National Longitudinal Study of Adolescent Health

Grand Sample Weight

❖ ❖ ❖ ❖ ❖

Roger Tourangeau
Hee-Choon Shin
National Opinion Research Center

Carolina Population Center
University of North Carolina at Chapel Hill

October 1999

This research was supported by grant P01-HD31921 from the National Institute of Child Health and Human Development. Further information may be obtained by contacting Jo Jones, PhD, Project Manager, 919/962-8412 (email: jo_jones@unc.edu) at the Carolina Population Center, CB# 8120 University Square, Chapel Hill, NC 27516-3997.

https://doi.org/10.17615/C67669

Design and Implementation of the In-School Sample

This section describes the design and procedures used for selecting schools. It also provides information on the calculation of sample weights and about procedures used to adjust sample weights for nonresponse.

Sample Design

The initial sample for the Add Health study consisted of 80 high schools and 52 associated feeder schools—middle schools and junior high schools that sent graduates to the sample high school. (An additional four feeder schools were selected but declined to take part in the study.) The high school sample was selected to represent all high schools in the United States; as a result, the high school students attending these schools constitute a nationally representative sample of the high school population. Similarly, the feeder school sample constitutes a nationally representative sample of schools whose graduates go on to enroll in high school. Within these 132 sample schools, all students in grades 7 through 12 were asked to complete the In-school Questionnaire.

Selection of the Initial High School Sample

The frame for selecting the sample of high schools was the QED database, thought to be the most comprehensive list of high schools available. For this study, a high school was defined as any school that included an 11th grade; as an operational matter, schools whose grade span could not be determined from the QED data were also included. The frame listed a total of 26,666 schools, including both public and private schools. Among the public schools were schools sponsored by the Department of Defense, the Bureau of Indian Affairs, and the Department of State.

The 80 sample high schools were selected systematically, with selection probabilities proportional to the school's enrollment—that is, high schools with higher enrollments (according to the QED data) had a greater chance of selection. Prior to sampling, the schools were sorted by size (125 or fewer, 126 to 350, 351 to 775, 776 or more students), school type (public, parochial, private), census region (Northeast, Midwest, South, West), level of urbanicity (urban, suburban, rural), and percent white (0, 1 to 66, 67 to 93, 94 to 100). Because the study used systematic sampling on a sorted list, it assured that the sample was representative along the dimensions used to sort the list; this technique is sometimes referred to as implicit stratification. Selecting schools with probabilities proportional to enrollment facilitated the selection of a nearly self-weighing core sample of students.

Replacement Schools

One of the sampled high schools was not, in fact, eligible for the study and others among the initial 80 selections refused to take part. Of the initial selections, only 52 were eligible and agreed to cooperate. The remaining 28 refusals were replaced by similar high schools. Replacement schools were found by sorting the frame file by eight variables:

1. School size (125 or fewer, 126 to 350, 351 to 775, 776 or more students);
2. School type (public, parochial, private);
3. Urbanicity status (urban, suburban, rural);
4. Percent white (0, 1 to 66, 67 to 93, 94 to 100);
5. Grade span (K-12; 7-12; 9-12; 10-12; vocational/technical; alternative; special education);
6. Percent black (0, 1 to 6, 7 to 33, 34 to 100);
7. Census region (Northeast, Midwest, South, West);
8. Census division (New England, Middle Atlantic, East North Central, West North Central, South Atlantic, East South Central, West South Central, Mountain, Pacific).

Within each category, the schools were sorted in a random order. The replacement school was the school that followed the initial sample school on the sorted file. As a result, the replacement school matched the selection it was replacing with respect to the characteristics listed above. If the first replacement school was ineligible or failed to cooperate, it was replaced by the next school on the list; the replacement process was repeated until a replacement was found that was both eligible and willing to cooperate.

Within some categories there either were not enough potential replacements, or the list of replacements was completely exhausted before an eligible school was recruited for the study. In such cases, similar categories were combined and the file was resorted. The combination in these cases worked as follows. First, the "census division" variable was collapsed. When that failed to yield at least five potential replacements that matched the initial selection on the remaining sort variables, the "percent black" variable was collapsed to two categories. If that still failed to produce five potential replacements that matched on all of the remaining sort variables, size categories were combined.

Feeder Schools

Once a high school had been recruited for the study, the school was asked to provide the names of each junior high or middle school expected to contribute at least five students to the entering class of the high school, as well as the approximate percentage of the high school's entering class coming from each feeder school. A single feeder school was selected for each high school; the feeder's probability of selection was proportional to the percentage of the high school's entering class that came from the feeder. (For example, a feeder school that contributed one-fourth of the entering freshmen at a sample high school had a selection probability of .25.)

Four of the sample high schools drew their entering classes from a very large number of schools, and thus had no eligible feeders. At the other extreme, 20 "high schools" had grade ranges that included 7th or 8th grades and were their own feeder schools. As a result of these exclusions, a total of 56 feeders were selected. Four of them refused to take part in the study.

In-School Student Selection

All students in the eligible grade range at the participating schools were supposed to complete In-school Questionnaires. There was no sampling of eligible students within the selected schools. However, four of the sample schools did refuse to allow In-school Questionnaires to be administered. (These schools did permit students to be sampled from school rosters for subsequent in-home data collection.)

Calculation of School-Level Weights

Survey weights are calculated for four main reasons:

1. to allow the sample totals to serve as estimates of population totals;
2. to compensate for differences in selection probabilities across different members of a sample;
3. to compensate for differences in response rates across different subgroups of the sample; and
4. to adjust for chance fluctuations of the composition of the sample from the composition of the population as a whole.

The weights calculated for the in-school component accomplish all four of these purposes.

The in-school weights were developed in nine separate steps. In the first three of these, weights were calculated for the sample high schools; in the next three, weights were calculated for the feeder schools. These school-level weights were only a preliminary to the development of student-level weights, which were calculated in the final three steps of the weighting process.

Step 1: Adjusting for the high school selection probabilities. As noted earlier, the sample high schools were selected with probabilities proportional to their measure of size (MOS)—the school's enrollment according to the QED file. For high school i, the selection probability was

$$P_i = \frac{MOS_i}{MOS/80} = MOS_i/183466.84 ,$$

where MOS_i is the number of students in high school i and MOS is the total number of students in all the QED high schools. (183466.84 is the actual total divided by 80.) The initial weights (W_{1i}) for the high schools are the reciprocals of these selection probabilities:

$$W_{1i} = 1/P_i .$$

Step 2: Adjusting to QED totals. Because of random sampling error, the sum of the school weights from the first step did not equal the total number of schools on the sampling frame. The initial school weights (W_{1i} from Step 1 above) were adjusted to match the totals on the QED file. Separate adjustments were made for four regions (Northeast, North Central, South, and West). The poststratification adjustment factor (PS) for region c was just the ratio of the total count from the frame for that cell (T_c) to the sum of the weights for the sample schools in the cell:

$$PS_c = \frac{T_c}{\sum_{i=1}^{m_c} W_{1ic}} ,$$

where m_c is the number of sample schools in region c. The weights (W_{2i}) adjusted to the frame totals are just the product of the initial weight and the poststratification adjustment factor:

$$W_{2ic} = PS_c \times W_{1ic} .$$

Step 3: Adjusting for ineligible high schools. The QED file contains a number of ineligible high schools (schools without an 11th grade). These ineligible schools were dropped from the sample and replacement schools were selected. As a result, the weights resulting from Step 2 overestimate the total number of eligible high schools (since the sum of the weights matches the QED total, which includes ineligible schools). In the third step, the high school weights were adjusted to compensate for school ineligibility. The overall eligibility rate estimated was based on the first 80 schools sampled:

$$E = \frac{\sum_{i=1}^{80} W_{2i} I_i}{\sum_{i=1}^{80} W_{2i}} ,$$

in which E is the weighted eligibility rate and the variable I is a flag taking on the value 1 for eligible schools and 0 for ineligible schools. Because of the small number of schools involved, only an overall adjustment for ineligibility was made.

The final high school weights (W_{3i}) are the product of the weight from Step 2 and the eligibility rate:

$$W_{3i} = E W_{2i} .$$

Step 4: Feeder school selection probabilities. The selection probability for a feeder school depended on the selection probability of each high school to which it sent graduating students:

$$P_{Fk} = \sum_{i=1}^{M_k} P_i \frac{n_{ki}}{n_i} , \qquad (1)$$

in which the summation is over the M_k high schools to which feeder school k sends graduating

students, P_i is the selection probability for high school i receiving those graduates (P_i is defined in Step 1 above), and n_{ki}/n_i is the proportion of high school i's entering class that came from feeder school k.

NORC distinguished several special cases of Equation (1) when it calculated feeder school weights. Some feeder schools send all of their graduates to a single high school. In such cases, the feeder school's selection probability becomes the product of the high school's selection probability and the proportion of its entering class that came from the feeder school:

$$P_{Fk} = P_i \frac{n_{ki}}{n_i} \quad . \qquad (2)$$

A second special case involves situations in which the feeder school and high school are in a one-to-one relationship—that is, all of the graduates from the feeder are expected to enter the high school and all of the entering students at the high school are expected to come from the feeder. In several communities, the high school and feeder school were in fact a single combined school and thus formed such a one-to-one relationship. In a few other cases, the high school and its sole feeder were in separate facilities but still formed a one-to-one relationship. In cases where the feeder sent all of its graduates to a single high school and that high school had only the one feeder, n_{ki} equals n_i and Equation (2) simplifies even further:

$$P_{Fk} = P_i \quad . \qquad (3)$$

For the remaining feeder schools, the weights were based on Equation (1). Because the direct application of Equation (1) requires collecting additional data (on how many graduates the feeder school sends to non-sample high schools), an approximate selection probability was calculated that did not require further data. Assuming that the size of the entering class (n_i) for a given high school is a constant fraction, f, of its measure of size (that is, its estimated total enrollment), the feeder's selection probability can be reexpressed as follows:

$$\begin{aligned} P_{Fk} &\approx \sum_{i=1}^{M_k} P_i \frac{n_{ki}}{f\,MOS_i} \\ &= \sum_{i=1}^{M_k} \frac{80 n_{ki}}{f MOS} \\ &= \frac{80}{f MOS} \sum_{i=1}^{M_k} n_{ki} \quad . \end{aligned}$$

These equations show that the feeder school's selection probability is approximately a constant times the size of its graduating class (that is, the sum across all linked high schools of the number of graduates it sends to each one). Roster data was used for the feeder school—specifically, the size of the highest grade in the school—to estimate the size of its graduating class. The constant of proportionality (f in the equations) was estimated separately for high schools grades 9 to 12 and those spanning grades 10 to 12. The average proportion of students in the lowest grade was used in the sample school high schools to estimate the value of f. Values near .25 and .33 would be expected for the two types of schools. In fact, the values observed for the sample high schools were .27 and .35, respectively.

Whether the feeder school's selection probability was derived using Equation (2), Equation (3), or the approximation of Equation (1), its initial weight was the inverse of that selection probability:

$$W_{1k} = 1/P_{Fk} \quad .$$

Step 5: Adjusting for high school ineligibility and deviations from the QED totals. As noted

earlier, because of the replacement of ineligible schools, the value for MOS was an inflated estimate of the total enrollment in all US high schools. Similarly, the estimates of the number of high schools by region deviated somewhat from the QED totals. Both of these deviations were reflected in the initial feeder school weight. To compensate for them, the initial feeder school weight was multiplied by the high school eligibility rate defined in Step 3 above and by the poststratification adjustment defined in Step 2:

$$W_{2kc} = E \times PS_c \times W_{1kc} \ .$$

This is equivalent to using $1/W_{3i}$ in place of P_i in Equations (1) to (3) above.

Step 6: Adjusting for feeder school noncooperation. In four cases, it was not possible to recruit a feeder school to take part in the study. The weights resulting from Step 5 thus need to be adjusted for school-level nonresponse within the sample of feeder schools. If the indicator variable J_k is defined to be 1 if feeder school k agreed to take part in the study and 0 otherwise, the feeder school response rate becomes

$$R_f = \frac{\sum_{k=1}^{m_f} W_{2kc} J_k}{\sum_{k=1}^{m_f} W_{2kc}} \ ,$$

where m_f (=56) is the number of sample feeder schools. The final feeder school weight (W_{3k}), adjusted for school-level nonresponse, was

$$W_{3k} = W_{2kc}/R_f \ .$$

for the cooperating feeder schools and 0 for the nonparticipants.

Calculation of Student-Level Weights

The student In-school Questionnaire data can be weighted using the final school weights (W_3 defined above) with adjustments for school-level and student-level nonresponse to the In-school Questionnaire. For purposes of this weight, the roster-only schools—which did not permit administration of In-school Questionnaires—were treated as school-level nonrespondents.

Step 7: Adjusting for school-level nonresponse. Define the indicator Q_i to be 1 if school i permitted in-school data collection and 0 if the school i was a roster-only school. The school-level response rate becomes

$$R_s = \frac{\sum_{i=1}^{m} W_{3i} Q_i}{\sum_{k=i}^{m} W_{3i}} \ ,$$

where m is the number of sample schools. This adjustment accounts for school-level questionnaire nonresponse separately for the high schools and junior highs. The initial student-level weight (W_{1ij}) for the In-school Questionnaire for student j in school i, adjusted for school-level nonresponse, was

$$W_{1ij} = W_{3i}/R \ .$$

Step 8: Adjusting for student-level nonresponse. The next step in computing the student In-school Questionnaire weights was an adjustment for student-level nonresponse. Provisional weighting cells were formed that represented each grade-sex combination within each school. An indicator variable (J_{ijc}) was created that reflected whether student j in school i and grade-sex cell c completed a questionnaire (J_{ijc}=1) or not (J_{ijc}=0). The cell-specific response rate for school i, R_{ic}, was

$$R_{ic} = \frac{\sum_{j=1}^{n_{ic}} W_{1ijc} J_{ijc}}{\sum_{j=1}^{n_{ic}} W_{1ijc}} ,$$

where n_{ic} is the total number of sample students in adjustment cell c within each school. (For our purposes, n_{ic} was the count of students who were either listed on the roster or who completed the questionnaire or both.) The resulting student weights (W_{2ijc}) for student j at school i in cell c were

$$W_{2ijc} = W_{1ijc} / R_{jc} .$$

for students who completed the In-school Questionnaire and 0 for students who did not. To avoid extreme adjustments, cells were collapsed when the weighted response rate was less than 67 percent. The cells were collapsed within grade first, then across grades.

Step 9: Adjusting to current population estimates. As a final step, the weights were adjusted to conform to population estimates, based on 1995 Current Population Survey figures, for each grade-sex combination. These adjustments used the same general procedure as in Step 2 above. The numerator of the adjustment was the population estimate for the grade-sex group; the denominator was the sum of the weights (W_{2ij} from Step 8 above) for In-school Questionnaire respondents in that cell. The final weight was the product of the weight from Step 8 and the adjustment factor.

Design and Implementation of the Wave I In-Home Sample

This section describes the design and procedures used for selecting adolescent respondents for the Wave I in-home sample. It also provides information on the calculation of sample weights and on the procedures used to adjust sample weights for nonresponse.

Selection of the Wave I In-Home Samples

For Wave I in-home data collection, 27,559 students were selected. This total includes a core sample, consisting of 16,044 students enrolled in grades 7 through 12 at the time of sample selection; in addition, it includes of all of the students (a total of 3,350) at two high schools—the PAIRS sample schools—where saturation samples were selected for Wave I in-home follow-up. Besides the core and PAIRS samples, two groups of supplemental samples were selected for Wave I in-home data collection. One group of supplemental samples—the non-genetic supplements—included students in various ethnic categories, as well as students identified as disabled from the In-school Questionnaire data. The second group of supplemental samples—the genetic supplements—consisted of individual students and pairs of students in various types of sibling relationships. This section describes the procedures for selecting these samples; the next section discusses the procedures for weighting the Wave I In-home Questionnaire data from each of the samples.

Sample Selection: Core and PAIRS Samples

The core sample consisted of roughly equal-sized samples drawn from 12 student-level strata; the strata were formed by cross-classifying students by their sex and grade. Overall sample-size targets were set for each stratum by dividing the total size of the core sample (16,044) by the number of strata (12), yielding a target of 1,333 selections per stratum. School-level targets were also set for each stratum by dividing the overall stratum target (1,333) by the number of schools with at least one student in the stratum. For example, if 75 schools included 9th grade males, the target sample size for 9th grade males at each of those schools was set at 18. The main frame for selecting the core sample was the set of rosters developed at the sample high schools and their linked feeders, with supplementary information coming from the In-school Questionnaires. (The frame included students who either completed a questionnaire or were listed on their school roster or both.) Using the roster and In-school Questionnaire data, each student was classified into a grade-by-sex cell. In some cases, missing grade or sex information had to be imputed to allow classification of each student by stratum.

Several practical issues complicated the implementation of the core sample design. First, the sample was selected in two waves. Most of the schools had submitted rosters and completed in-school data collection in time to meet the initial schedule. However, 13 schools provided data after the original deadline. Selections were made separately for the 119 initial and 13 late schools. This meant that the sample-size targets were estimated using data only from the initial schools. Second, in several schools, the frame data were inaccurate. In particular, students appeared to be misclassified by grade and were selected into the sample in error. (For example, students incorrectly classified as 12th graders were selected at a junior high.) These off-grade students were subsequently dropped from the sample. Third, a few schools did not allow in-school data collection, limiting the information on the students. In a few other schools, In-school Questionnaires were completed but identifying information was deleted from them, making it impossible to link the questionnaire and roster data. Finally, all of the students were selected for the core sample at any schools in which two-thirds or more of the students would have been randomly selected for the core. Application of this rule added an additional 138 students to the core sample.

The selection of the sample in two waves, the deletion of off-grade selections, and the selection of all students at small schools for the core sample produced some variation in the stratum sample sizes. Table 1 shows the final distribution of core selections by sex-grade strata.

Table 1. Number of Sample Core Selections by Grade and Sex

Stratum	Number of Core Selections
Males	7,885
7th Grade	1,223
8th Grade	1,277
9th Grade	1,385
10th Grade	1,387
11th Grade	1,379
12th Grade	1,234
Females	8,159
7th Grade	1,319
8th Grade	1,275
9th Grade	1,388
10th Grade	1,393
11th Grade	1,386
12th Grade	1,398
Total	16,044

At the two (purposively selected) PAIRS schools, all of the students were selected for Wave I in-home data collection. Within those schools, the core sample cases were necessarily a subsample of the cases selected for the PAIRS sample.

Sample Selection: Non-Genetic Supplements

Eligibility for the non-genetic supplemental samples was determined by race/ethnicity and by disability status. As with the core sample, sample selection for each of the supplemental samples was restricted to students enrolled in grades 7 through 12 in one of the sample schools at the time the sample was selected. A total of five non-genetic supplemental samples were selected:

High Education Blacks. This supplement included 1,318 black students, either of whose parents were college graduates.

Cubans. The Cuban supplement included 571 students of Cuban descent.

Puerto Ricans. Similarly, the Puerto Rican supplement included 559 students of Puerto Rican descent.

Chinese. The Chinese supplement included 500 students of Chinese descent.

Disabled. The disability supplement included 589 students who had difficulty using their limbs for the year prior to the survey and, as a result, used a cane, wheelchair, orthopedic shoes, an artificial limb, or some other mechanical aid. This sample was drawn for Wave I in-home only.

Students eligible for each of these supplemental samples were identified using data from the In-school Questionnaire. For example, students were classified as Chinese for sampling purposes if they indicated on

the In-school Questionnaire that they were Chinese; the relevant question was a follow-up to the main race question (question 6 in the In-school Questionnaire).

In the selection of the non-genetic supplements, systematic samples were taken from among the eligible students. Prior to selection, the list of eligible students was sorted by school and sex-grade stratum. As with the core sample, the non-genetic supplement samples were selected separately in the initial group of 119 sample schools and in the later group of 13 schools; the data from the initial batch of schools was used to determine the sampling rates applied in the final batch of 13. (The Chinese supplement was an exception; it was selected after data from all 132 schools were available.) The selection of the supplements was carried out independently of the selection of the core sample. As a result, cases could be (and were) selected for both the non-genetic supplement and core samples. For the same reason, a few cases were selected for more than one of the supplemental samples.

Sample Selection: Genetic Supplements

Four additional supplemental samples were selected based on the sibling relationships in which the student was involved:

> **Twins**. Any student who identified him or herself as a twin (in response to question 23 on the In-school Questionnaire) was included in the twin supplement; in addition, previously unreported twins discovered during the Wave I in-home data collection were added to the supplemental sample at that time. Altogether, this supplement included 2,658 students who claimed to be twins (or triplets). In addition, the basic twin sample was augmented by 367 twins identified at a set of supplemental schools not used for any of the other samples.
>
> **Other siblings of twins**. This supplemental sample included 208 non-twin siblings of the twins in the twin supplemental sample. To be eligible, the non-twin sibling also had to be enrolled in grades 7 through 12 at the time of sample selection. This sample was drawn for Wave I in-home only.
>
> **Other full siblings**. The full-sibling supplement included 255 pairs of brothers, 258 pairs of sisters, and 307 brother-sister pairs. Full siblings were eligible for this sample if neither member of the pair was a twin and both were in grades 7 through 12.
>
> **Half siblings**. The half-sibling supplement included 497 pairs of half siblings in which both members of the pair were enrolled in grades 7 through 12.
>
> **Non-related**. This supplement included 491 adolescents enrolled in grades 7 through 12 and who do not share a biological mother or father. In addition, previously unreported non-related adolescents discovered during the Wave I in-home data collection were added to the supplemental sample at that time.

Eligibility for the other siblings of twins, full siblings, half siblings and non-related supplements was determined based on responses to the household grid in the In-school Questionnaire. Each student was asked to list in the grid any other household members in grades 7 through 12; for each person listed, the student was asked to indicate the person's sex and whether he or she shared the same biological mother and father as the student. Included in the non-related supplement were 394 adoptees. A question in the In-school Questionnaire asked whether the adolescent was adopted and a follow-up question asked whether he or she lived with either biological parent. Adoptees not living with their biological parents were eligible for the non-related supplemental sample.

Twins and adoptees in the non-related supplement were selected with certainty—that is, anyone identified as eligible for the supplemental sample was selected for it. The selection process for the siblings of twins was also quite simple. Within the first group of 119 schools, a systematic sample of 180 twin siblings was

selected from the 590 twin siblings that were identified. Within the final 13 schools, all 28 of the students that were identified as siblings of twins were included in the twin-sibling supplement.

The sampling unit for the half-sibling and full-sibling samples was a pair of students. Based on the data from the In-school Questionnaire, pairs of half siblings and pairs of full siblings were identified. In the case of the full siblings, the pairs were further classified into brother-brother, sister-sister, and brother-sister pairs. The selection of the half-sibling pairs followed a simple design: 451 of the 2,443 pairs identified in the first 119 schools and all of the 146 pairs identified in the final 13 schools were selected for the supplement. The selection of pairs in the initial set of schools was carried out systematically.

The selection of the full-sibling pairs was a bit more complicated. The original sampling plan had assumed that a sufficient number of full-sibling pairs—250 of each type— would be identified from among the core, PAIRS, and non-genetic supplements so that no supplemental selection of full siblings would be needed. For the brother-sister pairs, that proved to be the case; some 307 pairs were identified from among the members of the other samples. For the brother-brother and sister-sister pairs, however, additional sampling was necessary. A total of 132 brother-brother pairs were added to the sample among 4,923 pairs of male full siblings identified; this produced an overall total of 255 brother-brother pairs. Similarly, 104 additional pairs of sisters were added to the full-sibling supplement, yielding a total of 258 sister-sister pairs. Altogether, the full-sibling pairs encompassed a total of 1,575 persons (many of whom were also selected for other samples).

Except for the full-sibling sample, the genetic supplements were selected without regard to the other samples. (With the full siblings, pairs were identified from among members of the other samples.) As a result, members of each supplement could also have been selected for one or more of the non-genetic supplements or for the PAIRS or core samples.

Table 2 shows the number of students selected for each sample.

–

Table 2. Sample Sizes for Wave I In-Home I Samples

Sample	Selections
Core	16,044
PAIRS	3,350
Nongenetic Supplements	
High Education Blacks	1,318
Cubans	571
Puerto Ricans	559
Chinese	500
Disabled	589
Genetic Supplements	
Twins	2,658
Twins at non-sample schools	367
Siblings of twins	208
Other full siblings:	
Members of brother-brother pairs	514
Members of sister-sister pairs	508
Members of brother-sister pairs	589

Non-related adolescents	491
Adoptees	394
Half siblings	1,177

Weighting of the Samples

Weights are used for various purposes in the analysis of survey data. The weights compensate for differences in the selection probabilities for different cases. For example, because large schools were more likely to be selected for the Add Health sample than small schools, Cubans attending large schools were more likely to be included in the Cuban supplement than Cubans attending small schools. Unless weights are applied, the results will reflect this over-representation of Cubans from large high schools and junior highs. In addition, the weights can compensate for differences in response rates across different subgroups of a sample. Even if the sample as selected is perfectly representative of the population, the sample of completed cases may depart from the population because of nonresponse. Weights can help offset the effects of differential rates of nonresponse across different subgroups of the sample. Finally, if the weights incorporate appropriate adjustments, they can help compensate for chance fluctuations of the sample from known population totals.

The next sections discuss the basic weighting procedures used with each of the Wave I in-home samples and then describe the specific issues that arose in connection with the core sample, the non-genetic supplements, and the genetic supplements.

Basic Weighting Procedure

Each of the Wave I in-home samples was weighted in four main steps. In the first step, a preliminary school weight was calculated to compensate for differences among the schools in their probability of selection for the Add Health sample; this weight (W_1) was just the reciprocal of the school's selection probability. In the second step, this preliminary school weight was adjusted for school ineligibility and school nonresponse (among the junior highs). In addition, the adjusted school weight (W_2) brought estimates based on the school sample (for example, regarding the distribution of high schools by region) into line with population figures derived from the Quality Education Data (QED) file used to select sample high schools. In the third weighting step, an initial student-level weight was calculated. This weight (W_3) compensated for differences in student selection probabilities across schools and across grades and sexes within a school. The final step of the weighting process attempted to compensate for nonresponse to the Wave I In-home Questionnaire. The final weight for each sample (W_4) thus incorporates adjustments for both school-level and student-level nonresponse. Each of the samples followed this four-step process, with occasional variations required.

To illustrate the weighting process, the development of the core weights is discussed in detail. For the remaining samples, departures from this basic procedure are highlighted.

Core Sample Weights

The school weights used in developing the core sample weights were the same ones used for the in-school weights.

High school weights. The sample high schools were selected with probabilities proportional to the school's enrollment according to the QED file. Thus, for high school i, the selection probability was:

$$P_i = \frac{MOS_i}{MOS/80},$$

where MOS_i refers to the number of students in the high school and MOS to the total number of students in all the QED high schools; 80 is the number of sample high schools. The initial weights (W_{1i}) for the high schools were the inverses of these selection probabilities:

$$W_{1i} = 1/P_i .$$

Because of sampling error, the sum of the initial high school weights did not match the total number of schools on the QED file used as the sampling frame. The initial school weights were poststratified to match the QED totals for each region (Northeast, North Central, South, and West). The poststratification adjustment factor (PS) for region c was just the ratio of the total count from the frame for that region (T_c) to the sum of the weights for the sample schools in the region:

$$PS_c = \frac{T_c}{\sum_{i=1}^{m_c} W_{1ic}},$$

where m_c is the number of sample schools in region c. The adjusted weights were just the product of the initial weight and the poststratification adjustment factor. The QED file contained a number of ineligible high schools (schools without an 11th grade); ineligible schools that were selected for the sample were replaced. The high school weights were adjusted for ineligibility, by multiplying them by the weighted eligibility rate observed within the first 80 schools sampled.

Feeder school weights. The selection probability for feeder i depended on the selection probability of the high schools to which it sent graduating students:

$$P_{i'} = \sum_{i=1}^{M_{i'}} P_i \frac{n_{i'i}}{n_i},$$

in which the summation is over the M_i' high schools to which feeder school i′ sends graduating students, P_i is the selection probability for high school i receiving those graduates, and $n_{i'i}/n_i$ is the proportion of high school i's entering class that came from feeder school i′. This expression is difficult to calculate because it requires data from all the high schools to which the feeder sent graduates. Instead, an approximation was used that assumed that the size of the entering class for a given high school was roughly a constant fraction (f) of its measure of size (that is, of its estimated total enrollment). In that case, the feeder's selection probability becomes

$$P_{i'} \approx \sum_{i=1}^{M_i'} P_i \frac{n_{i'i}}{f\,MOS_i} = \frac{80}{f\,MOS} \sum_{i=1}^{M_{i'}} n_{i'i} .$$

The feeder school's selection probability is just a constant times the size of its graduating class (that is, the sum across all linked high schools of the number of graduates it sends to each one). The size of the graduating class was estimated using the roster data from the junior high; the size of the highest grade in the feeder school was used to estimate the total number of graduates. The constant of proportionality (f in the equations) was estimated separately for high schools spanning grades 9 through 12 and those spanning grades 10 through 12.

The selection probability for the feeder school was far simpler to calculate when all of the graduates from the junior high were expected to enroll in a single high school. In such cases, the feeder school's selection probability becomes the product of the high school's selection probability and the proportion of its entering class that came from the feeder school. Even further simplification was possible when all of the graduates from the feeder are expected to enter the high school and all of the entering students at the high school are expected to come from the feeder. (In several communities, the high school and feeder school were in fact a single combined school and thus formed such a one-to-one relationship.) In such cases, the feeder's selection probability is just P_i, the selection probability for the linked high school.

Once a preliminary feeder school weight was calculated, it was adjusted to incorporate the same adjustment factors used with the high school weights. In addition, because in a few cases a feeder school could not be recruited to take part in the study, the feeder school weights were adjusted for school-level nonresponse.

Student-level weights. Within a given school, all students within the same grade-sex cell had the same selection probability for the core sample. The initial student-level weights (W_3) for members of the core sample were the product of the inverse of the within-school selection probability times the final school-level weight (W_2):

$$W_{3ijk} = W_{2i} \frac{N_{ij}}{n_{ij}}, \quad (1)$$

in which N_{ij} represents the total number of students in stratum i at school j and n_{ij} represents the number of core selections from that stratum and school.

The next step in computing the core weights was to adjust for nonresponse to the Wave I In-home Questionnaire. For each sex-school combination, a weighted response rate for members of the core sample was computed. The final student weight (W_4) was the product of the inverse of the response rate and the unadjusted student weight (W_3).

In principle, W_4 should yield unbiased estimates. The estimates based on the weights may, however, be improved if the weights are poststratified to agree with independent population estimates. In addition, there were a few extreme weights that increased the overall variability of the weights themselves and of the statistics derived from them. We therefore "trimmed" the weights (imposing a maximum value of 5,000 and reducing larger weights to that value). In addition, we calculated a second set of core weights in which the sum of the trimmed weights for each grade-sex-race (black vs. non-black) cell was brought into line with Census Bureau population estimates for that cell.

Weights for the Non-Genetic Supplemental Samples

The weighting procedure for the non-genetic supplements differed from that used in weighting the core sample in four respects.

First, because the students eligible for each supplement were identified using In-school Questionnaire data, no supplemental selections could be made at roster-only schools (which did not permit in-school data collection) or at schools where identifying information was deleted from the In-school Questionnaires. To compensate for this form of school-level nonresponse, the supplement weights were inflated by an additional school-level nonresponse adjustment factor. The adjustment factor was simply the inverse of the weighted proportion of schools that permitted the In-school Questionnaires to be completed with identifying information.

Individual students who failed to complete In-school Questionnaires at schools that permitted the questionnaires to be administered were also excluded from the supplemental samples. Thus, a second change to the procedures used in weighting the core sample was needed; the final student-level supplement weights also incorporated a nonresponse adjustment factor to compensate for student-level nonresponse to the In-school Questionnaire. These two changes from the procedures used in weighting the core sample meant that the preliminary student-level weight was inflated by three nonresponse factors:

$$W_{4hij} = \frac{W_{3hij}}{R_1 \, R_{2i} \, R_{3h}} .$$

The first factor, R_1, represented the proportion of schools permitting in-school data collection. The second one, R_{2i}, represented the student-level response rate to the In-school Questionnaire at school i. The third, R_{3h}, represented the Wave I in-home response rate for members of the supplemental sample in adjustment cell h. (Table 3 below lists the adjustment cells used in calculating these Wave I in-home nonresponse adjustments for each supplemental sample.)

The third departure from the core procedures involved the student selection probabilities (conditional on the selection of the schools) for each supplement. These probabilities (denoted by P_2) did not vary by school, but were constant. In a few cases, different probabilities were used within the initial 119 and the final 13 schools. Thus, the preliminary student weights (W_3) were calculated using the equation below in place of Equation (1) given earlier:

$$W_{3ij} = \frac{W_{2i}}{P_2} , \qquad (2)$$

in which W_{3ij} is the preliminary student weight for the supplement and P_2 is the relevant conditional selection probability. Table 3 lists the conditional selection probabilities used for selecting students for each of the non-genetic supplements.

Table 3. Nonresponse Adjustment Cells and Conditional Selection Probabilities, by Nongenetic Supplement

Supplemental Sample	Conditional Probability (P_2)	Weighting Cells
High Education Blacks	.3077	Male; Female
Puerto Ricans	.4435	Male; Female
Chinese	.6878	Male; Female
Cubans	.5376	Male; Female
Disabled	.6516 (initial 119 schools)	Male; Female
	1.000 (final 13 schools)	

The final departure from the core procedures involved trimming and poststratification. We did not trim the weights of any of the non-genetic supplements and, because external benchmarks were not available for these populations, we did not poststratify the weights for these samples either.

Weights for the Genetic Supplemental Samples

The procedures for weighting the genetic supplements closely paralleled those for the non-genetic supplements. The weights generally included the extra nonresponse adjustments for school- and student-level nonresponse to the In-school Questionnaire (since these data were needed to determine whether students were eligible for selection into the genetic supplements). In addition, no poststratification adjustments were attempted with the genetic supplement weights. The main differences in the procedures used with these supplements and those used in weighting the non-genetic supplements involved the calculation of the conditional student (or pair) selection probabilities—that is, the value of P_2 (see Equation [2] above) used in calculating the preliminary student weight. The procedures for calculating these probabilities differed across the different genetic supplements.

Adoptees and twin siblings. The adoptees and twin siblings were selected as individuals with a constant conditional selection probability. That probability was 1.0 for the adoptees (all of them were included in the Wave I in-home sample); for the twin siblings, the probability was .1999 for twin siblings identified in the first set of sample schools and 1.0 for those identified in the final 13.

Twins. If either twin was identified at a sample school, the pair was selected for the sample. The conditional selection probability for the twins was thus 1.0. When both twins attended the same school, the pair's selection probability was simply the school's selection probability and the preliminary pair-level weight was just the school's weight. When the twins attended different schools (i and i′), however, the pair's selection probability was more difficult to compute. If one of the twins attended a non-sample school, it was impossible to calculate the selection probability for that school without gathering additional data. Instead, based on the assumption that the other school had a selection probability equal to that of the sample school, the pair's selection probability was approximated by:

$$P_{ii'jj'} \approx 2P_i - P_i^2,$$

where P_i is the selection probability for the sample school.

Twins were identified both from In-school Questionnaire data and from the roster itself (via a string-matching algorithm that examined last names and dates of birth). Because In-school Questionnaire data were not needed to identify pairs of twins, the twin weights were not adjusted for In-school Questionnaire nonresponse. Twin pairs were considered respondents to the Wave I In-home Questionnaire only if both twins completed that questionnaire.

The additional pairs of twins selected from the special augmentation schools are available for analysis but were not weighted.

Half siblings. The half siblings were selected as a sample of pairs of students rather than of individual students. A pair could be identified if either half sibling in the pair attended one of the sample schools (and completed an In-school Questionnaire). If both members of the pair attended the same school (school i), then the selection probability for the pair was straightforward; it was the product of the school's selection probability and the subsampling rate ($P_{2,HS}$) among eligible pairs of half siblings. The value of $P_{2,HS}$ was set at .1846 (451/2443) in the first group of sample schools and 1.0 (146/146) in the final 13.

The selection probability for the pair was harder to calculate when the two half siblings attended different schools. With such half siblings, the pair could have been selected if the school attended by either one was selected for the sample. If the two schools did not form a high school-feeder school pair, the selection probability for the pair could be estimated using the same approximation used with the twins:

$$P_{ii'jj'} \approx (2P_i - P_i^2)\, P_{2,HS}, \quad (3a)$$

where P_i is the selection probability for the sample school and $P_{2,HS}$ is the selection rate for pairs of half siblings. If the two schools consisted of a high school and an associated feeder school, the pair-level probability became even more complicated to estimate, because the feeder school and high school had linked selection probabilities. The probability that the high school, the feeder, or both would be selected was the sum of the selection probabilities for the two schools less their joint probability of selection (that is, the probability that both would be selected). The following approximation was used to estimate the joint selection probability:

$$P_{ii'jj'} \approx \left(\frac{n_{i'i}}{n_i} + \frac{n_{i'} - n_{i'i}}{fMOS/80}\right) P_i\, P_{2,HS}, \quad (3b)$$

in which n_i represents the total number of students graduating from junior high i′ and $n_{ii'}$ represents the number of those graduates who enroll in high school i. Regardless of which approximation was used, to correct for the effects of the presence of ineligible schools on the frame, the final school weight (W_{2i}) was used in place of the inverse of the school probability in calculating the pair-level weight.

The correction for school-level nonresponse in the in-school data collection also depended on whether the two members of the pair attended the same or different schools. If R_1 denotes the overall school-level response rate to the In-school Questionnaire data collection and R_{2i} denotes the student-level response rate at school i, then the adjusted pair-level weight for pairs in which both members attended school i was:

$$W_{3ijj'} = \frac{W_{3ijj'}}{R_1(2R_{2i} - R_{2i}^2)}.$$

The first term in the denominator reflects the proportion of schools from which in-school data were successfully obtained; the second term reflects the proportion of pairs in which at least one of the two students completed an In-school Questionnaire (given the successful administration of In-school Questionnaires at the school). When the two members of the pair attended different schools, the pair-level weight adjusted for nonresponse to the In-school Questionnaire was:

$$W_{3ii'jj'} = \frac{W_{3ii'jj'}}{(2R_1 - R_1)\,[1-(1-R_{2i})\,(1-R_{2i'})]}.$$

As with the twins, a half-sibling pair was considered a respondent to the Wave I in-home data collection only when both members of the pair completed that questionnaire.

The full-sibling sample. The bulk of the full-sibling pairs came into the sample as a byproduct of the other samples. In addition to these "serendipitous" pairs, extra brother-brother and sister-sister pairs were added to meet sample-size targets for each full-sibling group. As a result, each brother-brother and sister-sister pair had two routes for entering the sample. Both siblings could have been selected for one of the other samples or the pair could have been selected as one of the supplemental pairs. The calculation of the full-sibling weights required estimating the probabilities for each of those routes.

For the pair to be selected as a byproduct of other sampling, both members of the pair had to be selected into one of the other samples. Rather than attempting to calculate the combined probabilities for each student across every sample, the weighting reflected only the student's selection probability for the core or PAIRS sample and for any non-genetic supplement for which he or she was eligible. If the student was not eligible for any supplement, the combined probability (P^*_{ij}) was identical to his or her selection probability for the core or PAIRS sample. For students at any of the saturation schools (including the PAIRS schools), in which all students were selected for Wave I in-home interviews, the students' combined probability was simply the inverse of the final school weight.

The pair's joint selection probability ($P_{1ijj'}$) depended on whether both siblings attended the same school or different schools. When both siblings (subscripted as j and j′) attended the same school (school i), their joint probability of selection was:

$$P_{1ijj'} = P^*_{ij} P^*_{ij'} W_{2i},$$

in which W_{2i} is the final school weight. If the two siblings attended different schools that did not make up a high school-feeder pair, then their joint probability of selection was:

$$P_{1ijj'} = P^*_{ij} P^*_{ij'}.$$

Finally, if one of the siblings attended a high school (school i) and the other its selected feeder (school i), the joint probability was:

$$P_{1ii'jj'} = \frac{n_{ii'}}{n_i W_{2i}} \left(P^*_{1ii'jj'} W_{2i}\right) \left(P^*_{1ii'jj'} W_{2i'}\right),$$

in which the first term on the right side of the equation represents the joint probability of selecting the high school and the sample feeder and the terms in parentheses represent the conditional probability of selecting each sibling given that his or her school was selected.

After the serendipitous pairs were formed, a supplemental selection of sister-sister and brother-brother pairs was taken and the extra full-sibling pairs were added to the sample. The selection procedure was exactly analogous to that for the half-sibling pairs (a list of unselected pairs was formed and a random selection was made from that list) so that the same methods can be used to estimate the pair's selection probability ($P_{2ii'jj'}$); the relevant formulas are given in Equations (3a) and (3b) above.

The pair's overall probability was just the sum of the probabilities for the two routes through which the pair might have been selected. The initial full-sibling pair weight was the inverse of this sum:

$$W_{3ii'jj'} = \frac{1}{P_{1ii'jj'} + P_{2ii'jj'}}.$$

The nonresponse adjustments to this preliminary weight parallel those for the half-sibling pair weights. For a pair of full siblings to be identified, one or the other member of the pair had to complete an In-school Questionnaire. Thus, the initial weights had to be corrected for pairs omitted because neither member completed an In-school Questionnaire. Pairs could be missed due to school- or student-level nonresponse to the In-school Questionnaire. As with the half siblings, the correction for

nonparticipation in the in-school data collection differed depending on whether the two members of the pair attended the same or different schools. In addition, the final weight incorporated an adjustment for nonresponse to the Wave I In-home Questionnaire. A pair of siblings was treated as responding only if both members completed the Wave I In-home Questionnaire.

A number of additional pairs of siblings were identified from the Wave I in-home data; these added pairs were available for analysis but were not weighted.

Combined Weight

We computed a final weight that allows cases from all of the samples to be combined. In theory, there are several methods for developing this combined weight. One method involves calculating the probability that each case would be selected for at least one of the Wave I in-home samples. Because some students were eligible for as many as four samples, this approach seemed intractably complex and potentially error-prone. We used a simpler method, based on a standard multiplicity approach, in which the weights for each case are summed across the different samples the case was selected for and this sum is divided by the number of samples the case was eligible for. This method also yields unbiased estimates for totals and means. Aside from its computational simplicity, the multiplicity approach has a second advantage—it incorporates relevant nonresponse and post-stratification adjustments automatically since it uses existing weights rather than starting with the selection probabilities.

Under the multiplicity approach, the combined weight (W^*_j) assigned to case j would be:

$$W_j^* = \frac{\sum^{S} W_{ij}}{s_j} \quad , \qquad (4)$$

where s_j is the number of samples for which case j was eligible and W_{ij} is the weight of that case for sample i. (W_{ij} would be treated as zero if the case was ineligible for sample i or was eligible but not selected for it.) The summation is across all the samples (including the core and PAIRS samples, and the High Education Black, Cuban, Puerto Rican, Chinese, disabled, twins, sibling of twins, full sibling, half sibling, and nonrelative supplements).

Three of the genetic supplement samples originally consisted of pairs (the twins, full siblings, and half siblings); these pairs were counted as respondents only if both members of the pair completed the in-home interview. Since the combined weight applies to individuals, nonresponse adjustments were recalculated for these three samples. The new nonresponse adjustment was based on whether the individual completed the in-home data collection. If $W_{ijj'}$ is the unadjusted pair-level weight for the pair consisting of students j and j′ and R'_i is the individual-level response rate for that sample (the weighted proportion of half siblings who completed the in-home data collection), then the new weight for person j for sample i would be:

$$W_{ij} = \frac{W_{ijj'}}{R'_i} \quad .$$

Instead of the original nonresponse adjustment, which compensated for pair-level nonresponse, the new nonresponse adjustment is used, which compensates for nonresponse by individuals. The value of W_{ij} replaces the original pair-level weight in Equation (4).

As with the core weight, the combined weights were trimmed to eliminate extreme values (weights greater than 6,000) and poststratified to agree with Census Bureau estimates of the size of each grade-sex-race (black vs. non-black) subpopulation.

Design and Implementation of the Wave II In-Home Sample

This section describes the design and procedures used for selecting adolescents for the Wave I in-home sample. It also provides information on the calculation of sample weights, and procedures used to adjust sample weights for nonresponse.

Selection of the Wave I In-Home Samples

For Wave I in-home data collection, 27,559 students were selected. This total includes a core sample, consisting of 16,044 students enrolled in grades 7 through 12 at the time of sample selection; in addition, it includes all of the students (a total of 3,350) in grades 7 through 12 at two schools—the PAIRS sample schools—where saturation samples were selected for Wave I in-home follow up. Besides the core and PAIRS samples, two groups of supplemental samples were selected for Wave I in-home data collection. One group of supplemental samples—the non-genetic supplements—included students in various ethnic categories, as well as students identified as disabled from the In-school Questionnaire data. The second group of supplemental samples—the genetic supplements—consisted of individual students and pairs of students in various types of sibling relationships. This section describes the procedures for selecting these samples in Wave I; the next section discusses the differences between the Wave I and Wave II samples. Procedures for weighting the Wave II In-home Questionnaire data from each of the samples groups are reviewed in the final section.

Sample Selection: Core and PAIRS Samples

The core sample consisted of roughly equal-sized samples drawn from 12 student-level strata; the strata were formed by cross-classifying students by their sex and grade. Using the roster and In-school Questionnaire data, each student was classified into a grade-by-sex cell. In some cases, missing grade or sex information had to be imputed to allow classification of each student by stratum.

Sample Selection: Non-Genetic Supplements

Eligibility for the non-genetic supplemental samples was determined by race/ethnicity and by disability status. As with the core sample, sample selection for each of the supplemental samples was restricted to students enrolled in grades 7 through 12 in one of the sample schools at the time the sample was selected. A total of five non-genetic supplemental samples were selected:

High Education Blacks. This supplement included black students, either of whose parents were college graduates.

Cubans. The Cuban supplement included students of Cuban descent.

Puerto Ricans. Similarly, the Puerto Rican supplement included students of Puerto Rican descent.

Chinese. The Chinese supplement included students of Chinese descent.

Disabled. The disability supplement included students who had difficulty using their limbs for the year prior to the survey and, as a result, used a cane, wheelchair, orthopedic shoes, an artificial limb, or some other mechanical aid. This sample was drawn for Wave I in-home only.

Students eligible for each of these supplemental samples were identified using data from the In-school Questionnaire. For example, students were classified as Chinese for sampling purposes if they indicated on the In-school Questionnaire that they were Chinese; the relevant question was a follow-up to the main race question (question 6 in the In-school Questionnaire).

In the selection of the non-genetic supplements, systematic samples were taken from among the eligible students. Prior to selection, the list of eligible students was sorted by school and sex-grade stratum. The selection of the supplements was carried out independently of the selection of the core sample. As a result, cases could be (and were) selected for both the non-genetic supplement and core samples. For the same reason, a few cases were selected for more than one of the supplemental samples.

Sample Selection: Genetic Supplements

Five additional supplemental samples were selected based on the sibling relationships in which the student was involved:

> **Twins**. Any student who identified him or herself as a twin (in response to question 23 on the In-school Questionnaire) was included in the twin supplement; in addition, previously unreported twins discovered during the Wave I in-home data collection were added to the supplemental sample at that time. In addition, the basic twin sample was augmented by twins identified at a set of supplemental schools not used for any of the other samples.
>
> **Other siblings of twins**. This supplemental sample included non-twin siblings of the twins in the twin supplemental sample. To be eligible, the non-twin sibling also had to be enrolled in grades 7 through 12 at the time of sample selection. This sample was drawn for Wave I in-home only.
>
> **Other full siblings**. The full-sibling supplement included pairs of brothers, pairs of sisters, and brother-sister pairs. Full siblings were eligible for this sample if neither member of the pair was a twin and both were in the 7th through 12th grade.
>
> **Half siblings**. The half-sibling supplement included pairs of half siblings in which both members of the pair were enrolled in grades 7 through 12.
>
> **Non-related.** This supplement included adolescents in which neither member shared a biological mother or father and both members were enrolled in grades 7 through 12. In addition, previously unreported non-related adolescents discovered during the Wave I in-home data collection were added to the supplemental sample at that time.

Eligibility for the other siblings of twins, full siblings, half siblings and non-related supplements was determined based on responses to the household grid in the In-school Questionnaire. Each student was asked to list in the grid any other household members in grades 7 through 12; for each person listed, the student was asked to indicate the person's sex and whether he or she shared the same biological mother and father as the student. For Wave II, the non-related supplemental group was composed primarily of adolescents who had been added to the sample during Wave I in-home. Also included in the non-related group were adoptees. A question in the In-school Questionnaire asked whether the adolescent was adopted and a follow-up question asked whether he or she lived with either biological parent. Adoptees not living with their biological parents were eligible for the non-related supplemental sample.

Twins and adoptees in the non-related supplement were selected with certainty—that is, anyone identified as eligible for the supplemental sample was selected for it. The sampling unit for the half-sibling and full-sibling samples was a pair of students. Based on the data from the In-school Questionnaire, pairs of half siblings and pairs of full siblings were identified.

Except for the full-sibling sample, the genetic supplements were selected without regard to the other samples. (With the full siblings, pairs were identified from among members of the other samples.) As a result, members of each supplement could also have been selected for one or more of the non-genetic supplements or for the PAIRS or core samples.

For further discussion on the sample selection for the Wave I in-home component, refer to "Selection of the In-Home Sample" in *The Prospective Longitudinal Study of Adolescent Health: In-Home Wave I Final Report.*

Wave II In-Home Sample

The Wave II in-home sample was comprised primarily of adolescents who participated in the first wave of the in-home component. In total, 17,913 adolescents were retained for Wave II. Following is a description of key changes between the Wave I and Wave II samples.

Wave I 12th Graders. Twelfth graders, who composed one-sixth of the Wave I sample, were not retained for Wave II. The exception to this were adolescents who were part of genetic pairs who were kept in the Wave II in-home sample.

Disabled and Siblings of Twins. These two supplemental groups were not retained for Wave II in-home.

Sample Reclassification. For Wave II, some sample members were categorized into different sample groups than selected into for Wave I. However, for weighting purposes, Wave I sample classifications were used.

Table 4 shows the number of students selected for each sample.

Table 4. Sizes for Wave II In-Home Samples

Sample	Selections[a]
Core	10,547
PAIRS	2,049
Nongenetic Supplements	
High Education Blacks	1,340
Cubans	420
Puerto Ricans	546
Chinese	344
Genetic Supplements	
Twins	2,315
Twins at non-sample schools	
Other full siblings:	
Members of brother-brother pairs	616
Members of sister-sister pairs	673
Members of brother-sister pairs	904
Non-related adolescents	1,156
Adoptee	112
Half siblings	965

[a]Figures represent Wave II sample classifications.

Screening for Siblings

In addition to the changes noted above, selected households were screened to determine if there were any adolescents in grades 7 through 12 not related to adolescents selected into the Wave I in-home sample or who were selected adolescents' half siblings, full siblings, or twins. Once identified, these adolescents were added to the appropriate supplemental sample group for Wave II in-home. The majority of adolescents who were thought to be half siblings or twins were found to be ineligible to participate, and were categorized as ineligible.

Weighting of the Samples

Weights are used for various purposes in the analysis of survey data. The weights are needed to compensate for differences in the selection probabilities for different cases. The members of the different Wave II samples did not necessarily have identical selection probabilities. Unless weights are applied, projections to the population they are intended to represent will be biased. The weights can also partly compensate for differences in response rates across different subgroups of a sample. Even if the sample as selected is perfectly representative of the population, the sample of completed cases may depart from the population because of nonresponse. Weights can help offset the effects of differential rates of nonresponse across different subgroups of the sample. Finally, if the weights incorporate appropriate adjustments, they can help compensate for chance fluctuations of the sample from known population totals.

The same basic procedure was used in weighting each of the Wave II samples; it is described in the following section. It describes the specific issues that arose in weighting the Wave II core sample, non-genetic supplements, and the genetic supplements.

Basic Weighting Procedure

For each sample, the corresponding Wave I in-home weights were adjusted to compensate for any additional Wave II nonresponse. The final Wave I in-home weights had already incorporated the necessary adjustments to compensate for differences in the selection probabilities of the different sample members and for subgroup differences in in-school and Wave I in-home nonresponse. Let W_{1i} denote the final Wave I in-home weight for a member of sample i (e.g., High Education Blacks); then the Wave II weight for that sample (W_{2i}) was calculated as:

$$W_{2i} = W_{1i} / R_{2i} , \qquad [1]$$

where R_{2i} was the response rate for that sample in Wave II. The response rate was calculated using W_{1i} to give a weighted estimate. This procedure allows researchers to focus on the subset of cases who completed both the Wave I and Wave II interviews. The weights thus facilitate longitudinal analyses (as well as analyses that utilize only Wave II data).

With the core sample and the non-genetic supplements (the Cubans, High Education Blacks, Chinese, and Puerto Ricans), the unit for the weight was the individual student. For the samples of twins, full siblings, and half siblings, the unit for the weight was a pair of students. (In addition, a weight for each twin as an individual was also created.) Respondents were classified as a pair if both members completed the Wave II interview. This same approach was used in weighting the Wave I pair data.

Core and Non-Genetic Sample Weights

Core sample. As in Wave I, two core sample weights were calculated. One was based on the Wave I core weight that incorporated an adjustment to census population figures for each grade-sex-race combination; the other was based on the Wave I core weight that omitted this adjustment. Both Wave II core weights included adjustments for Wave II nonresponse. Separate nonresponse adjustments (i.e., values for $1/R_{2i}$ in Equation (1) above) were calculated within each school by sex combination. To avoid extreme adjustment factors, cells were collapsed across sex when there were fewer than 20

respondents. This occurred in 16 of the schools.

Non-genetic samples. Wave II weights were computed for the Puerto Rican, Cuban, High Education Black, and Chinese supplements. Table 5 shows the overall size of each supplements and the cells used in computing nonresponse adjustments for each one. In each case, attempts were made to calculate separate response rates for each grade. In some cases, adjacent grades were combined to avoid extreme or unstable adjustments.

Table 5. Wave II Non-Genetic Supplements: Sample Sizes and Adjustment Cells

Supplement	Wave II Sample Size[a]	Nonresponse Adjustment Cells
Puerto Ricans	351	Grades 7; 8; 9; 10; 11 to 12
Cubans	333	Grades 7 to 9; 10; 11 to 12
High Education Blacks	687	Grades 7; 8; 9; 10; 11 to 12
Chinese	246	Grades 7 to 8; 9; 10; 11 to 12

[a]Figures reflect Wave I sample classifications.

Weights for the Genetic Supplemental Samples

Wave II weights were also calculated for pairs of twins, half siblings, and full siblings. The same procedure was used in calculating the pair weights as in calculating the weights for individuals—the Wave I weight was adjusted for Wave II nonresponse as described in Equation (1). A single nonresponse adjustment factor was calculated for both the twins and the half-sibling pairs, reflecting the overall weighted response rate for those two samples. For the full-sibling sample, separate adjustments were calculated for the brother-brother pairs, the sister-sister pairs, and the brother-sister pairs.

One additional weight was calculated. It can be applied to the *individuals* identified as twins who were included in the Wave II data collection. Separate nonresponse adjustments were computed for each sex by grade cell in weighting the individual twins.

Combined Weight for all the Samples

A final Wave II weight was developed that can be used to analyze data from all of the samples at once. As with the other Wave II weights, this weight is based on the corresponding Wave I weight and incorporates an adjustment for Wave II nonresponse. The Wave I combined weight was derived by summing the weights for each case across the different samples for which he or she was selected; this sum was then divided by the total number of samples for which the case was eligible. The Wave II nonresponse adjustment was calculated separately for each sample school.