National Longitudinal Study of Adolescent Health

Wave III
Education Data Weights

Stephen Roey

❖❖❖❖❖

Carolina Population Center
University of North Carolina at Chapel Hill

July 2005

This research was supported by a grant from the National Institute of Child Health and Human Development under grant R01 HD40428-02 to the Population Research Center, University of Texas at Austin; Chandra Muller (PI) and the National Science Foundation grant number REC-0126167, Chandra Muller (PI). Persons interested in obtaining data files from Add Health should contact Add Health, Carolina Population Center, 123 W. Franklin Street, Chapel Hill, NC 27516-2524 (addhealth@unc.edu).

Wave III Education Data Weights

There have been three waves of data collection in Add Health. Wave I was conducted in 1994-95 and collected information from a nationally representative sample of adolescents in grades 7 through 12. The sample covered 132 schools in 80 communities. The schools were drawn from a sampling frame of US schools sorted by size, school type (public, private, parochial), region of the country, location (urban, suburban, rural), and percent minority. Schools were selected systematically with probability proportional to enrollment. In addition to the main sample, 52 feeder schools that contributed students from high schools without 7th and 8th grades were randomly selected with probability proportional to percent of the high school's entering class.

In Wave I, 20,745 in-home interviews were conducted from adolescents who answered, in most cases, in-school questionnaires as well as interviews with their parents. These students constitute the longitudinal sample in Add Health. The longitudinal sample also includes oversampling of special groups. This sample was reinterviewed at home in Wave II in 1996. The Wave III sample consists of all Wave I respondents who could be located and reinterviewed in 2001-02. Only 15,170 original Wave I respondents were interviewed in Wave III.

The Adolescent Health and Academic Achievement Study (AHAA), conducted by the University of Texas at Austin and Westat, collected transcript information from Wave III respondents. Although special efforts were made to collect this information, some transcripts were missing for some students because of the following reasons:

- student did not agree to participate in the study,
- student did not attend high school,
- student was home-schooled,
- student attended school outside of the US,
- student did not provide adequate school information,
- school was closed,
- school refused to provide the student's transcript, and
- school provided incomplete or erroneous transcripts.

Estimates that incorporate transcript information can be computed using the Add Health analytical weights. However, these estimates will be biased because of the missing transcripts, therefore, new weights were computed to reduce the bias. Adjusted weights were created for two sets of respondents: longitudinal Wave I, II, and III respondents and cross-sectional Wave I and Wave III respondents. Table 1 shows the new adjusted weights. The procedure used to create these weights is described in the following paragraphs.

Table 1. Adjusted weights for transcript nonresponse

File	Weight name	Description
Restricted Use	TWGT3	Education Data longitudinal weight
	TWGT3_2	Education Data cross-sectional weight

In order to create new analytical weights, students with missing transcripts were considered nonrespondents. The Education Data weights were created by adjusting the Add Health weights for transcript nonresponse in three steps: assignment of disposition codes, adjust Add Health weights for transcript nonresponse, and benchmark (sample-based raking) the adjusted weights to control totals derived using the Add Health sample. These steps are described in the following paragraphs.

In the first step of weighting, sampled students[1] from either longitudinal Wave I, II, and III or cross-sectional Wave I and III were assigned one of the following response codes (*RSTATUS*) based on the transcript disposition code assigned during data collection:

ER **Eligible respondents.** This group consists of all eligible students with complete and usable transcript information.

ENR **Eligible nonrespondents.** This group consists of all eligible students with missing, incomplete, or unusable transcript information. This group also includes students who refused to participate in the transcript component of the study.

IN **Ineligible or out-of scope students.** This group consists of all sampled students who did not have transcript information because they never graduated, were home-schooled, or graduated outside of the US.

Table 2 shows the assignment of the response codes based on the transcript disposition code.

Table 2. Response code assignment for the Education Data weights

Response status (*RSTATUS*)	Transcript disposition code	Description
ER	1	Transcript received
ENR	2	Unable to locate, no record found
	3	Unable to locate, no longer in school database
	4	Unable to locate, unknown reason
	5	Refusal, student or school
	7	TRF or transcript not legible
	8	Not valid school
	9	Incorrect school given by student
	11	Unable to locate school or TRF

[1] Only students with a positive Add Health analytical weight were adjusted for transcript nonresponse.

	12	Other, not received
	13	No response from school
	14	Dropped from RTI
	15	No TRF signed
IN	6	Student never graduated
	10	Home-schooled or school of graduation in foreign country

Table 3 shows the number of sampled students for the different files by disposition code.

Table 3. Distribution of the number of sampled students by response status

	Wave I-III (Longitudinal)		Wave III (Cross-sectional)	
Response Status	Number of records	Percentage	Number of records	Percentage
ER	8,832	81.57	11,607	81.04
ENR	1,978	18.27	2,681	18.72
IN	18	0.17	34	0.24
Total	10,828	100.0	14,322	100.0

In the second step of weighting, the weights of students with transcript information (*ER,* eligible respondents) were adjusted to account for students with missing transcripts (*ENR*, eligible nonrespondents). In this adjustment, the weights of the students coded as *IN* (out of scope) was unchanged. It was assumed that all out-of-scope students have been found during the collection of the transcript data.

The transcript nonresponse adjusted student weight, $AD1W_i$, was computed as

$$AD1W_i = AD1F_c * AD0W_i,$$

where $AD0W_i$ is the Add Health weight and $AD1F_c$ is the transcript nonresponse adjustment factor computed as

$$AD1F_c = \begin{cases} \dfrac{\sum_{i \in ER,ENR} AD0W_i}{\sum_{i \in ER} AD0W_i} & i \in ER \\ 0 & i \in ENR \\ 1 & i \in IN \end{cases},$$

where the groups *ER*, *ENR,* and *IN* were defined in Table 2. The response adjustment was done within weighting classes (Brick and Kalton, 1996). Weighting class adjustments are effective in reducing nonresponse biases if the weighting classes are internally homogeneous with respect to the response propensity but as different as possible across classes without unduly inflating sampling variances (Kish, 1992). Nonresponse adjustments are computed and applied separately by weighting classes, where a weighting class is defined using characteristics known for both nonrespondents and respondents. The adjustment reduces bias if either response rates or the survey characteristics are more similar within the classes. Weighting classes were created using variables for Census region, race, grade, and school in the restricted use files. Because of fewer numbers of records, grade was excluded in the creation of the weighting classes in the public use files. Response rates tables were examined in order to determine which variables would be used to create the classes.

Very large adjustment factors or factors that are much different from others can occur in weighting classes with high nonresponse rates or with a small numbers of respondents. Combining weighting classes with few cases to form new classes with at least 30 respondents often compensates for large adjustment factors. However, there are times when weighting classes with more than 30 respondents have a large adjustment factor. If a class had a large adjustment factor, it was combined with a demographically similar class to form a new weighting class with a smaller adjustment factor. Census region was considered as a hard boundary and no weighting classes were collapsed across region.

Add Health analytical weights were poststratified to control totals computed by grade, race, and gender. In order to achieve a greater consistency with estimates produced using Add Health analytical weights, the Education Data nonresponse adjusted weights were raked to control totals derived from the Add Health analytical weights in the last step of weighting. This step also removed any residual bias not accounted in the nonresponse adjustments but included as part of the raking dimensions.

Raking (Brackstone and Rao, 1979, and Deville and Särndal, 1992) is an estimation procedure in which estimates are controlled to marginal population totals. Raking can be considered a multidimensional post-stratification procedure because the weights are post-stratified to control totals for different dimensions successively. The process is iterated until the control totals for all the dimensions are simultaneously satisfied within a specified tolerance. A sample-based raking approach was used for the Education Data weights. Brick and Kalton (1996) call the procedure a sample-based adjustment, and Lundström and Särndal (1999) refer to this as Info-S calibration. In this procedure a larger sample is used to benchmark a smaller sample through raking. In this case, the larger sample corresponds to Add Health respondents while the smaller sample corresponds to all Education Data respondents.

The raking estimator is design-unbiased in large samples and is efficient in reducing the variance of the estimates if the estimates in the cross-tabulation of the dimensions are consistent with a model that ignores the interactions between variables. For simplicity, assuming two dimensions (in the Education Data there were three dimensions shown in Table 4), the raked weight can be written as

$$\tilde{w}_{cd,i} = w_{cd}\hat{\alpha}_c\hat{\beta}_d ,$$

where w_{cd} is the pre-raked weight of an observation in cell (*c*,*d*) of the cross-tabulation, $\hat{\alpha}_c$ is the effect of the first variable, and $\hat{\beta}_d$ is the effect of the second variable. In this formulation, there is no interaction effect. In this sense, the weights are determined by the marginal distributions of the control variables. As a result, the sample sizes of the marginal distributions are the important determinants of the stability of the weighting procedure. Furthermore, raking permits the use of more variables or control totals than is possible with simple poststratification.

The final Education Data raked weight, $AD2W_i$, was computed as

$$AD2W_i = AD2F_k * AD1W_i$$

where $AD2F_k$ is the sample-based raking factor for dimension *k* computed to satisfy the condition that

$$\hat{C}_k = \sum_{\substack{i\in k \text{ and} \\ i\in ER,IN}} AD1F_k \cdot AD1W_i = \sum_{\substack{i\in k \text{ and} \\ i\in ER,IN}} AD2W_i = \sum_{\substack{i\in k \text{ and} \\ i\in ER,ENR,IN}} AD0W_i$$

where $\hat{C}_k$ is the control total for each dimension *k*. The total $\hat{C}_k$ is an estimated total computed by adding the sum of weights of the Add Health final weight for dimension *k*, for *k*=1 to 3. Table 4 shows the dimensions used to rake the sample. Control totals computed using fewer that 50 students were collapsed. Cells of raking dimensions with fewer than 30 respondents were also collapsed. Extensive collapsing was done for the public use file because of fewer records.

Table 4. Raking dimensions

Dimension	Description
1	Gender*Grade*Race
2	Region*Age group
3	Region*Race*Grade

After raking the sample for the first time, weights were examined to determine the presence of extreme weights. One outlier was detected and trimmed by attaching a trimming factor to the weight before raking. The trimmed weights were then re-raked. The re-raked weights were examined to verify the procedure was effective at reducing the outlier.

The method used to adjust for student nonresponse adequately adjusts for school nonresponse. The approach used reflects the effect of adjusting for school nonresponse because the schools were used to create the nonresponse adjustment classes in the original files. The sample size is smaller (fewer PSUs) and the estimates are less precise (i.e., fewer degrees of freedom) due to nonresponse.

REFERENCES

Brackstone, G. J. and J. N. K. Rao. 1979. "An Investigation of Raking Ratio Estimation." *Sankhya* C 41:97-114.

Brick, J. M. and G. Kalton. 1996. "Handling Missing Data in Survey Research." *Statistical Methods in Medical Research* 5:215-238.

Deville, J. C. and C.-E. Särndal. 1992. "Calibration Estimators in Survey Sampling." *Journal of the American Statistical Association* 87:376-382.

Kish, L. 1992. "Weighting for Unequal Pi." *Journal of Official Statistics* 8:183-200.

Lundström, S. and C.-E. Särndal. 1999. "Calibration as a Standard Method for Treatment of Nonresponse." *Journal of Official Statistics* 15:305-327.