The Add Health Study:
Design and Accomplishments

Kathleen Mullan Harris*

Carolina Population Center
University of North Carolina at Chapel Hill

2013

Abstract: This document was prepared to be used by those interested in a ready reference for the design features and accomplishments of Add Health (the National Longitudinal Study of Adolescent Health). It provides a summary of features incorporated in the first fifteen years of completed work on the Add Health Study. The reader is referred to our web site (www.cpc.unc.edu/addhealth) for additional information, and for availability of data. Parts of this document may be incorporated into other documents for grant applications, papers for publication, or public presentations, without further permission. This document will be useful for those planning to use existing Add Health data.

*Kathleen Mullan Harris, PhD, is Principal Investigator and Director of Add Health (kathie_harris@unc.edu). Address correspondence to Add Health, Carolina Population Center, CB# 8120, University Square, 123 W. Franklin St., Chapel Hill, NC 27516.

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. Special acknowledgment is due Ronald R. Rindfuss and Barbara Entwisle for assistance in the original design.

www.onlinelaege.com

INTRODUCTION

The National Longitudinal Study of Adolescent Health (Add Health) is a longitudinal study of a nationally representative sample of adolescents in grades 7-12 in the United States in 1994-95 who have been followed through adolescence and the transition to adulthood with four in-home interviews. Add Health was developed in response to a mandate from the U.S. Congress to fund a study of adolescent health and was designed by a nation-wide team of multidisciplinary investigators from the social, behavioral, and biomedical sciences. The original purpose of the research program was to help explain the causes of adolescent health and health behavior with special emphasis on the effects of multiple contexts of adolescent life.

This document describes the design features of Add Health, for data collected in Waves I, II, III, and IV, and includes reports of data quality assessments for the most recent wave of data collection at Wave IV in 2008-09. We encourage researchers and students of social science, public health, human development, biomedical sciences, and related fields to explore the possibilities in this rich dataset.

Data Access and Use

From its inception, Add Health has promoted the research opportunities and use of its data within the broad social and biomedical communities, resulting in prolific research production with unparalleled disciplinary breadth. Add Health has been a pioneer in the secure release of confidential data with an enlightened dissemination strategy that has significantly multiplied the research contributions the study has made to science and public policy. Add Health has no proprietary period; Add Health users around the world have access to new waves of data the same day as program investigators.

Add Health has become a national and global data resource for over 10,000 Add Health researchers, most of whom work in North America, but many who work with Add Health data in South America, Europe, Asia and Australia. Add Health researchers have obtained nearly 600 independently funded research grants and have produced nearly 2,000 peer-reviewed research articles that appear in over 300 different journals (see publications list at ***http://www.cpc.unc.edu/projects/addhealth/publications***). Since the Wave IV Program Project began in 2006, there were more than 1,300 publications. In addition, 19 books, 89 reports, 75 book chapters, and more than 450 masters' theses and doctoral dissertations are based on Add Health data.

Information about the Add Health study design, types of data, data documentation, codebooks, and access is available on our web site (www.cpc.unc.edu/addhealth), which averages over 56,000 hits per month (Harris et al. 2009). We also manage an interactive listserver of Add Health researchers who share important data discoveries, coding schemes, and measurement strategies, and discuss and solve data and analysis problems interactively. Attendance at the biennial Add Health Users Conference held at NIH and co-sponsored by NICHD regularly exceeds 150 participants, with an average of 60 paper presentations by Add Health researchers and didactic methodology sessions by Add Health staff. Paper abstracts from the recent 2012 Users Conference are posted at www.cpc.unc.edu/projects/addhealth/events. Access to and use of Add Health data are facilitated by the Add Health Dissemination Core.

THE ADD HEALTH SAMPLE AND DESIGN

Schools as Primary Sampling Units

Add Health used a school-based design. The primary sampling frame was derived from the Quality Education Database (QED). From this frame we selected a stratified sample of 80 high schools (defined as schools with an 11th grade and more than 30 students) with probability proportional to size. Schools were stratified by region, urbanicity, school type (public, private, parochial), ethnic mix, and size. For each high school selected, we identified and recruited one of its feeder schools (typically a middle school) with probability proportional to its student contribution to the high school, yielding one school pair in each of 80 different communities. More than 70 percent of the originally selected schools agreed to participate in the study. Replacement schools were selected within each stratum until an eligible school or school-pair was found. Overall, 79 percent of the schools that we contacted agreed to participate in the study. Because some schools spanned grades 7 to 12, we have 132 schools in our sample, each associated with one of 80 communities. School size varied from fewer than 100 students to more than 3,000 students. Our communities were located in urban, suburban, and rural areas of the country.

From September 1994 until April 1995, in-school questionnaires were administered to students in these schools. Each school administration occurred on a single day within one 45- to 60-minute class period. Add Health completed in-school questionnaires from over 90,000 students. The in-school questionnaire provided measurement on the school context, friendship networks, school activities, future expectations, and a variety of health conditions. An additional purpose of the school questionnaire was to identify and select special supplementary samples of individuals in rare but theoretically crucial categories. School administrators also completed a 30-minute questionnaire in the first and second waves of the study.

Core and Special Supplemental Samples

Add Health obtained rosters of all enrolled students in each school. From the union of students on school rosters and students not on rosters who completed in-school questionnaires, we chose a sample of adolescents for a 90-minute in-home interview constituting the Wave I in-home sample. To form a core sample, we stratified students in each school by grade and sex and randomly chose about 17 students from each strata to yield a total of approximately 200 adolescents from each pair of schools. (Students who did not participate in the in-school survey were eligible to be selected for participation in this sample.) The core in-home sample is essentially self-weighting, and provides a nationally representative sample of 12,105 American adolescents in grades 7 to 12. From answers provided on the in-school survey, we drew supplemental samples based on ethnicity (Cuban, Puerto Rican, and Chinese), genetic relatedness to siblings (twins, full sibs, half sibs, and unrelated adolescents living in the same household), adoption status, and disability. We also oversampled black adolescents with highly educated parents. The number of adolescents in each of these special samples who were interviewed at Wave I are shown in **Table 1** below. Note that individuals can be assigned to more than one group. For example, a Cuban twin pair would be counted as two of the 538 Cubans and one of the 767 twin pairs in Wave I.

Table 1. Case Counts in Add Health Wave I Data
(Notes: overlaps not removed; * numbers represent pairs of adolescents).

Core Sample	12,105
Cuban	538
Puerto Rican	633
Chinese	406
High-Education Black	1,547
Disabled	957
Full Sibling*	1,251
Half Sibling*	442
Non-related*	662
Adopted	560
Twin*	784
Saturated Sample	3,702

For two large schools and 14 small schools, interviews with all enrolled students were also attempted in Wave I. The two large schools were selected purposefully. One is predominantly white and is located in a small town; the other is characterized by marked ethnic heterogeneity and is located in a major metropolitan area. The 14 smaller schools are located in rural and urban areas, and both public and private schools are represented. We collected complete social network data in the saturated field-settings, providing unbiased and complete coverage of the social networks and romantic partnerships in which adolescents are embedded by generating a large number of romantic and friendship pairs for which both members of the pair have in-home interviews. The core sample plus the special samples produced a sample size of 20,745 adolescents in Wave I. Figure 1 (end of document) shows the Add Health study design for selecting the in-school and in-home Wave I samples. The Wave I in-home sample is the basis for all subsequent longitudinal follow-up interviews, and thus this innovative design remains a major strength of the longitudinal data as well.

Seventy-nine percent of all sampled students *in all of the groups* participated in Wave I of the in-home phase of the survey (20,745). A parent, usually the resident mother, also completed a 30-minute op-scan interviewer-assisted interview. Over 85 percent of the parents of participating adolescents completed the parental interview in the first wave. The parent questionnaire gathered data on such topics as heritable health conditions, marriage and marriage-like relationships, involvement in volunteer, civic, or school activities, health-related behaviors, education, employment, household income and economic assistance, parent-adolescent communication and interaction, the parent's familiarity with the adolescent's friends and friends' parents, and neighborhood characteristics.

In 1996, all adolescents in grades 7 through 11 in Wave I (plus 12th graders who were part of the genetic sample and the adopted sample) were followed up one year later for the Wave II in-home interview (N=14,738). We conducted the adolescent in-home interviews using audio-CASI technology (audio-computer assisted self interview) on laptop computers for sensitive health

status and health-risk behavior questions. Add Health was the first national study to use ACASI technology in an adolescent population. The use of ACASI and CASI techniques has been found to enhance the quality of self-reporting of sensitive and illegal information (Turner et al. 1998).

The school and Waves I and II in-home interviews constitute the adolescent period in Add Health and contain unique data about family context, school context, peer networks, spatial networks, and genetic pairs. The social context data, in particular, are unusual because we did not rely on self-reports to generate an image of an adolescent's world. Family context data come from parent questionnaires, from adolescent in-school and in-home questionnaires, and from interviews with additional adolescents living in the same household. For certain measures, such as parenting behaviors and parent-child relations, we have reports from both the child's and the parent's perspective (and in some cases from a sibling as well).

Contextual Data

School context data come from the administrator questionnaires (usually principals) who reported on school policies, the provision of health services, and other school characteristics. In addition, school context variables can be constructed by aggregating student responses from the in-school and in-home questionnaires, enabling researchers to describe schools with respect to their social demographic composition, the behaviors of their students, the health status of their students, and the attitudes of their students towards school.

Peer network data were obtained in the in-school questionnaire. Adolescents nominated their five best male and five best female friends from the school roster (using a unique id). Because nominated school friends also took the in-school interview, characteristics of respondents peer networks can be constructed by linking friends' data from the in-school questionnaire and constructing variables based on friends' actual responses. In the in-home Wave I and Wave II interviews, respondents nominated their best friend, as well as their romantic and sexual partners. If their friend or partner is also a member of the in-home sample, their data can be linked to construct friendship and partner contexts. In the 16 schools that were part of the "saturated" sample, all students in the school were also interviewed in the home. Complete friendship and sexual networks can therefore be constructed with these data.

Spatial data indicating the exact location of all households in the survey were collected using hand-held Global Positioning System (GPS) devices or recording actual addresses. These data make possible the interweaving of spatial and social networks, and the construction of community contexts. More than 2,500 attributes for community and neighborhood contexts at multiple spatial units of observation have been obtained and merged with the Wave I and Wave II survey data to describe the neighborhood and community contexts in which adolescents are embedded. Neighborhood and community data were gathered from a variety of sources, such as the U.S. Census, the Centers for Disease Control and Prevention, the National Center for Health Statistics, the Federal Bureau of Investigation, and the National Council of Churches.

Finally, the "genetic pairs data," based on more than 3,000 pairs of adolescents who have varying degrees of genetic relatedness (see Table 1), represent a fully articulated behavioral genetic design, and are unprecedented for a national study of this magnitude. These data

represent pairs of adolescents who took the exact same questionnaires, share the same home environment, and share, in most cases, the same school and neighborhood environment. Thus, from the outset, Add Health was designed to address biological contributions to health by permitting researchers to explore gene-environment interactions in relation to health and behavioral outcomes. In addition, the embedded genetic design created new opportunities for research on adolescents living in relatively rare family structures, such as blended step-families and surrogate-parent families, and research on adopted children with a remarkable sample of 560 adopted children in Wave I of Add Health. In all follow-up interviews, high priority has been placed on locating and re-interviewing pairs in the genetic sample to maintain the integrity of this sample for longitudinal research purposes.

The Transition to Adulthood: Wave III

With NICHD funding for a continuation of the program project, Add Health conducted a Wave III follow-up interview with original Wave I respondents as they entered the transition to adulthood. When adolescents finish high school, they enjoy greater independence and begin to explore new lifestyles. As a result, their social contexts change and their experiences broaden. Wave III data capture these expanding experiences by focusing on the multiple domains of young adult life that individuals enter during the transition to adulthood, and their well-being in these domains: labor market, higher education, relationships, parenting, civic participation, and community involvement. With the longitudinal data from adolescence, this third wave of in-home interviews allows researchers to map early trajectories out of adolescence in health, achievement, social relationships, and economic status and to document how adolescent experiences and behaviors are related to decisions, behavior, and health outcomes in the transition to adulthood. The fundamental purpose of this third follow-up was to understand how what happens in adolescence is linked to what happens in the transition to adulthood when adolescents begin to negotiate the social world on their own and develop their expectations and goals for their future adult roles.

Wave III data collection was conducted nationwide (including Hawaii and Alaska) between August 2001 and April 2002. Respondents were now aged 18-26 and in the midst of the transition to adulthood. Add Health completed interviews on 15,170 respondents at Wave III, resulting in a 76% response rate. (For more details, see Chantala et al. Wave III nonresponse report at ***http://www.cpc.unc.edu/projects/addhealth/data/guides/W3nonres.pdf***). In the interest of confidentiality, no paper questionnaires were used. As in earlier waves, data were recorded on laptop computers. For less sensitive material, the interviewer read the questions and entered the respondent's answers. For more sensitive material, the respondent entered his or her own answers in privacy. The average length of a complete interview was 134 minutes. The laptop interview took approximately 90 minutes and was immediately followed by the collection of biological specimens. Most interviews were conducted in respondents' homes.

In Wave III we continued to collect data on health and health related behavior that were measured at earlier waves, including repeated measures of diet, physical activity, access and use of health services, sexual behavior, contraception, sexually transmitted infections, pregnancy and childbearing, suicidal intentions and thoughts, mental health and depression, substance use and abuse, injury, delinquency, and violence. We again obtained physical measurements of height

and weight, and collected data on pubertal development, chronic and disabling conditions, and other forms of morbidity. Wave III contains new data specific to the late adolescent, young adulthood life stage on parent-child and sibling relations, contact with friends from high school, the role of mentors and mentoring relationships, personal income, wealth and debt, civic and political participation, children and parenting, involvement with the criminal justice system, and religion and spirituality. Extensive data were collected on relationships, including a complete history since Wave I and measures of relationship intimacy, quality, commitment, shared activities, length, exclusivity, and sexual, union, and fertility behaviors. Codebooks for all four waves of Add Health instruments can be downloaded from the Add Health website at http://www.cpc.unc.edu/projects/addhealth/codebooks.

Wave III Design Features

We incorporated several new design features into Wave III that tapped research topics particularly salient in the late adolescent-early adulthood life stage.

- Binge Sample: A special sub-sample of freshman and sophomores in 2- and 4-year colleges was chosen, along with a control group of non-college same-age peers, who were administered additional questions about binge drinking for an independently funded R01 grant on that topic.

- College Context: College names were recorded so that college context data can be linked and merged with respondent data for those in college.

- Biological Specimens: New data collection of biological specimens was included in Wave III. At the end of the Wave III interview, urine and saliva samples were collected for tests of HIV and curable STIs (sexually transmitted infections). Buccal cell DNA was also collected from a subsample of the genetic sample (full sibs and twins) for extraction, purification and subsequent genotyping. For more information, see the report on Wave III Biomarker Data Collection at http://www.cpc.unc.edu/projects/addhealth/data/guides/biomark.pdf.

- Several publications using these data have reported STD and HIV prevalence rates found in our national sample at Wave III (Miller et al. 2004, 2005; Morris et al. 2006).

- Couples Sample: Wave III contains a new "couples sample" in which Add Health respondents recruited their romantic partners to take the same Wave III interview as their Add Health partner. The couples sample is made up of slightly over 1,500 pairs of partners, with roughly 500 married couples, 500 cohabiting couples, and 500 dating couples.

For more details on the Add Health research design, design documents, available data sets, codebooks, and publications, please see Harris et al. 2009, available at http://www.cpc.unc.edu/projects/addhealth/design.

Additional Add Health Wave III Data

The high school transcripts of Add Health Wave III sample members and their partners in the

couples sample were collected in coordination with the Wave III fieldwork. This study, the Adolescent Health and Academic Achievement (AHAA), collected high school transcripts and other data from all Add Health high schools except two special education schools that did not maintain students' academic transcripts, and from approximately 1,400 additional schools where Add Health respondents last attended high school. Approximately 91% of Wave III respondents signed a valid transcript release form and high school transcripts were collected for most respondents (N= approximately 12,000). AHAA also collected detailed school information, course offerings, and school contextual measures from the schools last attended by the AHAA Wave III sample. The transcripts were coded using procedures designed for the National Educational Longitudinal Study (NELS) and the National Assessment of Educational Progress (NAEP). The AHAA data provide indicators of (1) educational achievement, (2) course taking patterns, (3) curricular exposure, and (4) educational contexts within and between schools. Constructed measures (e.g., course sequences, academic intensity, grade trajectories) have been merged with Add Health data and released to all Add Health users (see http://www.cpc.unc.edu/projects/addhealth/codebooks/wave3). (See http://www.prc.utexas.edu/ahaa/ for more information on AHAA.)

Add Health researcher Penny Gordon-Larsen at UNC developed a database of time-varying modifiable physical activity-related environmental factors for Waves I and III called ONEdata, "Obesity & Neighborhood Environment Database." Using a multidisciplinary approach that blends spatial analysis methodologies with traditional epidemiological methods, she has linked area-level data to the individual data in Add Health that will add community-level measures such as recreation facilities (public, private), transportation options, crime, land use, air pollution, walkability, climate, and cost of living. The environmental data come from US Geologic Survey, US Census, US Department of Labor Statistics, and others extant sources. These additional environmental data for Waves I and III were made publicly available in 2011 (see http://www.cpc.unc.edu/projects/onedata).

Social, Behavioral, and Biological Linkages in Health: Wave IV

In a second continuation of the Add Health program project, a fourth in-home interview was conducted in 2008 with the original Wave I respondents. The Wave IV study was designed as a follow-up of the nationally representative sample of adolescents first interviewed in 1994 and 1995. See Figure 2 (end of document) for the longitudinal design of Add Health from Wave I through Wave IV, showing interview years and sample sizes associated with each survey component. At Wave IV 15,701 original Add Health respondents were re-interviewed.

The scientific purpose of Wave IV is to study developmental and health trajectories across the life course of adolescence into young adulthood using an integrative approach that combines social, behavioral, and biomedical sciences in its research objectives, design, data collection, and analysis. At the time of the interview, the Wave IV participants were 24 to 32 years old[1] and settling into young adulthood. At the same time that the Add Health cohort was assuming adult roles and responsibilities, they were also developing crucial health habits and lifestyle choices that set pathways for their future adult health and well-being. At Wave IV, we administered a comprehensive personal interview that included physical measurements and biospecimen collection. By integrating biological information into models of health and human development,

[1] 52 respondents were 33-34 years old at the time of interview.

the Wave IV design stimulates interdisciplinary research teams that bridge the social and biomedical sciences (Zerhouni 2003).

Through Wave IV data collection, we obtained longitudinal survey data on the social, economic, psychological, and health circumstances of our respondents, as well as longitudinal geographic data. Several features of Wave IV data collection represented new directions in Add Health, including methods to obtain more objective measures of health status and health behavior to capture prevailing health concerns, and methods to obtain biological markers of future chronic health conditions and disease. Wave IV employed innovations in the collection of biological measures in a field setting on a large national sample that were both practical and groundbreaking. We collected DNA on the entire national sample and obtained indicators of cardiovascular health, metabolic syndrome, and immune functioning using noninvasive procedures. The combination of longitudinal social, behavioral, and environmental data collected over 10 years with new biological data expands the breadth of research questions that can be addressed in Add Health regarding pre-disease pathways, gene-environment interactions, the relationship between personal ties and health, factors that contribute to resilience and wellness, and environmental sources of health disparities.

Wave IV Data Collection

Wave IV data collection was carried out by RTI International under sub-contract to the University of North Carolina at Chapel Hill. Data collection was conducted nationwide in all 50 states from January 2008 to February 2009. All Wave IV protocols were pre-tested on 300 respondents in three states from April to June, 2007. We located 92.5% of the Wave IV sample and interviewed 80.3% of eligible sample members.

Survey data were collected using a 90-minute CAPI/CASI instrument: Less sensitive questionnaire sections were administered with the assistance of an interviewer (computer-assisted personal interview, or CAPI). More sensitive questionnaire sections were self-administered using CASI technology (computer-assisted self interview). Immediately following the 90-minute interview, interviewers took physical measurements and collected biological specimens and a medications log, which took around 30 minutes. Most interviews were conducted in respondents' homes.

The Wave IV response rate (80.3%) is an improvement over Wave III (77.4%) and is comparable with other national longitudinal studies that have even shorter intervals between interviews. For example, the 2009-10 round of the annual NLSY97 had an 84% retention rate, and the 2010 round of the biennial NLSY79 had a 75.9% retention rate. Moreover the Add Health response rate far exceeds other national studies with longer intervals between waves (e.g., last NSFH interview in 2001-2003 had a 55% response rate; MIDUS II 2004-2006 had a 75% retention rate); regional longitudinal studies (LA FANS 2006-2008 had a 62.5% response rate); and national cross-sectional studies (2006-2010 NSFG had a 77% response rate).[2]

2. https://www.nlsinfo.org/content/cohorts/nlsy97/intro-to-the-sample/retention-reasons-non-interview; http://www.nlsinfo.org/content/cohorts/nlsy79/intro-to-the-sample/retention-reasons-noninterview; www.ssc.wisc.edu/**nsfh**/**wave3**/**fieldreport**.doc; www.midus.wisc.edu/newsletter/MIDUS_Final.pdf;www.rand.org/content/dam/rand/pubs/working_papers/2012/RAND_WR240.20.pdf; www.cdc.gov/nchs/data/nhsr/nhsr049.pdf.

Consistent with previous waves, there was differential attrition by gender, race, and immigrant status, with higher response rates for female, white, and native-born respondents at Waves IV. Response rates also increased as parental education and socioeconomic status levels at Wave I increased. To investigate the effect of non-response on study estimates at Wave IV, we used demographic, behavioral, health and attitudinal variables from Wave I to measure the extent that differences between respondents and non-respondents introduce bias at Wave IV. Results indicated that total and relative bias is small in magnitude for nearly all measures after study estimates were adjusted with final sampling weights. One exception is cognitive ability whereby significantly more non-responders than respondents had very low verbal aptitude (measured by AHPVT—Add Health Picture Vocabulary Test score), but this could also be a statistical artifact of the low prevalence of all respondents (i.e., 2.5%). Overall, Wave IV non-response bias is negligible and the Wave IV sample adequately represents the same population surveyed at Wave I when final sampling weights are used to compute population estimates (see Brownstein et al. 2010, available at http://www.cpc.unc.edu/projects/addhealth/data/guides/W4_nonresponse.pdf).

Wave IV Survey content

In the development of the Wave IV instrument, we consulted with our expert advisors on the Survey Research Advisory Committee and with representatives of NIH Institutes who had contributed co-funding support to the Wave IV Program Project. The Wave IV survey maintained longitudinal elements from previous waves and added new questions and sections that were developmentally appropriate for the lives of young adults. Figure 3 (end of document), shows questionnaire content across waves of interviews in Add Health, reflecting different developmental stages of the cohort. Bolded text indicates new content. Questions that reflected continuity with earlier content included updates of sociodemographic transitions central to the movement into adulthood and information about continuity and change in multiple indicators of physical and mental health status and health care; in social, emotional, spiritual, and physical contexts, including contact and relationships with parents; in sexual patterns and reproductive health; and in risk taking, social, and antisocial behavior.

New data were collected in a number of domains, reflecting funded research needs and salient aspects of the young adult years of the 20s. Survey questions were expanded on educational transitions, economic status and financial resources and strains, sleep patterns and sleep quality, eating habits and nutrition, illnesses and medications, physical activities, emotional content and quality of current or most recent romantic/cohabiting/marriage relationships, and maltreatment during childhood by caregivers. An inventory of the "Big 5" personality dimensions was added, as were indicators of interpersonal and occupational stressors, loneliness, and attitudes about parenting. Expanded questions on substance addiction and dependency and items on intersections and balance between work and family responsibilities were also added. Finally, two memory tasks were added, supplementing earlier measures of oral vocabulary (AHPVT).

Wave IV collected information on the dates and circumstances of key life events occurring in young adulthood, including a complete marriage and cohabitation history, full pregnancy and fertility histories from both men and women, an educational history of dates of degrees and school attendance, contact with the criminal justice system, military service, and various employment events, including the date of first and current jobs, with respective information on

occupation, industry, wages, hours, and benefits. We also collected a residential history of state locations since Wave III and recent mobility, and updated US citizenship status when applicable.

Wave IV Survey Data Quality

Our complex survey instrument performed exceptionally well at Wave IV, yielding outstanding data quality. There were little missing data due to respondents' refusal to answer questions or responding "don't know." Across all items (every section) in the Wave IV survey instrument, the weighted mean percentage of "don't know" responses was 1.68 and the weighted mean percentage of refused answers was 0.74. Even with hundreds of skip instructions, there were only a few, minor instrument programming errors. In addition, consistency across responses was excellent. For example, of respondents who reported getting health insurance through work, only 1.62% did not report a current job (but did work in the past), and only 0.1% reported never having a job. In the entire sample, only 1.60% of respondents listed a personal income that is higher than the total household income they reported.

Further, information about romantic/sexual relationships and fertility history is challenging to capture and represented the most complex portions of the survey instrument. In the Wave IV interview respondents reported information about 30,263 relationships, 21,966 pregnancies, and14,749 live births. As one indicator of data quality we compared respondents' summary reports (e.g., how many times have you ever been married?) with counts generated from completing relationship and pregnancy history tables. We found that 97% matched on total number of pregnancies, 95% matched on total number live births, and 95% matched on total number of living children. Further, 93% of respondents matched on all 3 reports. We also gave respondents opportunity to confirm information provided such as birth dates of children. Only 0.68% of baby birth dates were missing after respondents had the opportunity to correct information provided earlier in the interview.

Wave IV Geographic Data

At Wave IV we collected geographic data in two forms: home addresses and latitude/longitude coordinates from Global Positioning System (GPS) devices. Information systems specialists who are part of the Geographic Information Systems (GIS) Spatial Core of the Carolina Population Center have cleaned and checked the validity of the Wave IV geocodes by comparing the geocodes from the latitude longitude measures to the geocodes obtained from coding the addresses and by spatially mapping respondent locations at Wave IV. The spatial group has also developed a database of the longitudinal geocodes for Waves I, II, III and IV by evaluating and comparing addresses, geocodes, and latitude longitude measures across waves for consistency with other Add Health data which allows us to track moves and construct distance mover variables between waves. These mover variables are available with the contextual data along with grouping variables that allow researchers to geographically aggregate participants by state, county, tract, and block group.

Ancillary studies are underway (see Appendix B) that use the Wave IV geocodes to add contextual data to the longitudinal environmental data in Add Health that currently contains over 8,500 variables. Using the Wave IV geocodes and data from the American Community Survey (ACS), contextual variables are being constructed that parallel the Add Health Wave III Census contextual data, including population statistics, race, educational attainment, labor force and employment status, occupation, income and poverty status, housing characteristics, and

industries at the state, county, tract, and block group levels. Working with scientists from the Environmental Protection Agency, the Wave III and IV latitude and longitude coordinates are being utilized to include locational data about air quality and toxicologically-based air pollutant groups. Ancillary geographic data have also been added at Waves I and III on physical attributes of the community such as recreation facilities (public, private), distance to facilities, transportation options (e.g., public transportation), crime, presence of sidewalks, climate, and economic factors including local prices.

Wave IV Biological Data

We expanded our biological data collection at Wave IV to include additional physical measurements and biospecimen collection to measure early markers of disease risk in the young adult cohort. Anthropometric data included longitudinal measures of weight, height, body mass index [BMI], and a new measure of waist circumference. Cardiovascular measures new to Add Health included blood pressures and pulse rate. Trained and certified interviewers also obtained whole blood spots via finger prick, then dried and shipped them to study laboratories for assay of a lipid panel (total cholesterol [TC], high-density lipoprotein cholesterol [HDL-C], total triglycerides [TG]); glucose; glycosylated hemoglobin (Hb_{A1c}); high sensitivity C-reactive protein (hsCRP); and Epstein-Barr virus (EBV) antibody. Salivary buccal cells were collected and genomic DNA extracted from them for genotyping (described below). In the pre-test, respondents also collected three consecutive samples of their own saliva during the day after the interview. Although the saliva was assayed for cortisol, low adherence to protocol and reliability precluded self-collection and assay of saliva for cortisol in the main study (see Halpern et al. 2011). Lastly, all prescription and select over-the-counter medications were inventoried at the time of the exam to help identify / classify cardiovascular diseases, interpret laboratory results, evaluate treatment patterns, their putative antecedents, temporal changes or consequences.

Given the size and geographic spread of the Wave IV sample, and the non-clinical setting of our interviews, we chose methods to collect biological data that were noninvasive, innovative, cost-efficient, and practical for population-level research (McDade et al. 2007). Trained and certified interviewers used a finger prick to obtain whole blood spots that were dried and shipped for laboratory analysis. Buccal cell DNA in saliva was collected for genotyping a set of genetic markers. Table 2 shows the biomarker measures and biological specimens used for the biological domains covered below.

Anthropometric Measures: Interviewers measured the weight, height and waist circumference of Wave IV participants according to standardized protocols. Please see Entzel et al. 2009, "Add Health Wave IV Documentation: Cardiovascular and Anthropometric Measures" (available at http://www.cpc.unc.edu/projects/addhealth/data/guides/Wave%20IV%20cardiovascular%20and%20anthropometric%20documentation%20110209.pdf) for additional details on the anthropometric protocols at Wave IV.

Cardiovascular Measures: Interviewers measured the systolic blood pressure, diastolic blood pressure, and pulse of Wave IV participants according to a standardized protocol. Please see Entzel et al. 2009, "Add Health Wave IV Documentation: Cardiovascular and Anthropometric Measures" (available at http://www.cpc.unc.edu/projects/addhealth/data/guides/Wave%20IV%20cardiovascular%20and%20anthropometric%20documentation%20110209.pdf) for additional details on the blood

pressure protocol and Nguyen et al. 2011 for reliability and validity analysis of blood pressure measures.

Metabolic Measures: Dried blood spots obtained from a finger prick were assayed for lipids, glucose, and glycosylated hemoglobin (HbA1c). Please see Whitsel et al. 2011, "Add Health Wave IV Documentation: Measures of Glucose Homeostasis" (available at http://www.cpc.unc.edu/projects/addhealth/data/guides/Glucose_HbA1c.pdf) for a description of the blood spot protocol and data quality analysis of metabolic measures.

Measures of Inflammation and Immune Function: High sensitivity C-reactive protein (hsCRP) and Epstein-Barr virus (EBV) were assayed in dried blood spots. Please see Whitsel et al. 2012, "Add Health Wave IV Documentation: Measures of Inflammation and Immune Function" (available at http://www.cpc.unc.edu/projects/addhealth/data/guides/add-health-wave-iv-documentation-measures-of-inflammation-and-immune-function) for data quality analysis of inflammation and immune function measures.

Genetic Measures: In collaboration with the Institute for Behavioral Genetics (IBG) in Boulder, CO, Add Health collected, extracted, quantified, and stored DNA samples from all respondents in Wave IV. Genotyping for a set of candidate genes was conducted at IBG. Please see Smolen et al. 2013, "Add Health Wave IV Documentation: Candidate Genes" (available at http://www.cpc.unc.edu/projects/addhealth/data/guides/DNA_documentation.pdf) for a description of the DNA collection protocols and details on the processing and genotyping of DNA specimens.

Medications Data: Interviewers collected data on respondent use of prescription and select over-the-counter (aspirin-containing and non-steroidal anti-inflammatory) medications. Please see Tabor and Whitsel, 2010, "Add Health Wave IV Documentation: Prescription Medication Use" (available at http://www.cpc.unc.edu/projects/addhealth/data/guides/medication_documentation.pdf) for details on the collection protocol and the therapeutic classification of prescription medication.

Compliance rates for the collection of biomarkers have always been high in Add Health because respondents express confidence in the rigorous security system Add Health maintains to ensure its original pledge of confidentiality to them. At Wave IV, over 99% of respondents agreed to anthropometric and blood pressure measurements. Biospecimen consent was two-tiered, respondents agreed to provide the biospecimen for a) currently planned Add Health Program Project research; and (b) for archival for future testing "related to long term health." Compliance was high for all specimens: 96% consented to DNA collection and 95% consented to blood spot collection for purposes of planned program project research; 80% agreed to archive their blood spots and 78% agreed to archive their DNA for future analysis. There were few race and ethnic differences in consent rates with the exception that Black and Asian respondents were somewhat less likely to agree to biospecimen archival than Hispanic and white respondents.

Wave IV Biological Data Quality: Intra-Individual Variation (IIV) Study

We embedded an innovative *Intra-Individual Variation (IIV) Study* at Wave IV to assess the short-term reliability of the anthropometric, cardiovascular, metabolic, inflammatory and immune measures among a race/ethnicity- and sex-stratified random sample of 100 Wave IV

participants. Each participant was examined twice, approximately one week apart, typically by the same interviewer, and at approximately the same time of day. The trained and certified interviewers collected anthropometric measures, cardiovascular measures, and dried whole blood spots following standard data collection protocols at each of the two examinations. Blood spots from both exams were shipped to study laboratories and processed by technicians masked to participant identity. A nested, random-effects model was used to partition the variance of each measure into its between-participant, between-visit, and within-visit components. Reliability was then computed as the ratio of the between-participant to total variance, i.e. an intra-class correlation coefficient (ICC). Reliability of the anthropometric measures, Hb_{A1c}, and EBV were uniformly high, with ICCs approaching unity (0.97-1.00). As found in other settings, ICCs for blood pressures, TG, and hsCRP were somewhat lower (0.70-0.81). ICCs for pulse rate, the remaining lipids, and glucose were comparatively low (0.31-0.47), in part due to post-prandial fluctuation. Add Health is the first major national social science study to include this innovation in its design and assessment of biological data collection.

DNA Data Quality

The quality of DNA data was assessed on a number of dimensions. The average yield of DNA was 33 ± 25 ng/µl (mean ± SD), and ranged from zero (37 samples) to nearly 400 ng/µl. Of the 15,140 respondent samples, only 2.45% provided less than three ng of DNA per µl and we were unable to obtain reliable genotypes from most of these. The best overall measurement of the quality of the DNA samples is their utility in genotype determinations. All of the Wave IV DNA samples were of excellent quality and have been used for the assessment of over 100,000 VNTR genotypes in the first genetic data release. Data for five polymorphisms, the dopamine transporter (DAT1), dopamine D4 receptor (DRD4), serotonin transporter-linked polymorphic region (5HTTLPR--both diallelic and triallelic), and Monoamine Oxidase A promoter (MAOA-uVNTR) have been analyzed in duplicate, and are currently available as part of the public-release Wave IV Add Health data. Genotyping is currently underway for the polymophisms, dopamine D5 receptor (DRD5), dopamine D2 receptor (DRD2), Catechol O-methyltransferase (COMT), and Monoamine Oxidase A STR (MAOA[GT]), and should be available for the next release of Add Health Wave IV data later in 2012. Finally, a panel of 48 "functional" SNPs with emphasis on dopamine and serotonin pathways will be completed this fall and released in the beginning of 2013. All of the Add Health Wave IV buccal DNA samples were preamplified with the Invitrogen (Carlsbad CA) REPLI-g® method and analyzed for the polymorphisms listed above. Results for genomic and whole-genome amplified DNA were identical. Please see Smolen et al. 2012, "Add Health Wave IV Documentation: Candidate Genes" (available at http://www.cpc.unc.edu/projects/addhealth/data/guides/DNA_documentation.pdf) for more details on the Wave IV genetic data.

This expansion of biological data at Wave IV builds on the original design of Add Health which included important features for understanding biological processes in health and developmental trajectories across the life course of young people, including an embedded genetic sample with more than 3,000 pairs of adolescents with varying biological resemblance (e.g., twins, full sibs, half sibs, and adolescents who grew up in the same household but have no biological relationship). Add Health continued to add biomarkers with testing of saliva and urine for sexually transmitted infections and HIV at Wave III, and biomarkers of cardiovascular health, metabolic processes, immune function, and inflammation at Wave IV. Add Health therefore has critical objective indicators of health status and disease markers in young adulthood, well before

chronic illness or its complications emerge in later adulthood. Figure 4 (end of document) shows the array of biological data collected in Add Health across all waves.

Genome Wide Association Study

A Genome Wide Association Study (GWAS) will be conducted on approximately 12,200 Wave IV saliva samples from respondents who agreed to archive their specimens for future research. Funding has been obtained to assay the majority of these specimens and genome-wide genotyping is underway. No results will be released until all samples have been assayed and their quality confirmed. When released, the genetic data and a subset of the phenotypic data will be available through dbGaP (database of Genotypes and Phenotypes).

Longitudinal Contextual Data in Add Health

With the collection of geocodes at all interview waves, there is a wealth of longitudinal environmental data in Add Health. The Wave IV geocodes are currently being cleaned and updated, but we expect to merge the same contextual data at multiple spatial units (e.g., census tracks, census block groups, counties and states) at Wave IV for a complete set of contextual measures across waves and key developmental stages of the Add health cohort. In Table 3, we present a sampling of the rich array of contextual data available in the Add Health data sets. Due to their large number, we aggregated a subset of the more than 2,600 contextual variables into the categories shown in the table, each of which subsumes as many as several dozen related but discrete variables. These contextual factors operate at one or more ecological levels, from micro to macro, and range from intimate to broad contexts that capture psychosocial factors, factors that affect health both narrowly- and broadly-defined, and elements of the built environment. The variables are also drawn from survey responses across four waves of Add Health, including Add Health adolescent respondents at Waves I and II, young adults at Waves III and IV, parents of Add Health respondents at Wave I, school administrators at Wave I, Add Health ancillary studies, and administrative data sets (e.g., US Census, Centers for Disease Control and Prevention, National Center for Health Statistics, Federal Bureau of Investigation, National Council of Churches, Common Core of Data, Private School Survey).

In sum, by all indicators, the Wave IV data collection was highly successful: data collection was completed on time and within budget; our 80.3% response rate was higher than it was in the previous wave and higher than most longitudinal cohort studies, despite a 6-7 year interval since we last interviewed our respondents; attrition was minimal and introduced negligible bias into population estimates based on the Wave IV sample; item non-response was extremely low and reliability high; innovative and non-invasive in-home collection of biological data and specimens resulted in extremely high compliance rates; the validity and reliability of derived biomarkers based on an embedded intra-individual study ranged from excellent to moderate, with the ability to calculate measurement error; DNA collection resulted in high yield and over 100,000 VNTR genotypes in the first genetic data release (with more to come); and Wave IV geocodes were added to the longitudinal geocode database to construct mover variables between waves and facilitate multiple Ancillary Studies that add new environmental data to Add Health. On all indicators, Wave IV improved over Wave III.

Availability of Data. Please see http://www.cpc.unc.edu/addhealth, for details of currently available data. Requests for information should be sent to addhealth@unc.edu.

LITERATURE CITED

McDade, T.W., S. Williams, and J. J Snodgrass. 2007. "What a Drop Can Do: Dried Blood Spots as a Minimally Invasive Method for Integrating Biomarkers into Population-Based Research." *Demography* 44:899-925.

Miller, W. C., H. Swygard, M. M. Hobbs, C.A. Ford, M. Morris, M. S. Handcock, J. L. Schmitz, M. S. Cohen, K. M. Harris, and J. R. Udry. 2005. "The Prevalence of Trichomoniasis in Young Adults in the United States." *Sexually Transmitted Diseases* 32(10):593-598.

Miller, W.C., C.A. Ford, M. Morris, M.S. Handcock, J.L. Schmitz, M.M. Hobbs, M.S. Cohen, K.M. Harris, and J.R. Udry. 2004. "The Prevalence of Chlamydial and Gonococcal Infection among Young Adults in the United States." *Journal of the American Medical Association* 291(18):2229-2236.

Morris, M., M. S. Handcock, W. C. Miller, C. A. Ford, J. L. Schmitz, M. M. Hobbs, M. S. Cohen, K. M. Harris, and J. R. Udry. 2006. "Prevalence of HIV Infection Among Young Adults in the U.S.: Results from the Add Health Study." *The American Journal of Public Health*. 96(6):1091-1097

Nguyen, Q. C., J. W. Tabor, P. P. Entzel, Y. Lau, C. Suchindran, J. M. Hussey, C. T. Halpern, K. M. Harris, and E. A. Whitsel. 2011. "Discordance in National Estimates of Hypertension Among Young Adults." *Epidemiology* 22(4):532-541.

Resnick, M.D., P.S. Bearman, R.W. Blum, K.E. Bauman, K.M. Harris, J. Jones, J. Tabor, T. Beuhring, R. Sieving, M. Shew, M. Ireland, L. Bearinger, and J.R. Udry. 1997. "Protecting Adolescents from Harm: Findings from the National Longitudinal Study of Adolescent Health." *Journal of the American Medical Association* 278(10):823-832.

Zerhouni, E. 2003. "The NIH Roadmap." *Science* 302:63, 64, 72.

Table 2. Add Health Wave IV Biomarkers		
Biological Measure	**Biological Specimen**	**Method of Measurement**
Anthropometric		
Weight [kg]	NA	Health-o-meter 844KL digital scale
Height [cm]	"	Steel measure tape & Carpenter's sq
BMI [kg/m^2]	"	Weight ÷ $Height^2$
Waist Circumference [cm]	"	SECA 200 tape measure
Arm Circumference [cm]	"	"
Cardiovascular		
SBP [mmHg]	NA	Microlife oscillometric BPM
DBP [mmHg]	"	"
Pulse [beats/min]	"	"
PP [mmHg]	"	SBP – DBP
MAP [mmHg]	"	(SBP + 2 × DBP) ÷ 3
Metabolic	Finger Prick	
TC [mg/dL]	Whole Blood	Blood Spot Assay
HDL [mg/dL]	"	"
TG [mg/dL]	"	"
LDL [mg/dL]	"	TC – HDL – (TG ÷ 5)
TC:HDL	"	TC ÷ HDL
Non-HDL [mg/dL]	"	TC – HDL
Hb_{A1c} [%]	"	Blood Spot Assay
Glucose	"	"
Inflammatory		
$_{hs}$CRP [mg/L]	"	Blood Spot Assay
Immune		
EBV [ELISA units]	"	Blood Spot Assay
Genetic		
DAT1	Buccal Cell DNA	VNTR and SNP Panel
DRD4	"	VNTR and SNP Panel
DRD2	"	VNTR and SNP Panel
5HTT	"	VNTR and SNP Panel
5-HT2A	"	VNTR and SNP Panel
MAOA-uVNTR	"	VNTR
MAOA[GT]n	"	STR
DRD5	"	STR
COMT	"	STR
IGF1	"	STR

Table 3. Longitudinal Contextual Data in Add Health	**Contexts**																			
	Family				**Dyadic relationships**				**Peer / Social networks**				**School / College / Workplace**				**Neighborhood / Community/State**			
	w1	w2	w3	w4	w1	w2	w3	w4	w1	w2	w3	w4	w1	w2	w3	w4	w1	w2	w3	w4
Social																				
Household composition *marital status; roster*	+	+	+	+	+	+	+	+	+	+			+				+	+	+	+
Relationships – parent, sibling, peer & partner	+	+	+	+	+	+	+	+	+	+	+	+	+							+
Economic/Work *income, work stressors, unemployment*	+	+	+	+			+	+					+	+	+	+	+	+	+	+
School/Education *school type; achievement*	+	+	+	+			+	+	+	+	+		+	+	+	+	+	+	+	+
Race/Ethnic/Sex composition *integration; saturation*	+	+	+	+	+	+	+	+	+				+	+	+		+	+	+	+
Legal *crime, welfare regulations*																				
Health																				
Healthcare facilities, prevention programs & utilization	+	+	+	+	+	+	+	+	+	+	+	+	+	+			+	+	+	+
Fertility, morbidity & mortality *STD incidence*											+		+				+	+	+	
Alcohol & Tobacco availability, prevention & control	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+		+	+	+	
Health behavior *Peer & parent substance use*	+	+	+	+	+	+	+		+	+	+		+	+						
Physical Environment																				
Natural environment *distance to parks; day length*	+		+		+		+		+		+		+				+		+	+
Built environment *urbanicity, street connectivity*	+		+		+		+		+		+		+				+	+	+	+
Air quality *pollution*	○		○		○		○		○		○		○				○	○	○	○
Housing type & quality	+	+	+	+	+	+	+	+	+	+	+	+	+	+			+	+	+	+

Notes. Content examples are given in italics in the first column. W1 = adolescent in-school & in-home, parent, & school administrator surveys; geocodes.
W2 = adolescent in-home, school administrator & geocodes; W3 = young adult in-home & partner sample surveys, geocodes, biomarkers; W4 = young adult in-home, biomarkers, geocodes. Additional variables from administrative datasets.
+ variable available; ○ planned variable construction.

Figure 1. Sampling Structure for Add Health

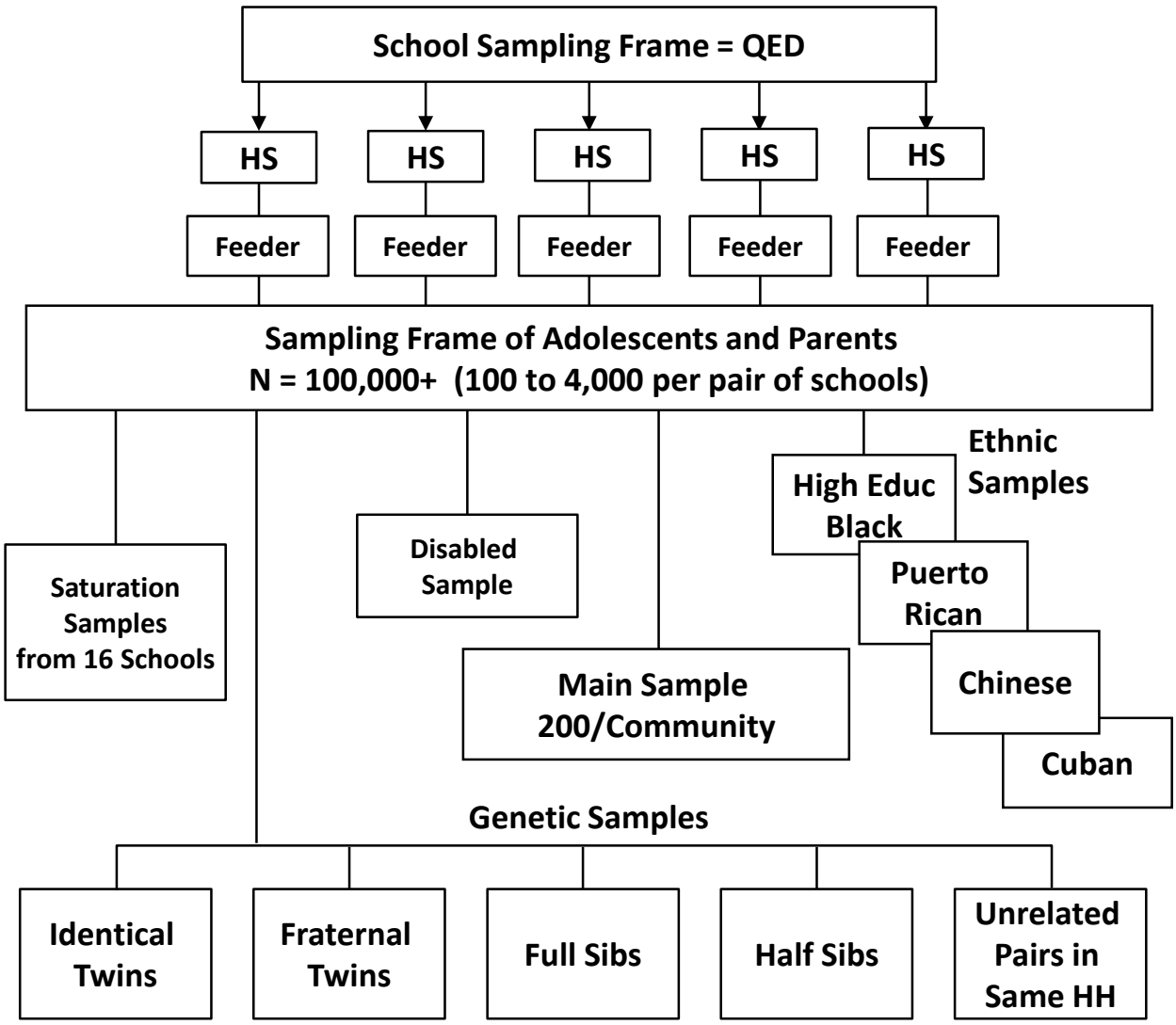

Figure 2. Add Health Longitudinal Design

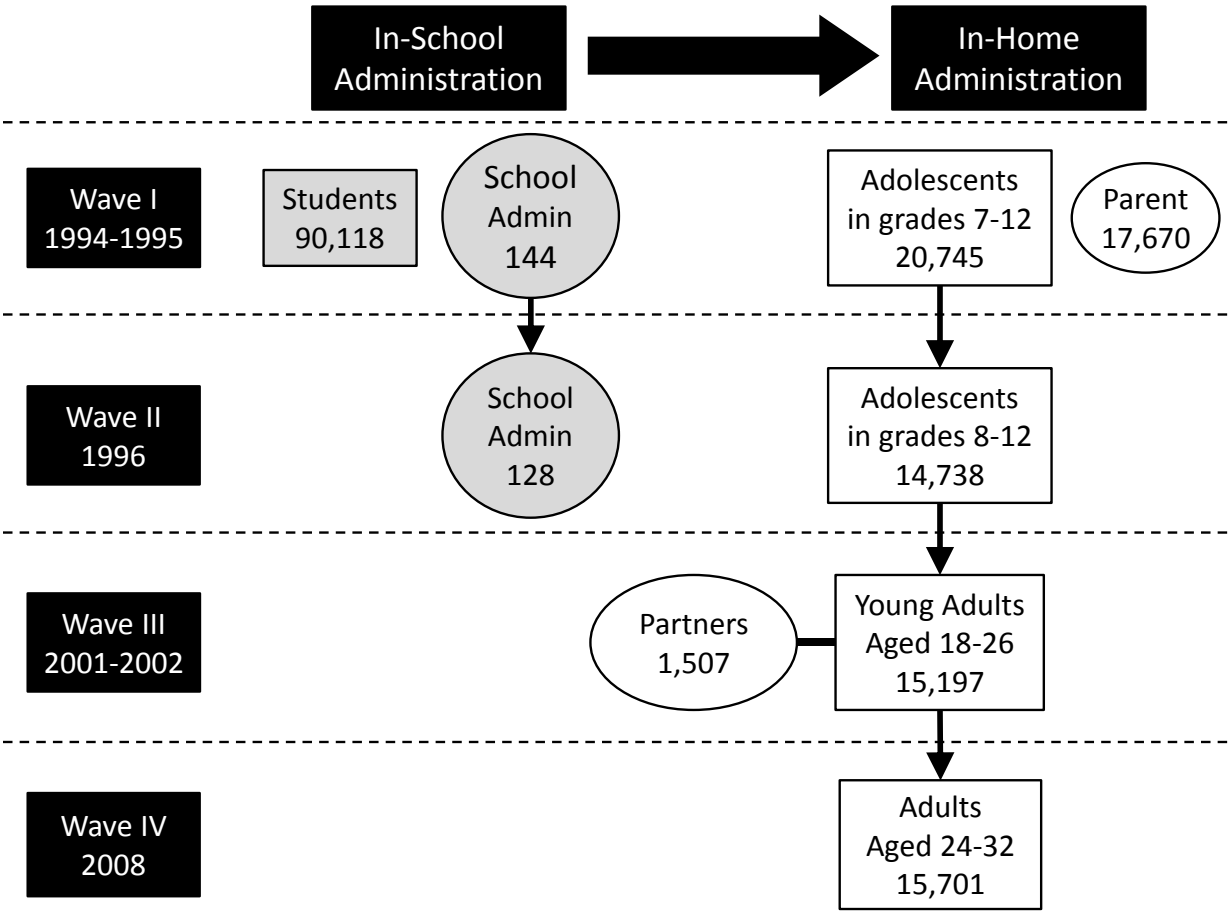

Figure 3. Questionnaire Content Across Waves

Waves I, II

- Demographic
- Family, siblings, friends
- Education, work
- Physical and mental health
- Daily activities and sleep
- Relationships
- Sexual, & fertility histories
- Substance use
- Delinquency and violence
- Attitudes, religion
- Economics, expectations
- Psychological, personality

Wave III

- Demographic
- Family, siblings, friends
- Education, work, ***military***
- Physical and mental health
- Daily activities and sleep
- Relationships
- Sexual, & fertility histories
- Substance use
- ***Involvmt w/criminal justice sys***
- Attitudes, religion
- Economics, expectations
- Psychological, personality
- ***Children and parenting***
- ***Civic participation***
- ***Gambling***
- ***Mentoring***

Wave IV

- Demographic
- Family, siblings, friends
- Educ, work, ***military (records)***
- Physical and mental health
- Daily activities and sleep
- Relationships
- Sexual, & fertility histories
- ***Substance use and abuse***
- Involvmt w/criminal justice sys
- ***Work attitudes and chars***, relig
- Economics, expectations
- ***Big 5 Personality, stressors***
- ***Children and parenting***
- Civic participation
- ***Cognitive function***
- ***Psychosocial factors***

Figure 4. Add Health Biological Data Across Waves

Adolescence	Young Adulthood	Adulthood
Wave I-II (Ages 12-20)	**Wave III (Ages 18-26)**	**Wave IV (Ages 24-32)**
Embedded genetic sample of 3,000 pairs		
Physical development		
Height, weight	Height, weight	Height, weight, waist
	STI tests (urine)	Metabolic (lipids, HbA1c, glucose
	HIV test (saliva)	Immune function (EBV)
	DNA (buccal cell)	Inflammation (CRP)
		Cardiovascular (BP, P)
		DNA (buccal cell)
		Medications

Add Health and WLS were included in the SSGAC GWAS of educational attainment. For each of these datasets, polygenic were computed using summary statistics from GWAS meta-analyses from which the target dataset for polygenic scoring was excluded. Within each datasetÂ 2. Harris KM. The add health study: Design and accomplishments. Chap Hill Carol Popul Cent Univ N C Chap Hill. 2013 National Longitudinal Study of Adolescent Health: Strategies to Perform a Design-Based Analysis Using the Add Health Data. Carolina Population Center, University of North Carolina at Chapel Hill: Chapel Hill, NC. Chapman, DP, Whitfield, CL, Felitti, VJ, Dube, SR, Edwards, VJ, Anda, RF (2004).Â The Add Health Study: Design and Accomplishments. Carolina Population Center, University of North Carolina at Chapel Hill: Chapel Hill, NC. Harris, LF, Roberts, SC, Biggs, MA, Rocca, CH, Foster, DG (2014). Perceived stress and emotional social support among women who are denied or receive abortions in the United States: a prospective cohort study. BMC Women's Health 14, 1â€“11. Herrenkohl, TI, Hong, S, Klika, JB, Herrenkohl, RC, Russo, MJ (2013). Using data from all four waves of Add Health study, this paper focuses on two outcomes: (1) depression and (2) levels of antidepressant use by race/ethnicity, immigrant generation, and linguistic acculturation levels during adulthood. Findings reveal that depression is prevalent among Mexican Americans, other Hispanics, and Asian Americans. Furthermore, Mexican Americans and Asian Americans have reported a lower level of antidepressant use than whites, with Asian Americans attaining the lowest level when immigrant generation, language acculturation levels, and other socioeconomic factors are h...Â The Add Health Study: Design and Accomplishments. Kathleen Mullan Harris. 2013. VIEW 1 EXCERPT. Add Health Wave IV Documentation: Prescription Medication Use.