National Longitudinal Study of Adolescent Health

Wave III
Education Data
Design and Implementation of the Adolescent Health and Academic Achievement Study

Chandra Muller, Jennifer Pearson, Catherine Riegle-Crumb,
Jennifer Harris Requejo, Kenneth A. Frank, Kathryn S. Schiller,
R. Kelly Raley, Amy G. Langenkamp, Sarah Crissey,
Anna Strassmann Mueller, Rebecca Callahan,
Lindsey Wilkinson, and Samuel Field

❖ ❖ ❖ ❖ ❖

Carolina Population Center
University of North Carolina at Chapel Hill

January 2007

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

WAVE III EDUCATION DATA
DESIGN AND IMPLEMENTATION OF THE ADOLESCENT HEALTH AND ACADEMIC ACHIEVEMENT STUDY

I. INTRODUCTION

The Adolescent Health and Academic Achievement Study (AHAA) is the educational component of the National Longitudinal Study of Adolescent Health (Add Health) – a school-based survey of a nationally representative sample of adolescents in 7th through 12th grade from 132 public, private, and parochial schools (Bearman, Jones, and Udry 1997). Add Health sample members were drawn from a random sample of 80 high schools with an 11th grade, stratified by region, urbanicity, size, type, and racial composition. Participants were also selected from 52 middle or elementary schools that "fed" into these 80 high schools. The feeder schools were randomly selected with probability proportional to the percentage of students they contributed to their respective high school's entering class. (See http://www.cpc.unc.edu/addhealth for full documentation of the Add Health school and student samples, and wave-based study framework).

While giving depth to the social context of the lives of adolescents, Add Health has limited information on the academic trajectory of youth. The focus of AHAA is to contribute this information by collecting official high school transcripts from all Wave III Add Health respondents, and critical contextual information about the schools these respondents last attended. The AHAA study was funded by a grant from the National Institute of Child Health and Human Development under grant R01 HD40428-02 to the Population Research Center, University of Texas at Austin; Chandra Muller (PI). Aspects of the curriculum component of AHAA, such as the coding of math and science textbooks, were funded by a grant from the National Science Foundation under grant REC-0126167 to the Population Research Center, University of Texas at Austin; Chandra Muller and Pedro Reyes (Co-PI). The official AHAA website (http://www.prc.utexas.edu/ahaa/) provides full documentation of the theoretical underpinnings of the each of the education components. This website also lists publications with AHAA and other current research-related activities as well as useful information for users.

The next sections outline the AHAA sample and study design, and provide an overview of the various AHAA components and accessible data files.

II. AHAA SAMPLE DESIGN

The AHAA study is a separate data collection designed to create an educational data set that can be studied independently or in relation to Add Health, facilitating an analysis of the course-taking patterns of students who attended Add Health high schools. Importantly, AHAA allows analyses of the relationships between Add Health indicators capturing adolescent social and health-related behaviors and outcomes, and

measurements of students' academic experiences as recorded on student transcripts. AHAA's structural compatibility with the 1987, 1990, 1994, 1998, and 2000 National Assessment of Educational Progress (NAEP) High School Transcript Studies (HSTS) also makes it possible to compare data collected for AHAA with data gathered for these other existing national-level data sets.

Several important distinctions of AHAA in contrast to the NAEP HSTS studies are worth noting. First, when possible, any coursework taken at an Add Health high school was identified as having been taken at that school, even if a student later transferred to another high school. This step allows analysts to place students in their respective Add Health schools prior to transferring, and minimizes the amount of missing data reported for school-specific information. Second, in contrast to these previously conducted high school transcripts studies, which simply demanded students' transcripts from schools without explicit student permission, the AHAA study received respondents' permission, which may have made the collection of transcripts more successful. Third, unlike the NAEP studies in which transcripts were collected only from high school graduates, the AHAA collected transcripts for students who did and did not complete a high school degree.

In sum, the AHAA design is similar to the approach used in the NAEP HSTS in these important ways:

1. The studies were minimally burdensome on participants. Study respondents were asked only to give the study permission to collect their transcripts. All data items came from extant documents (transcripts, course catalogs, syllabi, course records, etc.) routinely maintained by schools.

2. Schools were reimbursed for the costs associated with producing the transcripts and for providing course catalogs.

3. The coding scheme and processes used for the AHAA study were modified from those developed and implemented for the NAEP transcript studies.

4. Schools were asked to complete a School Information Form (SIF).

5. Respondent confidentiality was of paramount importance.

Major differences between AHAA and previously conducted NAEP transcript studies include:

1. Although all Add Health students were associated with the 80 schools from the original sample, they graduated from over 1,150 schools. AHAA was student based in that student records (for Add Health respondents) were collected from the final schools that students attended. For NAEP HSTS studies, the transcript collection was school based.

2. To have full descriptive information about the grade 9-12 courses taken by students included in AHAA and Add Health's longitudinal data collection, course catalogs for 8-10 years were needed for AHAA. Catalogs for only 4 years were usually needed for NAEP.

3. The NAEP HSTS studies processed data for, and have been focused on high school graduates only. The AHAA produced data for all study participants who attended high school, including those who did not graduate.

4. In NAEP HSTS, student courses, including transfer courses, were linked to the course catalog from the high school of graduation and/or a generic course catalog. In AHAA, students who graduated from a non-Add Health school had their transfer courses linked back when possible to the original Add Health school they initially attended.

5. In NAEP HSTS, participating students and schools were identified prior to data collection, and school materials and student transcripts could be collected during one request. In AHAA, the list of participating students and signed TRFs were gathered in a series of waves spanning around 5 months.

6. In AHAA, schools were asked to provide lists of the textbooks used in their math and science courses to enable assessment of students' curricular exposure in these key academic subjects. This information was not requested in previously conducted national-level transcript studies.

The following subsections describe in detail the student and school selection criteria for inclusion in AHAA.

2.1 Selection of Students – Eligibility Criteria

Students included in the AHAA study are those defined by the Research Triangle Institute (RTI) as participants in all three data collection waves of the National Longitudinal Study of Adolescent Health (Add Health). Identified romantic partners of these Add Health study participants were also selected to participate in AHAA. During Wave III of the Add Health data collection process, respondents were asked to sign a Transcript Release Form (TRF) authorizing Add Health to request official transcripts from the high schools they last attended. These students were also asked to provide the names of the cities and states of the original Add Health and any other high schools they attended during their high school careers. This original student list was modified throughout the study by dropping those who were later deemed ineligible for study inclusion. See the data collection procedures section below for an accounting of the reasons some students who signed TRFs did not ultimately qualify for participation in AHAA.

As previously noted, graduation from high school was not an eligibility requirement for study participation. AHAA collected transcripts from student respondents and their partners who did and did not earn high school diplomas.

2.2 Schools in the Sample

For a full accounting of how schools were randomly selected for inclusion in the National Longitudinal Study of Adolescent Health (Add Health), see the University of North Carolina's Add Health website http://www.cpc.unc.edu/addhealth. Because the data collection procedures used for AHAA were student-based in that transcripts were collected from the final school respondents (including selected partners) attended, 78 original Add Health high schools plus over 1,150 other high schools students last attended were included in AHAA. Transcripts were not collected from two original Add Health schools that served only special education students and did not keep transcript records.

III. AHAA DATA COLLECTION MATERIALS

The central goal of AHAA was to collect official high school transcripts from Wave III Add Health sample members. In order to code transcript data with a high degree of accuracy and validity from school to school, as well as ensure AHAA data compatibility with previous NAEP transcript studies, it was crucial to have as much information as possible about the content of the courses reported on student transcripts. Course content information was obtained from a variety of documents. Some schools and districts provided course catalogs that were highly detailed. Others submitted lists with more limited information. These materials aided in the identification of the level and appropriate Classification of Secondary School Curriculum (CSSC) code for each course taken by Add Health/AHAA students. (Detailed information about the application of CSSC codes to the transcripts collected for AHAA is provided in the Data Collection Procedures section.) See the AHAA website (http://www.prc.utexas.edu/ahaa/) for more information about CSSC codes.

Schools vary in the length of class periods, the amount of credit awarded for courses, and the number of credits required for graduation. The same procedures implemented for the NAEP transcript studies were used for AHAA in order to standardize and render comparable courses taken across and within participating Add Health and non-Add Health schools. These procedures included collecting and analyzing completed School Information Forms (SIFs), and interviewing school administrative staff.

A second goal of AHAA was to gather information about school characteristics. Specifically, details about school features such as the types of special programs offered to students were collected through SIFs and interviews with administrative staff.

AHAA also collected textbook lists and course syllabi of all math and science courses offered at participating schools. The purpose of these efforts was to enable a

textbook-based analysis of students' potential exposure to math and science curricula during their high school careers.

Full documentation of the types of data collection materials gathered from participating schools is provided in the school-level disposition section of the primary component of AHAA. The primary component also documents the specific types of information available per participating student.

3.1 Student Transcripts

Westat collected a range of transcripts per school. On average, 100-200 transcripts were collected from the original 78 Add Health high schools while 1-25 transcripts were typically collected from non-Add Health schools. These numbers matched our expectations given that approximately 100-200 students in each Add Health school participated in Wave III, and the number of students who transferred to non-Add Health schools was generally small.

3.2 Supplemental Materials from Add Health and Eligible Non-Add Health Schools

The supplemental materials requested from participating schools included:

- Sample transcripts and a transcript checklist identifying the location of information needed from the transcript;
- Course catalogs from 8-10 years (annotated as requested before leaving the school);
- The School Information Form (SIF) – Smaller schools and/or schools requiring more assistance were given an abbreviated SIF;
- The number of transcripts released, compared to the number requested from the school. Since the transcript requests were made in waves, transcripts not received in earlier waves were often re-requested from participating schools during later data collection efforts; and
- Other materials containing relevant information about school course offerings such as course records.

3.3 Textbook Lists and Course Syllabi

In addition to the supplemental information collected from schools to aid in coding transcripts, we collected a list of textbooks (title, author, publisher, date of publication) and course syllabi used for all current mathematics and science courses taught at most Add Health high schools and a number of the larger non-Add Health schools. (See the Curriculum Component for comprehensive information about this aspect of AHAA).

IV. AHAA DATA COLLECTION PROCEDURES

4.1 Overview

AHAA was conducted by the University of Texas at Austin. Westat and its subcontractor, Intelligent Automation, Inc. (IAI), were responsible for the collection, coding, data entry, and processing of high school transcripts and other school information for AHAA. Wave III Add Health survey respondents were asked to sign a TRF authorizing Add Health to request transcripts from the high schools they last attended; from August 2001 through June 2002 AHAA collected information about their high school last attended and transcripts for these consenting respondents. In addition, AHAA collected high school transcripts and other information for many of the Wave III partners.

Importantly, the transcript data collection procedure used for AHAA was student-based: transcripts were collected from the final high school respondents attended. This meant that transcripts were collected not just from the original Add Health schools, but from the more than 1,150 high schools Add Health respondents last attended. Transcripts were not collected from two original Add Health schools that served only special education students and did not keep transcript records; however, a few respondents who entered the Add Health sample through one of these two schools do have transcript records in the AHAA database because they last attended another school that did keep transcript records.

Information about schools and curriculum was collected from the last schools attended by AHAA sample members and used primarily for coding transcript materials. School administrators were asked to complete an SIF with information about school grading practices, policies, and any special programs available to students. Administrators were also asked to provide course catalogs, other supplemental materials pertaining to course descriptions to aid in the coding of student transcripts, and textbook lists and syllabi for all math and science courses offered by their schools. Over 900 schools completed the SIF and 139 provided textbook lists. The textbook lists compiled by schools served as the basis for AHAA's textbook-centered analysis of students' potential exposure to math and science curricular materials.

Finally, secondary data sources were attached to each school last attended by Wave III Add Health respondents even in the rare instance when transcripts were not collected from the school. Details about the institutional characteristics of schools gleaned from the 1990-91, 1993-94, 1994-95, and 1999-2000 Common Core of Data (CCD) surveys and the 1995-96 Private School Survey, for example, were linked to Add Health/AHAA schools. Likewise, measures tapping features of the local educational market, and capturing information about the attributes of the residential population of the districts and broader commuter zones surrounding Add Health/AHAA schools were derived from 1990 and 2000 census data. Additional sources of secondary data attached to Add Health/AHAA schools include the 2000 Office of Civil Rights data.

4.2 Initial Steps – Contacting Schools and Districts

Each school was initially mailed a package containing a letter describing the study and listing the toll-free hotline number to call with questions or for assistance, the list of student(s) for whom transcripts were requested, a copy of the student(s)' signed permission letter(s) (TRFs), a postage-paid return envelope, and transcript cost reimbursement forms.

Follow-up telephone calls were made two weeks after the initial mailing to make sure schools received the packages, and to answer any questions school administrators had about AHAA. Since many schools were closed and/or school administrators were not available when these telephone calls were made, repeated attempts were made throughout the summer and fall to contact and gain cooperation from the schools. In addition, numerous schools that originally had indicated they would cooperate did not send in supplemental materials and student transcripts for many months. Some of these schools faced difficulties submitting requested materials due to limited resources in the summer and early fall. Others underwent changes in administration and required new appeals for cooperation. On-site visits to schools/districts were occasionally necessary to collect requested materials. Decisions about and coordination of on-site visits were made on a case-by-case basis.

4.3 Field Operations – The Wave Approach

The planning and implementation of the field operations were adjusted to the timing and delivery of the signed TRFs. The field operations maintained flexibility to adapt to the needs of the project. For example, the specific request for data materials from each school was based upon (1) whether the school was an Add Health school; (2) the number of transcripts expected from the school; and, (3) the level of cooperation attained from the school. The establishment of contacts with AHAA schools/districts and decisions about on-site visits were also made as necessary throughout the duration of the study.

With the passage of time and changing school administrations, Add Health schools needed to be reminded periodically of their participation in the study. Some schools that had no recollection of AHAA required information regarding its significance before they would cooperate. Letters from involved agencies and academic centers helped elicit school cooperation in these cases.

After collecting transcripts, Westat field staff recorded the AHAA study ID number on each transcript and masked any identifying information (i.e. name, address, SSN). This is the same procedure used in NAEP and other studies to guarantee respondent confidentiality.

Field staff had a log of all students for whom transcripts were requested. This log matched the list of student names sent to the schools and had space for field staff to record information about the status of each transcript sought. For example, if a transcript

was incomplete or not locatable, this was noted in the log. Field staff used this list to report to Westat's home office on the status of field activities.

All materials received from participating schools were carefully receipted by clerical staff. This process proved to be complicated because transcripts were requested in waves and supplemental school materials such as catalogs, SIFs, and textbook information arrived at various times and in various forms such as hardcopy, electronic, or website download.

4.4 Training of Data Collectors/Field Staff

Westat followed the same procedures used for the NAEP HSTS studies when selecting and training field and administrative staff for AHAA. Many of the selected staff members had previous experience with the NAEP transcript studies.

New field staff were trained during a full-day in-person training session. Experienced staff completed a home-study package and a telephone review with the field manager. All field and office staff members were required to sign a confidentiality statement.

4.5 Obtaining School-Level Information

4.5.1 Course Catalogs

To have full descriptive information about the grade 9-12 courses taken by AHAA/Add Health students in core academic and elective courses, course catalogs for 8-10 years spanning the 1990s were requested from school administrators. (Four years of course catalogs were usually needed for NAEP HSTS).

Because of the importance of course catalogs for ensuring accurate and uniform course coding, we attempted to collect course catalogs from all original Add Health high schools and any non-Add Health high schools with significant numbers of study sample members. In total, catalogs were collected from all 78 Add Health schools and from approximately 140 non-Add Health schools. Initially 25 study participants attending a non-Add Health school functioned as the threshold for triggering a request for that school's course catalog (and math and science textbook lists). Given that transcripts and other pertinent information were received from schools in waves, it was not known until the end of the Wave III data collection which and how many schools met this criterion. Subsequent to the conclusion of Wave III, it became apparent that only a few schools qualified, and the student-threshold number was lowered to 5. As a result, school catalogs were requested from all non-Add Health schools where 5 or more AHAA participants last attended high school.

In the rare cases when school catalogs were unavailable, course catalogs were collected from relevant districts. "Model catalogs" that include courses most typically offered in different types of high schools (i.e. large, academically oriented schools; small

rural schools; culturally diverse urban schools; high-tech vocational schools) were constructed for some small non-Add Health schools that did not submit course catalogs.

Because study materials were collected from schools in waves during the data collection process, multiple course catalogs from the same year were sometimes collected from schools. Although these catalogs were coded separately, the results in terms of the application of appropriate CSSC codes to courses recorded on student transcripts were consistent across them.

It is important to note that course catalogs were modified to facilitate the coding of student transcripts, and do not have data on course offerings that would be appropriate for independent analytical use.

4.5.2 SIFs, Sample Transcripts, Transcript Checklists

As in the NAEP transcript studies, SIFs were collected in order to obtain useful information about participating schools' graduation requirements and the format of their course schedules. Given the high degree of variability in the amount of credit schools award for certain courses, the number of credits they require for graduation, and the length of class periods, the SIF forms facilitated the process of standardizing coursework taken across and within AHAA schools. (Complete documentation of the schools that did and did not submit completed SIFs is located in the primary data component of AHAA).

From the SIF we collected the following:

- School contact information;
- Course catalog checklist, listing all years for which catalogs were submitted;
- Carnegie unit conversion data so that credits from similar courses in different schools can be expressed in a standardized measure;
- Graduation requirements (including tests and credit requirements);
- Types of diplomas offered;
- Special programs offered to students (i.e. magnet programs, religious education, performing arts, special education, ESL);
- Title I eligibility status of school; and
- Transcript review checklist, which explains where data may be found on the transcripts from a school.

Sample transcripts and transcript checklists illustrating the location of pertinent information on the transcripts were requested from all participating schools to serve as templates for data entry personnel.

4.5.3 Textbook Lists

The textbook list collection efforts focused on Add Health schools and large non-Add Health schools, and on gathering materials that would sufficiently represent the math and science course-taking experiences of the various AHAA/Add Health student cohorts. Requests for textbook lists were frequently made as a follow-up to the transcript collection efforts. The heads of school mathematics and science departments often needed to be contacted in order to obtain the textbook information. To minimize the burden on school staff, and because of the relative stability of math and science course offerings over time and slow turnover rate of textbooks in these subjects, textbook lists were collected for the math and science courses offered by schools in the single academic calendar year 2001-02.

4.6 Identifying the Student Sample and Obtaining Transcripts

4.6.1 Response Rates and Attrition Analysis

The response rate for transcript collection efforts refers to the difference between (1) the in-scope Wave III participants who originally agreed to participate in AHAA, and (2) the resulting valid participants. Approximately 14,070 signed a valid TRF and from August 2001 through June 2002, AHAA collected high school transcripts for most of them (over 12,250). In addition, about 1,260 partners of Add Health respondents signed a valid TRF, and AHAA collected 940 of their high school transcripts.

The students (respondents and partners) who signed valid TRFs, but did not ultimately qualify for participation in AHAA were considered ineligible for the following reasons:

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

More information about the student response rates and the implications of student attrition from the study is provided in the Weighting and Estimation of Sampling Variance section at the end of this document.

4.6.2 Obstacles to Transcript Data Collection

There were some obstacles to data collection. The passage of time during the data collection process and changes in school administration meant that Add Health and non-Add Health high schools often needed to be reminded of their participation in AHAA. Some schools with no recollection of the study insisted we re-submit information regarding its significance. In these cases, letters from participating agencies and academic centers such as the University of Texas and Westat solicited further school cooperation. In-person visits were made on a case-by-case basis when necessary to encourage schools reluctant to release transcript records. School closures during the summer also slowed data collection. In addition, transcript records older than five years proved difficult for some schools to retrieve. Many schools lacked the resources to recover these older records, which were frequently stored at other sites such as district offices. This meant that permission to access and retrieve older transcript records had to be made at both the school and district levels. Lastly, in a few cases, schools merged or closed down completely, further complicating data collection efforts.

Many school administrations and faculty were overwhelmed with requests for transcripts, catalogs, textbook lists, and other supplemental materials. Incentives were provided to schools (on a school-by-school basis) and personal contacts were made with school staff to encourage participation. Communication with superintendents and district staff members was also often necessary, particularly when requests were made for older transcript records.

4.6.3 Receipt and Review of Transcript Data from Data Collectors

All transcripts received were receipted by a clerk who verified that the AHAA study ID number was on the transcripts and then deleted individually identifying information (i.e. name, address, SSN, etc.) if this had not been done at the school. This receipt clerk, like all staff with access to study materials, signed a confidentiality pledge as a condition of employment.

V. DATA PROCESSING PROCEDURES

5.1 Catalog and Transcript Coding Procedures

5.1.1 Training Catalog and Transcript Coders

Westat made every effort when selecting its coding staff to hire trained educators knowledgeable about secondary school curriculum. The goal of the training program was to ensure staff members were fully prepared to consistently and accurately code catalogs

and transcripts. Coders were trained using the same basic procedures and training materials used in the 2000 NAEP HSTS.

A significant portion of the coders' training was spent learning how to assess the content of courses. Specifically, the training emphasized the importance of evaluating a school's curriculum based on all available information. Catalog coders were also instructed to rely upon their knowledge and expertise as educators when determining the appropriate CSSC codes to assign to courses. To facilitate this, coders were encouraged to consult with one another during the coding process, particularly when a coder had an area of expertise (such as mathematics or vocational education), or an academic or practical background in a relevant area.

5.1.2 Course Coding Procedures

Course Coding is a multi-step procedure. Course titles appearing in each school's course catalog were first keyed into the Transcript Coding System (TCS) – a reliable software program designed for the 2000 NAEP HSTS study to facilitate efficient, accurate data entry, and to accommodate the variance that exists among transcripts and course catalogs. (Comprehensive documentation about the TCS is provided in the 2000 NAEP HSTS codebook). Westat fine-tuned the TCS for the AHAA study; these modifications are discussed above in the sample design section which describes how AHAA differs from previously conduced NAEP transcript studies.

The resulting list was then checked, verified, and revised as necessary by a catalog coder and supervisor. Using TCS, catalog coders then assigned a CSSC code to each course described in the catalogs. Following this step, another aspect of the TCS was used to match each course title appearing on a transcript from a school to a course title included in the school's course catalog. The TCS then assigned the linking school catalog identification to the transcript course title. TCS also prompted the catalog coder to set all flags pertaining to a course, such as those designating honors, AP/IB, or remedial status.

When school catalogs were not available, the best available source of information was used to obtain relevant course offering information. For instance, if district catalogs were available and applicable, the course descriptions in them were used to determine the content and CSSC code of courses listed as offered at participating schools. If no catalog or course list was available from the school, "model catalogs" with generic course listings and descriptions were appended to the course titles found on transcripts from these schools.

Transfer courses presented a challenging set of problems. In the NAEP HSTS, catalogs were not collected from transfer schools. However, because numerous Wave III respondents graduated from non-Add Health schools and many partners never attended an Add Health school, it was essential to collect catalogs from transfer schools to obtain accurate information about these students' academic records. If school catalogs were unavailable from these transfer schools, district or state catalogs were used as a proxy

means of determining the content and CSSC code of the courses listed on transcripts collected from them. Importantly, all transfer credits appearing on transcripts from students (respondents and partners) enrolled in non-Add Health schools were identified and linked back to the original Add Health school the student initially attended. Westat's TCS maintained a list of all transfer course titles and their associated codes. When an identical title was encountered, it was automatically linked to the same code. This technique had the advantage of ensuring consistency in coding transferred courses across schools and reducing the number of courses that had to be coded manually.

5.1.3 Summary of Transcript Coding Procedures

In order to provide high quality, accurate, and consistent coding for AHAA, the CSSC was used to code all courses appearing on student transcripts, as well as all courses offered at original Add Health schools and eligible non-Add Health schools. This coding scheme, which has been refined and standardized over the years, was used for High School and Beyond, the National Educational Longitudinal Study of 1988 (NELS), and all of the NAEP HSTS.

The specific procedures employed for coding transcripts were based on the highly successful procedures used in these previously conducted national-level transcript studies, and were designed to achieve the following objectives:

- Guaranteeing AHAA data compatibility with the data produced in the 1987, 1990, 1994, 1998, and 2000 NAEP High School Transcript Studies; and

- Ensuring a high degree of uniformity and accuracy across coders and from school to school.

The coding process emphasized the need to evaluate school curricula based on all available information (including course catalogs, completed SIFs, and other supplemental materials collected from schools), and to code each course based on its *content* rather than its *title*. As a result, appropriate CSSC codes were applied to courses following careful scrutiny of course descriptions from school catalogs. Courses taken in small non-Add Health high schools that did not submit catalogs were assigned CSSC codes based on Westat's knowledge of secondary school curriculum and extensive experience in coding high school transcripts. Specifically, a set of "model catalogs" that included courses most typically offered in different types of high schools (i.e. large academically oriented schools; small rural schools; culturally diverse urban schools; high-tech vocational schools) were developed and applied to code courses taken at these non-Add Health schools that appeared to be similar in content.

Significantly, the AHAA coding procedures were designed to maximize the amount of information useful for future analytic work, and involved coding two levels of transcript data – student characteristics and course characteristics:

- **Characteristics of the Student.** For example, graduation date, type of diploma, exit status;

- **Characteristics of the Courses that the Student Took.** For example, CSSC code; course title; grade; credit awarded; year and semester taken; whether it was a regular, honors, AP, IB, or remedial course; and whether it was a course designed exclusively for students with special needs.

Additional efforts were made, such as adopting the same formats and procedures for editing, coding, error resolution, and documentation, to ensure AHAA data compatibility with NAEP. For example, whenever possible, common variable names, labels, values, and missing codes were assigned. These efforts should enable analysts familiar with the NAEP HSTS to work easily with the AHAA and compare study cohorts with minimal effort.

5.2 Data Entry

5.2.1 Training Data Entry Personnel

Data entry staff underwent training in both the use of the TCS system for transcript data entry and in interpreting the extensive variety of transcript formats. Actual transcripts obtained in the 2000 NAEP HSTS were used as demonstration materials to illustrate different formats and types of information recorded on transcripts. Trainees also used these transcripts during practice exercises to gain familiarity and skill in using the TCS system. Besides extensive hands-on practice, training topics included consistent use of standard abbreviations and numerical entries; entry of courses by semester rather than by year (if they appear semester-by-semester on transcripts); and formal reporting of questions, problems, and issues to catalog coders for resolution.

The training provided to transcript data entry personnel was designed to ensure comparability between AHAA and previously conducted high school transcript studies, and stressed the importance of accuracy.

5.2.2 Data Entry Procedures

A major portion of the transcript coding process involved entering data from the transcripts into the TCS databases. However, data entry from transcripts is not a straightforward task. Transcripts vary significantly in their content, layout, structure, legibility, and style.

Before being entered into the appropriate database, all data received from a school was reviewed manually by a catalog coder. The catalog coder “mapped” information from the various materials provided by the schools onto the appropriate fields in the database. Specifically, catalog coders precoded a sample transcript from each AHAA school, illustrating the location of information for each data field, and the location and

value of any flags that should be set for listed courses. Data entry personnel then entered all transcript data and key verified it. Each student's course credits were automatically converted to Carnegie units. The total number of credits was then compared with the minimum required to graduate, and a report of any unusual cases was produced. The catalog coders reviewed each file and Carnegie unit report for consistency and accuracy before the data entry task was considered complete.

Data entry personnel entered information from each eligible transcript from AHAA schools. They entered the information exactly as it appeared on the transcripts, except that they used a set of standard abbreviations and Arabic numbers in course titles. The TCS allows data entry operators to "point and click" to select (rather than type) data elements that have been standardized. Data entry personnel directed any questions or problems to their assigned catalog coder, who reviewed their work for completeness and accuracy throughout the data entry process. When all transcripts for a school were entered, the TCS system changed the status of the school file from "incomplete" to "ready for verification."

All transcript data, except free-form text fields, were 100 percent key verified, using the verification component of the TCS system. This portion of the TCS system is essentially a "re-do and match" process. Data are re-entered (blind to the first entry), and the computer stops when it encounters a non-match between the original data and the current data. Verifiers can then override the original entry with the verified entry. Free-form text fields and test name fields were displayed and reviewed by verifiers, but not key verified.

5.3 Quality Control Measures for Transcript Data Entry and Coding

The quality control measures implemented to ensure accurate and consistent transcript data entry and coding included:

- Selecting coding personnel who were trained educators and knowledgeable about secondary school curriculum;
- Thoroughly training all field and coding personnel;
- Developing and using the computer-based systems (TCS) for data entry and coding that were designed to reduce error and maximize uniformity;
- Implementing multiple quality assurance measures at every stage of coding; and
- Performing numerous automated checks.

Double coding of transcript data, and spot reviews by the senior transcript data entry manager of a subset of transcript records from every school, resulted in coding reliability at 90 percent or better. Automated checks to maintain the quality of data entry

on transcripts were also completed and consisted of checking the reasonableness of the data entered against information from the school catalog. Frequency reports and tabulations checking for outliers were produced and reviewed. For example, cases where students appeared to have no completed credits were reviewed by supervisors.

Additional procedures to verify the accuracy and completeness of all coding efforts included a two-step review process. The first step consisted of generating a report listing the courses that were uncoded, coded as "uncodeable," or coded with an "other" code. The curriculum specialist reviewed these cases and recoded them to the fullest extent possible. The second step involved examining each TCS file a final time, paying close attention to title matching and catalog coding. When problems were identified, verification of the catalog coding was carried out by the curriculum specialist.

The quality control measures described above were performed on the intermediate data files created by the TCS systems, and on the delivery files produced after the intermediate files were merged. Printouts of both intermediate and delivery files were reviewed by the curriculum specialist. Particular attention was given to courses coded with low frequency codes, with a large number of different codes, and with remedial, honors, and AP/IB flags.

5.4 Textbook Coding and Linking Procedures

Documentation of the procedures used to code textbooks and link coded textbook information to student transcripts is provided in the curriculum component of AHAA.

VI. WEIGHTING AND ESTIMATION OF SAMPLING VARIANCE

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

Table 1. Adjusted weights for transcript nonresponse

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

In order to create new analytical weights, students with missing transcripts were considered nonrespondents. The Education Data weights were created by adjusting the Add Health weights for transcript nonresponse in three steps: assignment of disposition

codes, adjust Add Health weights for transcript nonresponse, and benchmark (sample-based raking) the adjusted weights to control totals derived using the Add Health sample. These steps are described in the following paragraphs.

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

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

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

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

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

Table 2. Response code assignment for the Education Data weights

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

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

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

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

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

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

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

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

$$AD1W_i = AD1F_c * AD0W_i ,$$

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

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

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

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

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

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

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

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

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

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

$$AD2W_i = AD2F_k * AD1W_i$$

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

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

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

Table 4. Raking dimensions

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

After raking the sample for the first time, weights were examined to determine the presence of extreme weights. One outlier was detected and trimmed by attaching a

trimming factor to the weight before raking. The trimmed weights were then re-raked. The re-raked weights were examined to verify the procedure was effective at reducing the outlier.

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

REFERENCES

Brackstone, G. J., and Rao, J. N. K. (1979). An investigation of raking ratio estimation. *Sankhya* C 41:97-114.

Brick, J. M., and Kalton, G. (1996). Handling missing data in survey research. *Statistical Methods in Medical Research* 5:215-238.

Deville, J. C., and Särndal, C. E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association* 87:376-382.

Kish, L. (1992). Weighting for unequal pi. *Journal of Official Statistics* 8:183-200.

Lundström, S., and Särndal, C-E. (1999). Calibration as a standard method for treatment of nonresponse. *Journal of Official Statistics* 15:305-327.

VII. GUIDE TO THE AHAA DATA COMPONENTS

An extensive set of constructed variables from the AHAA study is available to the user community. These variables are organized into data files associated with the components described briefly below, each substantively unique and representative of distinctive domains of measurement.

More detailed information about each study component is included in the users' guides. Specifically, the component-specific users' guides consist of a discussion of the component's theoretical and analytical significance, an in-depth presentation of all constructed indicators, an explanation of the conventions used to name the indicators, an overview of the applied missing codes, and a reference guide to all related data files. Codebooks are also available for most AHAA data sets.

The official AHAA website, http://www.prc.utexas.edu/ahaa/, provides up-to-date information about the types of AHAA data files currently available, and those that will be made available at a future date. This website also lists publications and other research-related activities using AHAA data.

7.1 Academic Courses Component

Component Description

More than two decades of research on stratification in schools has shown that students' exposure to curriculum leads to a variety of outcomes. Access to advanced courses is directly related to future opportunity to learn, performance on achievement tests such as college entrance exams, and to college enrollment and success. Thus, examining students' academic achievement in high school provides not only valuable information on inequality during adolescence, but also on the foundation of social and occupational stratification in adulthood. The constructed Academic Courses indicators enable detailed analyses of the academic performance of the multiple Add Health/AHAA cohorts. Specifically, these indicators measure aspects of students' course-taking enrollment and achievement in each year and cumulatively across all years of high school.

The range of academic courses measures available include course sequences, course type, course grades, course failures (failure index variables), semesters attempted, and credits earned variables for the key academic subjects of math, science, foreign language, English, history/social science, and physical education. Academic courses measures are also available for the composite category "overall" which refers to all coursework (including non-core and non-academic courses) taken by students per year and by the end of high school.

7.2 Linking Component

Component Description

Due to the multi-cohort design of Add Health, students' high school careers overlap differently with the Add Health survey years. Thus, AHAA created a set of variables that consist of survey-to-transcript matching indicators. Importantly, these variables enable analysts to link students' course-taking information to the school year 1994-95 – when the in-school survey and Wave I were conducted – and therefore connect existing survey data from Add Health to academic data from AHAA. Specifically, these variables allow analysts to discern when each student was in high school, the duration of each student's high school career, what grade level each student was in during the first Add Health survey year, and each student's transcript-indicated grade level when high school course-taking began.

7.3 Curriculum Component

Component Description

The impact of coursework on student educational and health-related outcomes is intrinsically related to curricular content. The textbook coding is based on curriculum frameworks developed for the Third International Math and Science Study (TIMSS) and is used to create summary measures for the topics covered (content) and the types of tasks students are asked to do with specific topical information (performance expectations). Using information from textbooks and other instructional materials for high school–level mathematics and science courses, William Schmidt and colleagues

developed a method to measure two fundamental aspects of math and science curricula: (1) content, and (2) performance expectations. See http://ustimss.msu.edu/ for more information. Applying this method to AHAA transcript data, AHAA constructed several relatively fine-grained indicators of the learning opportunities made available to AHAA participants through their math and science courses. These indicators are based on detailed coding of textbooks schools reported using in each course in these two subjects. These measures can be linked to students' academic performance, attainment, and to their health-related attitudes and behaviors as reported in the Add Health surveys.

7.4 Academic Networks Component

Component Description

Extensive educational research has shown that the internal academic organization of schools shapes within-school academic and social processes and students' outcomes. AHAA adapted and developed two systematic approaches for measuring academic networks within high schools for two academic years, 1994-95 and 1995-96. (1) Course-overlap indicators, akin to the friendship data in Add Health that captures dyadic ties between two individuals, were constructed, enabling analysts to examine the effects of the relationships that occur when students take similar courses. (2) Local positions were estimated using network methodology to produce non-overlapping course-taking clusters for all the original Add Health schools. Students are placed in one local position per academic year, based upon the students' transcript-recorded course-taking history. These local positions represent particularly salient, intermediate social contexts within the larger school environment. Through the use of hierarchical-linear modeling techniques, these local positions can be used to estimate the effects of the social milieus within schools on student academic, social, and health-related outcomes.

7.5 Contextual Data Component

Component Description

Institutional characteristics and the surrounding context of schools clearly impact students' academic experiences. Therefore, data from the Common Core of Data (CCD) survey, the Private School Survey (PSS), the U.S. census, and the 2000 Office of Civil Rights data which describe school and/or district features were attached to participating schools.

7.6 Transitions Component

Component Description

As adolescents move through school, they are confronted with transitions that impact their educational trajectory. Most students transition from middle school to high school by changing schools. In addition, some students transfer to a new school during high school. These transitions can be used to trace students through middle school and high school as well as estimate the effects of schools on academic and health-related outcomes for incoming students. Available measures include middle school students'

transitioning to high school, student transfers between Wave I and Wave II, and district indicators of last high school attended for transfer students.

7.7 Primary Data Component

Component Description

The primary or raw data indicators are based on information collected from participating schools, and listed directly on student transcripts. One grouping of primary measures conveys comprehensive information about the specific materials gathered from schools during the data collection process. Analysts can examine these measures to determine which and how many data collection instruments were submitted by each school for AHAA. Student-level disposition variables were also produced to enable assessment of the availability of data per Add Health/AHAA student. A third set of primary indicators concern school characteristics as ascertained from school information surveys completed by school administrators. The final grouping of primary or raw data indicators relate to pertinent items recorded on student transcripts such as details about student high school exit or graduation status, and standardized college entrance exam information (which is available for only a very limited number of students).