100.00%	17,867,330	100.00%
Eligibility Unknown	1,509	39.66%	6,919,825	38.73%
Eligible[2] (Known + Estimated)	3,607	94.81%	16,941,804	94.82%
Completed	2,013	55.80%	9,621,259	56.79%
Not Completed	89	2.47%	420,727	2.48%

Estimated Eligible[3]	1,505	41.73%	6,899,818	40.73%
Ineligible[4] (Known + Estimated)	198	5.19%	925,525	5.18%
Deceased	182	92.11%	844,785	91.28%
Duplicate	7	3.54%	31,616	3.42%
No Eligible Parent Relationship	5	2.53%	29,118	3.15%
Estimated Ineligible[5]	4	1.81%	20,007	2.16%

1. The parent weighted counts and percentages are approximate because they are based upon child-level base weights. Parent-level base weights were calculated only for responding parents.
2. The estimated number of eligible cases in the entire PS sample includes the number of the known eligible cases plus an estimated number cases with unknown eligibility that are eligible.
3. The estimated number of cases with unknown eligibility that are eligible is calculated by $e\times$(number of cases with unknown eligibility), where e is the percentage of eligible cases among cases of known eligibility (excluding those who deceased prior to 9/19/2015 or were duplicates in the sample). In this table, e=(2013+89)/(5+2013+89)=99.76%.
4. The estimated number of ineligible cases in the entire PS sample includes the number of the known ineligible cases and the estimated number of ineligibles among cases with unknown eligibility.
5. The estimated number cases with unknown eligibility that are ineligible is calculated by $(1-e)\times$(number of cases with unknown eligibility).

Some cases were either not fielded at all or fielded but then retired early in data collection for various reasons. This includes the following types of cases:

- Very low propensity to respond
- No name, or insufficient name (e.g., missing last name)
- Incorrect PO Box addresses with unknown physical addresses.

Altogether, there are a total of 363 of these cases in sample. Table 2 provides a breakdown of these cases by general type. These cases will be treated as nonrespondents in the weighting so that their characteristics will be represented in the final estimates.

In addition, some cases were believed to be deceased based upon obituary searches or family provided information. However, this type of verification was not always possible and the case was coded as "pending deceased." As an example, a mail return with "deceased" written on the envelope and no other evidence would be coded a "pending deceased." Deceased and pending deceased cases were not fielded.

All deceased status codes will be will be confirmed by matching the sample member's personal identifying information to the NCHS National Death Index (NDI) in a separate operation that will be conducted in early 2018. If an Add Health parent was in fact deceased when the sample was drawn, then the case will be classified as ineligible. However, if a nonresponding sample member was alive at the time the sample was drawn, the case will be classified as a noninterview regardless of deceased status. Likewise, an Add Health parent of a deceased Add Health child will also be classified as either a noninterview or as an ineligible sample member based upon the time of death of the child using a similar same rule.

Finally, as an additional check on eligibility, all nonresponding Add Health parents and their corresponding AHSMs will be match to the NDI to determine if they are living and thus if they are eligible for the PS. Thus, it is conceivable that some nonrespondents may be deemed ineligible based upon their NDI match status.

Table 2. Cases Not Worked or Retired

Type of Case	**Number**
Total Sample Cases	3805
Retired due to low response propensity[1]	*59*
No-name/unknown address cases not worked[2]	*304*
Total cases not worked or only partially worked	*363*

[1]Low propensity cases are defined as cases whose estimated response propensity (based upon a model) was less than or equal to 0.40.

[2]No-name cases are cases that do not have identifying information making tracing and interviewing them impossible.

4. Weighting

One of the unique features of the PS is the possibility of using three types of units of analysis: the child (or AHSM), the parent or the spouse partner. In general, if the dependent variable in an analysis is a AHSM characteristic, then the child weight should be used. When it is a characteristic of the parent, the parent weight should be used. No separate weight has been derived for spouse partners. Rather, the parent weight should be used when the dependent variable is a characteristic of the spouse partner.

Here are some situations where the child-level weight is appropriate:

1. To estimate the proportion of children in the population that have some characteristic, including the proportion of children whose parents have some specified characteristic.
2. To model a child's outcome as a function of the child's characteristics and that child's parent characteristics for the entire Parent Study sample.
3. To repeat the analysis in (1) for a subgroup of the PS child population.
4. To repeat the analysis in (1) or (2) after merging information from the Add Health for all children in the PS and including characteristics of the child from the Add Health as explanatory variables
5. To model an Add Health outcome using PS study characteristics for parents and/or children as explanatory variables and the sample is confined to AHSMs that are in the PS.

The parent weight would be more appropriate for the following analyses:

1. To estimate the proportion of parents that have some characteristic, including the proportion of parents with children that have some specified characteristic.

2. To model a parent's outcome as a function of the parent's characteristics and/or that parent's child's or children's characteristics for the entire Parent Study sample.
3. To repeat the analysis in (1) for a subgroup of the PS parent population.
4. To repeat the analysis in (1) or (2) and include characteristics of the child from the Add Health as explanatory variables after merging information from the Add Health for all children in the PS.

In this section, we discuss the development of the PS child weight and parent weight and describe the additional weighting adjustments to be performed once data collection has been completed.

Using (1), we can write

$$\Pr(i^{\text{th}}\text{child in PS sample}) = n\omega_{AH,i}^{-1}\frac{\text{MOS}_i}{\text{TMOS}} \tag{4}$$

for $n = n_1 + n_2$ where n_1 is the number of parents selected in the first replicate and n_2 is the number of parents in the second replicate[1]. Thus, the Parent Study weight for the i^{th} child, denoted by w_{1i} for child i, is then the inverse of right-hand side of equation (4). Conditionally on the ith child being selected, the parent of ith child is selected with probability 1. However, for an analysis where the parent is the analysis unit, the child weight is not appropriate because some parents have multiple children, each of which could have resulted in the parent's selection for the study. To appropriately adjust for this in the weighting, a parent weight should be used which is developed in Section 4.3.

4.1 Nonresponse Adjustment

The WTADJUST procedure in SUDAAN® (2012) (a generalized exponential model [GEM] module (see Research Triangle Institute 2012) will be used to adjust the selection weight to compensate for nonresponse error. We shall use (unweighted) models for estimating ρ_i, the i^{th} parent's response propensity of the form:

$$\hat{\rho}_i = \frac{\exp\left(\mathbf{x}_i\hat{\boldsymbol{\beta}}\right)/U}{1+\exp\left(\mathbf{x}_i\hat{\boldsymbol{\beta}}\right)} \tag{5}$$

where $\mathbf{x}_i$ is the ith row of the $n \times p$ matrix $\mathbf{X} = [\mathbf{X}_1 \mid \mathbf{X}_2]$, $\mathbf{X}_1$ is an $n \times p_1$ matrix of covariates that are mostly related to the response mechanism, $\mathbf{X}_2$ is an $n \times p_2$ matrix of covariates that are mostly related to the key PS outcome variables, $\boldsymbol{\beta}$ is a p dimensional column vector of coefficients to be estimated by GEM using explanatory variables, $\mathbf{X}$, dependent variable r_i which is a binary indicator variable that is 1 if the ith parent responded and 0 otherwise

[1] As previously noted, n is currently 3,805 sample members with no exclusions. However, this number could be reduced once final eligibility is determined based upon the NDI results and ineligible sample members are excluded.

satisfying $\mathrm{E}(r_i \mid \mathbf{X}) = \rho_i$, $p = p_1 + p_2$ and U is the maximum size of the nonresponse adjustment factor (i.e. $1/\hat{\rho}_i$). Some potential variables for inclusion in $\mathbf{X}_1$ and $\mathbf{X}_2$ are: total number of contacts attempts made to the child at Wave III and IV, number of children of the parent, region and urban-rural setting at the child level in Wave V, and parent's education level (W1PRHGTE), parent's relation (PC1), family structure variable (W1FAMST) and family type (i.e., FAMTYP25). Demographic variables of both the parent and the child such as age, gender, race and ethnicity will also be tested for possible inclusion in the models. In addition, any effects on response propensity because of fielding the sample as two replicate subsamples will be addressed by adding a replicate sample indicator to the models and allowing this indicator to interact with other response propensity predictors.

Appendix B contains some preliminary information on some of these demographic variables. Additional variables will be considered in the weighting analysis. One complicating factor for this approach is that these data are only available for parents and children who responded at Wave I. To address this issue, response propensity models will be fit that are conditional on Wave I response. Variables will be selected for the final response propensity models via a classification/regression tree method.

An equivalent approach to (5) for estimating response propensities is to write the models in the form of calibration estimating equations following RTI International (2016). Then an estimate of ρ_i is obtained by solving the following simultaneous equations for $\hat{\rho}_i$:

$$\sum_{i \in R} \hat{\rho}_i^{-1} \dot{\mathbf{x}}_{1i} = \sum_{i \in S} \dot{\mathbf{x}}_i \tag{6}$$

where R is the subset of parents responding to the Parent Study, $\dot{\mathbf{x}}_i$ is the ith column of $\mathbf{X}'$, the $\hat{\rho}_i$ are selected to minimize $\sum_{i \in R} |1 - \hat{\rho}_i^{-1}|$ and S is the full Parent Study sample.

The nonresponse adjusted weight is thus

$$w_{2i} = w_{1i} / \hat{\rho}_i \tag{7}$$

To control the size of the adjustment factors, $1/\hat{\rho}_i$, the values of $\hat{\rho}_i$ will be sorted and grouped in to deciles of approximately equal numbers of cases. The value of $\hat{\rho}_i$ in each decile will be replaced by the corresponding decile group's response rate. Then this rate will replace $\hat{\rho}_i$ in (7).

4.2 Population Calibration

The nonresponse adjusted child weight can be further improved by weight calibration. Let $\mathbf{z}_i = (z_{i1}, z_{i2}, \ldots, z_{ip})$ denote a p dimensional row vector of auxiliary variables with known (or precisely and unbiasedly estimated) totals, $Z_k, k = 1, ..., p$, respectively. The effectiveness of each available parent or child characteristic will be explored in the weighting process. The following

variables will be tested for inclusion in the calibration equations: the three-way interaction term among age, race, sex, and family structure strata variable (i.e. W1FAMST). Appendix B provides additional information regarding the creation of these variables.

Define

$$Z_k = \sum_{i \in S_I} \omega_{Ii} z_{ik} \tag{8}$$

where ω_{Ii} is the grand sample weight for the i^{th} child on the frame and the sum is over all eligible children on the Parent Study frame (i.e., all children in Wave I after removing deaths, children of parents who did not fit the eligibility criteria and other ineligible children, denoted by S_I). Thus, Z_k is the Wave I estimate of the Parent Study frame population total of z_{ik}. Using WTADJUST in SUDAAN, we seek constants a_i such that the following calibration equations hold

$$\hat{Y} = \sum_{i \in R_C} a_i w_{2i} z_{ik} = Z_k \tag{9}$$

where R_C denotes the set of children for which data were collected from the responding parents and $k = 1,\ldots, p$. The final weight is then defined as

$$w_{\text{FINAL},i} = a_i w_{2i} z_{ik} \tag{10}$$

for $i \in R_C$.

Because Parent Study eligibility is not known for all units on the Add Health Wave I frame, S_I defined in (8) is also unknown and thus, Z_k must be estimated by accounting for known ineligibility in the Parent Study sample. Let S_{AH} denote the full Add Health Wave I sample, P_{AH} denote the set of children excluded prior to sampling (for example, because their parents did not satisfy the Parent Study eligibility criteria) and let D_{PS} denote the corresponding set children who were determined to be ineligible after they were selected for the Parent Study sample (for example, deceased children or children of deceased or otherwise ineligible parents). Denote the Parent Study sampling frame as $F_{PS} = S_{AH} - P_{AH}$. Thus, an estimate of Z_k is

$$\hat{Z}_k = \sum_{i \in F_{PS}} \omega_{Ii} z_{ik} - \sum_{i \in D_{PS}} w_{1i} z_{ik} \tag{11}$$

Now, the calibration equation in (9) can be computed substituting $\hat{Z}_k$ for Z_k .

4.3 Parent Weight

For analyses where the parent is the analysis unit, the child weight is not appropriate because, while a child has only one chance of selection, parents may have multiple chances selection, one for each eligible child to whom they are linked. A parent's selection probability can be expressed as the product of two probabilities: the probability the parent has at least one child in the Add Health survey and, thus, they themselves are in the Add Health survey and the probability that parent is in the Parent Study sample given the parent is in the Add Health survey. The former probability is the probability that at least one child linked to the parent is selected for the Add Health because that would automatically lead to the parent being selected. The latter probability is given by (2). Thus, for the *j*th parent, the selection probability is

$$\begin{aligned}\Pr(\,j\text{th parent in PS}) &= \Pr(\,j\text{th parent in AH}) \times \Pr(j\text{th parent in PS} \mid j\text{th parent in AH}) \\ &= \left[1 - \prod_{i \in F_j}\left(1 - \frac{1}{\omega_{AH\,i}}\right)\right] n \frac{\text{MOS}_j}{\text{TMOS}} \end{aligned} \tag{12}$$

where F_j is the set consisting of all eligible children on the Parent Study frame that are linked to the *j*th parent. Denote the probability in (12) by π_j . Then the selection weight for this parent is $w_{Pj}^{*} = \pi_j^{-1}$. Note that when the parent has only one child, the parent's selection weight is equal to the child's selection weight. However, for parents of multiple children, a smaller weight is assigned to the parent to reflect that parent's increase probability of selection.

It is important that the weight assigned to the parent is consistent with the weights assigned to the children of the parent. It is possible that the application of (12) could result in inconsistent child versus parent weights if parent weights are computed independently of the child weights. To avoid this problem, equation (12) will not be used to derive the parents' weights. Instead, the weights for parents will be derived from the children's weights using the formula

$$w_{\text{Pj}} = \left[1 - \prod_{i \in F_j}\left(1 - \frac{1}{w_{\text{FINAL},i}}\right)\right]^{-1} \tag{13}$$

where $w_{\text{FINAL},i}$ is the *i*th child's weight defined in (10). The parent weights computed using (13) will be calibrated to force agreement with the weight totals for the entire Wave 1 sample of eligible parents using the calibration factor, f_P, defined as

$$f_P = \frac{\sum_{j \in S_I} \omega_{Pj}}{\sum_{j \in S} w_{Pj}} \tag{14}$$

where S_I is the Wave I sample after removing deaths and other ineligible children,

$$\omega_{\text{Pj}} = \left[1 - \prod_{i \in F_j}\left(1 - \frac{1}{\omega_{I\,i}}\right)\right]^{-1} \tag{15}$$

for all $j \in S$. Thus, the final parent weight will be $w_{PF,j} = f_P w_{Pj}$. Further calibration of the parent weights (for example to age, race and gender totals) may be possible so long as consistency between the child-level and parent-level weights can be maintained.

5.0 Quality Control Verification of the Weights

The Parent Study weighting process is complex and requires multiple steps. To ensure that the weights were computed as specified and produced the desired results, a comprehensive QC of the weighting process was conducted (Appendix C). For both parent and children's weights, the sample weight totals matched the corresponding Wave I weight totals on the variables used in the calibration process, after accounting for deaths and other out of scope units. In addition to checking the calibration totals, other totals that were not used in the calibration were inspected to ensure they were consistent with prior wave totals. Other checks and analyses were conducted throughout the weighting process to verify each step of the weighting process.

Glossary of Terms

Add Health parent: Biological, step, or adoptive parent of an Add Health child, who currently lives in the U.S., and who participated in the Wave I of Add Health.

AH: Add Health.

Child is interchangeably used to denote an Add Heath sample member.

MDES - Minimum Detectable Effect Size: The smallest effect size (as defined by Cohen, 1988) that can be declared significant with a Type I error of 5 percent and a Type II error of 20 percent.

MOS - Measure of Size: Relative size (and hence relative probability of selection) of a sampling unit in a selection mechanism where units are chosen proportionally to a certain quantity or measure.

PPS – Probability Proportional to Size Sampling: Random sample selection mechanism, where frame units are chosen with probabilities proportional to a certain quantity or measure.

PS: PS.

Replicate: The total sample is partitioned into two subsamples or replicates. The second replicate is held in reserve and will be fielded according to the unknown survey quantities estimated from the first replicate.

Sampling Unit: The Add Health parent.

Second Parent: Current, live-in, partner of the Add Health parent.

UWE - Unequal Weighting Effect: Increase in variance due to the variation in the final selection (or base) weights.

APPENDIX A:
STEPS FOR PS SAMPLE FRAME CONSTRUCTION

The following steps describe how to arrive at the baseline file for selecting the samples for the PS. The N = numbers after each item describe the size of the file at the completion of the step.

1. Interviewed at Wave I: N = 20,745.

   RTI received this master sampling data file, r01_samplevars. sas7bdat, on 9/23/13. It contains a record for everyone who participated in the Wave I in-home interview.

2. Delete cases where dsp3 = missing: N = 20,058.

   Beginning with Wave III, these deleted cases are those that Add Health decided not to follow up, N = 687. In general, these are Wave I cases without a sampling weight who do not have a pair ID.

3. Delete cases where dsp3 = 459: N = 19,962.

   A Wave III disposition code of 459 identifies respondents confirmed deceased at Wave III.

4. Delete cases where dsp4 = 459: N = 19,831.

   A Wave IV disposition code of 459 identifies respondents confirmed deceased at Wave IV. Wave V cases that were deceased prior to sampling (as determined by the National Death Index data base) will also be deleted and considered as ineligible for the purposes of weighting.

5. From the file created by steps 1-4, select cases where PC1 = bio/step/adoptive. N = 16,087.

   This step retains only those cases in which the person who answered the Wave I parent questionnaire is a biological mother or father (codes 1 and 9), a stepmother or stepfather (codes 2 and 10), or an adoptive mother or father (codes 3 and 11).

6. From the above remaining cases, select cases where PC1DEC = 0: N = 15,562.

   If the Add Health sample member reports at Wave II, III, or IV that the person who answered the parent questionnaire is deceased, then those cases are deleted. NOTE: Two Wave II deceased cases are not included in the master data file, but must be hard-coded using their particular IDs (variable AID).

7. The resulting file has 15,562 cases to be used as the baseline for the PS sampling.

APPENDIX B:
VARIABLES THAT WILL BE USED AS POTENTIAL INDEPENDENT VARIABLES IN NONRESPONSE ADJUSTMENT AND POPULATION CALIBRATION

B.1. Demographic Variables of Eligible Children

Demographic variables of eligible children that will be used as potential independent variables in nonresponse adjustment and population calibration are: region, state, urbanicity, gender, race and age. Missing values will be imputed prior to adjustment. The following provides some relevant information on the variables that will be used in the weighting process. This information will be updated and extended in the post-weighting documentation.

- State (**state_w5**)
  Variable **state_w5** was created based on the contact addresses at Wave V, when we were performing the propensity modeling in 2015.

- Urbanicity (**urban_w5**)
  Variable **urban_w5** was created based on the contact addresses at Wave V, when we were performing the propensity modeling in 2015.

- Region (**region_w5**)
  Variable **region_w5** is created based on variable **state_w5**.

- Gender (**gender**)
  Variable **gender** was created based on the original AddHealth Wave I to Wave IV data. The SAS code used to create this variable is:

  ```
  gender=RSEX_W4;
  if gender=. then gender=RSEX_W3;
  if gender=. then gender=RSEX_W2;
  if gender=. then gender=RSEX_W1;
  if gender=. then gender=RSEX_1994;
  ```

- Race (**race_r**)
  Variable **race_r** was created based on the original AddHealth Wave I data. The SAS code used to create this variable is:

  ```
  if race=. then race=Prace;
  if race=. then race=6;
  race_r=race;
  if race in (3,4,5) then race_r=3;
  if race=6 then race_r=4;
  ```

- Age (**age, age_c**)

Variable **age** (that will be used in the nonresponse adjustment as a potential continuous independent variable) was created based on the original AddHealth Wave I to Wave IV data. The SAS code used to create this variable is:

```
byear=RBirthYear_W4;
if byear=.  then byear=RBirthYear_W3;
if byear=.  then byear=RBirthYear_W2;
if byear=. then byear=RBirthYear_W1;
if byear=1996 then byear=1981; *there is only case with byear=1996, and we recode it as byear=1981;
age=2015-byear;
```

Variable **age_c** (that will be used in population calibration) was created based on variable **age**. The SAS code used to create this variable is:

```
if age in (32,33,34) then age_c=1;
else if age=35 then age_c=2;
else if age=36 then age_c=3;
else if age=37 then age_c=4;
else age_c=5;
```

B.2. Level of Effort (LOE) to Contact in Waves III and IV (z_{III} and z_{IV}) of Eligible Children

The LOE variable is the number of contacts (either by phone or mail) made to the person for a survey interview. Quantiles (viz., sextiles) of the LOE distributions at Waves III and IV were used to create the 6-point scale variables z_{III} and z_{IV}, which indicate the level of difficulty for getting this person to participate the survey. Table B.1 presents the definitions of the 6-point scale variables z_{III} and z_{IV} based on their corresponding LOE variables. The number of contacts in the table correspond to the 1/6, 2/6, 3/6 etc. quantiles of the distribution of LOE variable for each wave. Variables z_{III} and z_{IV} will be used in nonresponse adjustment as potential categorical independent variables.

Table B.1. The 6-Point Scale Variables (z_{III} and z_{IV}) Based on the Level of Effort in Waves III and IV

Number of Contacts Wave III	Number of Contacts Wave IV	6-Point Scale Variables (z_{III} and z_{IV})
1 ~ 2	1 ~ 2	Very Easy
3 ~ 4	3 ~ 4	Moderately Easy
5	5 ~ 8	Slightly Easy
6 ~ 7	9 ~ 12	Slightly Difficult
8 ~ 12	12 ~ 20	Moderately Difficult

12 ~ Maximum + Nonresponse	20 ~ Maximum + Nonresponse	Very Difficult

B.3. Demographic and Other Variables of Parents

The following variables of parents are used as potential independent variables in nonresponse adjustment and population calibration:

- Parent Relation and Parent's Gender (**PC1, Pgender**)

Variable **Pgender** was created based on variable PC1 (Parent Relation) in the baseline data file of the parent study. The SAS code used to create this variable is:

```
if PC1 in (1,2,3) then Pgender=2; else Pgender=1;
```

Both variables PC1 and Pgender will be considered as potential independent variables in nonresponse adjustment.

- Parent's Age Group (**Page_c**)

Variable **Page_c** was created based on variable PA2 (parent's age when they did the parent questionnaire in 1994) in the baseline data file of the parent study. The SAS code used to create this variable is:

```
Page=PA2;
if Page<=30 then Page_c=1;
if Page>=31 and Page<=40 then Page_c=2;
if Page>=41 and Page<=50 then Page_c=3;
if Page>=51 then Page_c=4;
if PA2=. or PA2=996 then Page_c=9;
```

- Parent's Race/Ethnicity (**PRACE**)
- Family Structure (**FAMTYP25**)
- Family Structure (**W1FAMST**)
- Parent's Education at Wave I (**W1PRHGTE**)
- Number of Children of the Parent

APPENDIX C:

Quality Control Summary Results of Final Weights for the Add Health Parent Study

1. Weight Totals: Overall and by Subgroups

For the child-level weights, the overall weight totals as well as the weight totals by children's age, gender, race, parents' race and family structure strata variable (W1FAMST) were checked against the corresponding estimated control totals after removing deaths and other out of scope units from the baseline data file of the parent study. These totals should agree exactly because the child-level weights were calibrated to these estimated control totals as part of the weighting process. Weight totals that differ by more than rounding are indicative of a potential problem in the calibration/post-stratification process. In addition, we checked weight totals by three other variables (children's region and urbanicity, and parents' age group) that were not used in the calibration for consistency to the extent that they should be consistent (e.g. the differences should be within 10% of the estimated control totals for the majority groups). As can be seen in Table 1, the overall weight total as well as the weight total by variables used in the calibration of the child-level weights agree exactly with the estimated control totals and their weight totals by variables that were not used in the calibration remained close to the corresponding estimated control totals. In the last column, the weight totals of the parent-level weights are presented. They remained close to the corresponding estimated control totals for the child-level weights as well.

Table 1. Parent Study Weight Totals: Overall and by Subgroups

Variable	Estimated Control Total for the Child-level Weight[f]	Totals for Child Weight (n=2,244)	Totals for Parent's Weight (n=2,013)
Variables used in the calibration			
Overall	17,000,746.63	17,000,746.63	14,786,793.00
Age×Gender×Race[a]			
111	248,900.75	248,900.75	231,091.31
112	266,499.35	266,499.35	240,717.15
113	179,805.73	179,805.73	179,213.33
114	1,609,124.70	1,609,124.70	1,438,029.45
121	281,721.02	281,721.02	223,972.19
122	289,583.04	289,583.04	247,034.10
123	137,471.44	137,471.44	101,486.38
124	1,669,762.45	1,669,762.45	1,512,796.93

Variable	Estimated Control Total for the Child-level Weight[f]	Totals for Child Weight (n=2,244)	Totals for Parent's Weight (n=2,013)
211	186,584.71	186,584.71	173,854.50
212	198,855.41	198,855.41	179,924.46
213	101,297.40	101,297.40	79,431.25
214	1,036,992.20	1,036,992.20	847,752.06
221	185,954.52	185,954.52	174,517.17
222	225,973.44	225,973.44	200,067.23
223	77,503.74	77,503.74	71,902.46
224	1,033,440.68	1,033,440.68	828,168.92
311	188,707.39	188,707.39	157,238.07
312	187,184.19	187,184.19	161,095.60
313	90,845.71	90,845.71	55,499.75
314	956,145.99	956,145.99	820,130.87
321	167,663.45	167,663.45	171,244.37
322	190,772.87	190,772.87	168,075.38
323	84,352.90	84,352.90	80,560.44
324	1,008,977.22	1,008,977.22	870,773.02
411	169,634.04	169,634.04	173,257.05
412	192,925.45	192,925.45	181,801.16
413	86,587.59	86,587.59	73,672.28
414	946,949.63	946,949.63	841,387.92
421	139,379.21	139,379.21	123,796.82
422	216,604.95	216,604.95	182,843.42
423	88,859.55	88,859.55	83,539.66
424	897,338.43	897,338.43	749,136.53
511	239,945.02	239,945.02	198,687.23
512	283,796.00	283,796.00	239,040.40
513	92,586.08	92,586.08	66,231.87
514	1,357,541.80	1,357,541.80	1,181,571.71
521	229,734.72	229,734.72	211,744.90
522	260,009.32	260,009.32	243,371.23
523	111,374.22	111,374.22	88,493.84
524	1,083,360.32	1,083,360.32	933,640.60
Parent's Race (PRACE)[b]			
Hispanic, All Races	1,790,944.03	1,790,944.03	1,655,752.09
Black or African American, Non-Hispanic	2,172,246.05	2,172,246.05	1,936,822.16
Other	12,056,642.55	12,056,642.55	10,386,084.94
White, Non-Hispanic	980,914.00	980,914.00	808,133.82

Variable	Estimated Control Total for the Child-level Weight[f]	Totals for Child Weight (n=2,244)	Totals for Parent's Weight (n=2,013)
R01 Cluster Strata (W1FAMST)			
Stratum 1: Two biological parents	9,794,627.15	9,794,627.15	8,396,423.88
Stratum 2: Single mother (bio-mom & extended family, no pop, other pop)	4,326,762.61	4,326,762.61	3,911,752.73
Stratum 3: Residual (bio-pop or step/adoptive mom and/or pop)	2,611,624.74	2,611,624.74	2,241,621.09
Stratum 4: Not recognized (other 'non-family')	267,732.13	267,732.13	236,995.29
Variables not used in the calibration			
Region[c]			
Northeast	2,219,506.39	2,049,562.40	1,834,160.23
Midwest	4,869,708.33	5,188,450.18	4,217,346.89
South	6,594,898.14	6,245,519.84	5,625,903.41
West	3,111,894.20	3,407,488.49	3,005,498.27
Urbanicity[d]			
Urban	12,114,254.64	12,137,367.19	10,685,489.13
Large Rural	1,439,907.63	1,473,012.09	1,267,286.09
Small Rural	3,245,725.67	3,280,641.63	2,730,133.57
Parent's Age Group (age in 1994)[e]			
30 and under	246,266.02	236,746.89	219,163.16
31-40	8,188,128.15	7,931,210.67	6,844,751.98
41-50	7,469,794.04	7,648,582.19	6,768,624.83
50 and above	784,877.47	891,800.63	789,651.84

a: The first digit stands for value in the age group variable, the second digit stands for value in the gender variable, and the third digit stands for value in the race variable.

b: The original PRACE variable in the baseline file was recoded so that parent with multiple children has the same PRACE value across different children and the missing PRACE values were imputed based on children's race/ethnicity variable.

c: There are a few cases (around 200) in the baseline file having their values in region missing.

d: There are a few cases (around 200) in the baseline file having their values in urbanicity missing.

e: This is parent's age when they did the parent questionnaire in 1994. There are a few cases (around 50) in the baseline file having their values in parent's age missing.

f: The control totals are derived based on equation (11) in the weighting specification document (Biemer and Liao, 2018).

2. Unequal Weighting Effects

To evaluate the weight variation, the unequal weighting effects (UWEs) were computed for each set of weights as: UWE = $n\Sigma w^2/(\Sigma w)^2$, where n is the sample size and w is the weight and the sum extends over all sample units having positive weights. If the UWE becomes unreasonably large after weighting adjustment, a weight trimming (smoothing) adjustment can be implemented to reduce weight variation. The UWEs of the base weights for the parent study (prior to any weighting adjustment), the final child-level weights and parent-level weights are displayed in Table 2. No weight trimming was done because the UWEs of the final weights are deemed as acceptable.

Table 2. Unequal Weighting Effects

UWE for Base Weight at the Child Level Among Respondents (n=2,244)	UWE for Child's Weight (n=2,244)	UWE for Parent's Weight (n=2,013)
1.038	1.219	1.262

a: the base weight is the inverse of the selection probability in the Parent Study described in equation (4) of the weighting specification document.