WARNING

FEDERAL RESTRICTED-USE DATA

UNAUTHORIZED ACCESS TO LICENSED INDIVIDUALLY IDENTIFIABLE INFORMATION IS A VIOLATION OF FEDERAL LAW AND WILL RESULT IN PROSECUTION.

NLTS2 Database Structure

Linked Contents

Note: All samples sizes included in this unrestricted version of the data dictionary are rounded to the nearest 10, per IES Disclosure Review Board requirements for restricted datasets.

Sources of Data

Data from each data collection source and point in time are stored in a separate file or table. All files are linked by a youth ID. Please refer to the Database Overview for a complete description of data collection sources.

In addition to the variables extracted from each of the data collection sources, new variables were created for analysis. Variables created for analyses that are featured in reports and presentations are included in the database. If the variables were created from data within a single data source, those variables are included in the same file as the source data. Variables created from multiple data sources may be in separate files according to their purpose and how they are linked to other files. Multiple data sources include longitudinal comparisons (e.g., a variable constructed from the same item for the same youth collected in different waves of data collection). Another type of multiple-source variable is one created from information collected in the same wave but from different sources (e.g., a variable created to indicate that a youth received mental health services from any source).

Weights were created for every data collection source in every wave and for some combinations of data collection sources. Both main weights and replicate weights are included in the database so that researchers have options as to which statistical package to use. For statistical procedures that produce weighted standard errors, such as SAS Survey or SPSS Complex Samples, there is a separate sample file that includes the sample cluster and strata data required for these procedures. Other statistical packages require just the replicate weights. Main and replicate weights for each source and wave are stored in the file associated with that source.

Individual youth have different combinations of data available from the different sources and in the different waves. Some youth will have data from a Parent Survey, for example, in all Waves, while others may not have a full complement of data from the different sources in the different waves. Unless missing data are imputed, to be included in a longitudinal analysis (e.g., analyzing a pattern of youth employment using data from Waves 2 through 5 of the parent/youth surveys), it may be necessary for respondents to have been surveyed in all waves being analyzed. To be included in an analysis for a single wave of data collection (e.g., youth employment status from the Wave 2 parent/youth survey) requires only the completion of the interview or mail survey questionnaire for that wave. The "Weights" section later in this document discusses use of weights when multiple instruments and/or imputations are used in the analysis.

See Quick References.pdf in Appendix D for a listing of data sources.

Weighted Standard Errors

Because the sample is both stratified and clustered, weighted standard errors need to be calculated that account for the sampling design. Standard procedures such as SAS® Proc Means, SAS Proc Freq, SPSS® Descriptives, and SPSS Crosstabs produce weighted percentages and means and regression coefficients correctly but do not calculate the associated standard errors appropriately; the calculation of standard errors in these procedures assumes a simple random sample, whereas these data are from a stratified and clustered sample. There are statistical packages that will calculate the standard errors correctly: some require the use of replicate weights, and some require sample design data and a single weight. SUDAAN®, WesVar®, SAS,

and Stata® can use either replicate weights or sample design data with a single weight. SAS and SPSS have procedures—based on Taylor linearization—that require sample design variables with a single weight.

It is important to note that weighted standard errors are estimates and that each procedure may calculate similar but slightly different standard errors using the same data and weights. The programs that generated the data tables on the NLTS2 website were written before SAS or SPSS had procedures for calculating weighted standard errors using this type of data. Standard error estimates for the tables on the NLTS2 website were calculated in Base SAS using a formula that calculated an estimated weighted standard error and then multiplied it by a safety factor of 1.25:

$$Adjusted\ SE = SE \times 1.25\sqrt{\frac{M^2 + V}{M^2}}$$

M = mean of the weights
V = Variance of the weights
SE = Unweighted standard error

Estimated weighted standard errors that account for sampling design—although having slight variation with different calculation methods—are the only appropriate standard errors to report using these data.

A data file is included in the database that provides sample design information for each youth in the sample. The data file includes the linking youth identifier [ID] to link to other files and variables for stratum [STRATUM] and cluster [CLUSTER]. This file remains the same and will not change from wave to wave. To calculate weighted standard errors with Taylor linearization procedures, merge these variables with the analysis file. To calculate standard errors with other methods—such as balanced repeated replication (BRR) and jackknife—simply use the replicate weights included in each file.

Both SAS and SPSS—statistical analysis software programs commonly used by researchers—have procedures that calculate weighted standard errors for stratified samples, but these are not part of their base software package. SAS Proc Survey procedures are part of the SAS Statistics module and include procedures such as Proc SurveyFreq, Proc SurveyMeans, and Proc SurveyReg. SPSS recently added a separate module called "Complex Samples," which includes frequencies, crosstabs, descriptives, and regressions. Both SPSS and SAS require a weight, a stratum variable, and a cluster variable for each case or observation in an analysis file.

For SAS Proc Survey procedures, include the statement "strata STRATUM CLUSTER;" with the Proc Survey statements. This statement defines the sample by specifying the stratification and clustering using the "STRATUM" variable and the "CLUSTER" variable.

For SPSS, it is necessary to create a plan file associated with the analysis file. Open the analysis file in SPSS. From the Analyze menu, select "Analyze>Complex Samples>Prepare for Analysis." Select "Create a Plan File" and "Browse" to specify the name of the file and the location it will be stored. Once the name has been specified, click "Next." In the next pop-up box, select "STRATUM" for strata, "CLUSTER" for cluster, and the analysis weight for sample weight, and click "Next." In the subsequent pop-up box, select the "WR" option and click "Finish." To perform analysis, from the Analyze menu, select "Complex Samples" followed by

the analysis procedure, such as "Frequencies" or "Crosstabs." In the pop-up menu, either select "Use default file" or browse for the plan file and select the plan file created in the previous step. The name will have an extension ".csaplan" for the plan file. Once the plan file is specified, the subsequent pop-up screens will look much the same as the SPSS analysis specifications screens used in the standard procedures. Each weight used will have a separate associated plan file.

The "WR" option specifies that the sampling was done with replacement; however, these data were sampled without replacement (WOR). The proportion of sample members in each stratum was low, so there is little difference between what one obtains with and without replacement. Using WR when the sample is actually WOR gives slightly conservative estimates.

There are several other packages that produce weighted standard errors, such as SUDAAN, WesVar, and Stata. For these and other packages, refer to the software documentation as to whether it is appropriate to use replicate weights or sample design information with a single weight. Some packages allow for a choice of methods; the simplest for the user may be to use replicate weights because this method does not require adding sample design information to the analysis data.

See Quick References.pdf in Appendix D for programming examples.

Weights

Weights have been developed for each data collection source within each wave and are listed under "File Descriptions" as well as in the data dictionaries. There is a single weight for each data collection source within a wave with the exception of the Parent/Youth interview/survey data in Waves 2 to 5. There are two main weights and associated sets of replicate weights for every wave of Parent/Youth interview/survey data. One weight is appropriate to use for items that are asked of the parent/guardian-only or for collapsed items that are combined parent/youth responses. The other weight—the youth weight—is appropriate to use for youth-only items or responses. For example, in Wave 3 the weight to use for parent/guardian-only items and combined parent/guardian/youth items is "np3Wt" and for youth-only items, the appropriate weight is "np3YouthWt". The weights for the parent/guardian or parent/guardian/youth combined items follow the same pattern for each wave.

In addition to weights for individual data collection sources within each wave, weights were developed for youth to be included in analyses that require any of the following combinations of data from:

1. Any Parent/Youth interview/survey
2. Any Parent/Youth interview/survey when the youth was out of high school
3. One or more of the transcript, school program surveys, or general education teacher surveys (for example, a student with only a transcript would have a non-zero weight, as would a student who had a transcript and a school program survey)
4. Any wave of Parent/Youth interview/survey and a school program survey
5. Any post-high school Parent/Youth interview/survey and one or more of the transcript or school program surveys (for example, a student with a post-high school Parent/Youth interview/survey and a school program survey would have a non-zero weight)
6. Any data collection for any wave.

Generally, when data from multiple sources are combined in an analysis and *there are no imputations*, the appropriate weight to use for the combined data is the weight from the data collection source with the smallest sample size. For example, when combining Wave 1 Parent interview/survey data with Wave 2 Parent/Youth interview/survey data, use the Wave 2 Parent/Youth weight "np2Wt" because there are fewer respondents in the Wave 2 Parent/Youth database than in the Wave 1 Parent database. When combining Wave 2 Direct Assessment data with Wave 3 Parent/Youth data, the appropriate weight to use would the weight for the Direct Assessment, "Wt_na," because there are fewer respondents in the Direct Assessment database than in the Wave 3 Parent/Youth interview/survey database.

When *missing data are imputed* for an analysis, the researcher must make two separate decisions in determining which weight to use: 1) which instruments must be present to qualify the youth for participation in the study, and 2) which additional instruments will have entire instruments imputed. For example, in an analysis based on data from a Parent/Youth interview/survey conducted when the youth was out of high school and a School Program survey, a researcher might require that the youth have both instruments to be included in the analyses. In that case, imputation would be used to fill in missing items within each instrument when both surveys are available. Youth who do not have both surveys would not be included in the analyses. In contrast, a researcher might instead only require that youth have a post-high school Parent/Youth interview/survey to be included in the analyses. In that case, imputation would be used to fill in missing items in the available Parent/Youth interview/survey, missing items in *available* School Program surveys, as well as *entire* School Program surveys when they are missing (as long as the youth has the Parent/Youth interview/survey).

When missing data are imputed and the set of instruments that would qualify the youth to be in the analysis are one of the combinations described on page 26 (e.g., any data from any Parent/Youth interview), the weights developed for those combinations of instruments and waves should be used. For example, suppose that a researcher is conducting an analysis on students who have any wave of Parent/Youth interview/survey and a School Program survey. There is a weight for this combination of survey instruments. This weight was calculated under the assumptions that (1) missing items within the Parent/Youth and School Program surveys will be imputed and (2) the entire Parent/Youth and School Program surveys will not be imputed for any student who is missing one or both of those surveys. Another example would be a researcher who is conducting an analysis of students who have one or more of the following: a school transcript, a School Program survey, or a General Education Teacher Survey.[1] There is a weight for this combination of survey instruments. That weight was calculated under the assumption that the researcher will be imputing all missing items in those three surveys for any student who has any one of those surveys. This not only includes imputation within a survey; if a student only has a School Program, the weight assumes that the researcher will impute the entire school transcript and General Education Teacher survey for that student. If a weight for the specific combination

[1] If a student is known from other sources (e.g., high school transcript) not to have a general education class, then the appropriate imputation of the General Education Teacher survey is to code the variables as "not applicable." Similarly for any items involving skip logic, if the preceding actual or imputed item implies that the conditional item should have been skipped (e.g., wages for those who have never had a job), then the appropriate imputation for the condition item should be "not applicable."

of instruments to be used in the analysis is not available, the weight that best approximates the data to be used in an analysis should be selected. For example, if a researcher wants to conduct an analysis of students who have any Parent/Youth interview/survey and a General Education Teacher Survey, there is no weight specifically for that combination of data. Similarly if a researcher wants to conduct an analysis for students who have any of the School Program, General Education Teacher Survey, or transcript, but was only willing to impute entire instrument missing data for the School Program survey, there is no fully appropriate weight. In either of these cases, the researcher will need to conduct the following calculations to determine which weight is the most appropriate and should therefore be used.

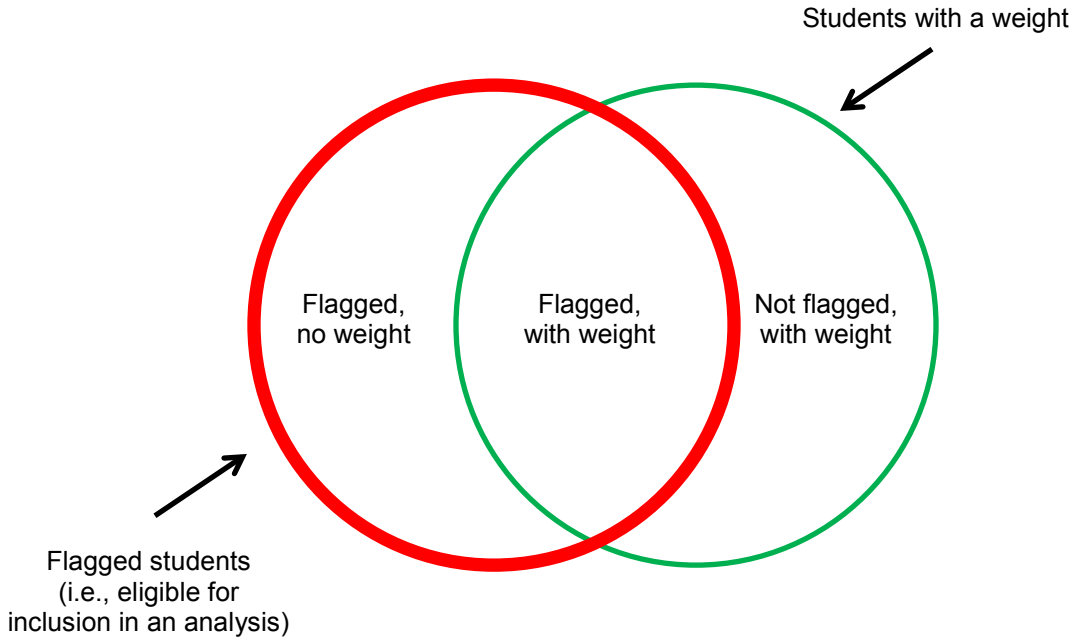
1. The first step in identifying the most appropriate weight is to construct an indicator variable that has a value of 1 for students who should have a weight and a value of 0 for youth who should not have a weight. For example, suppose a researcher wants to conduct an analysis of youth with learning disabilities (LD) who have not been declassified, are not in a special school, and who have *both* a School Program Survey and a post-high-school Parent/Youth interview/survey. In this case, the researcher will want to construct an indicator variable that is 1 for any youth with LD who has both a School Program Survey and an out-of-school Parent/Youth interview/survey. For purposes of selecting the most appropriate weight, youth who have an indicator variable value of 1 will be denoted as "flagged" youth. Note that the flagged youth include students who have been declassified and who are in a special school. Later in the analysis stage, those students can be excluded.
2. The second step is to apply to the distribution of flagged and not-flagged youth two criteria for selecting the most appropriate weight. The first criterion is that the researcher should select a weight that minimizes the number of flagged youth (who therefore qualify to be part of the analysis) who do not have a weight (or equivalently, that their weight is zero) and therefore would be dropped from the analyses. The second criterion is that the researcher should select a weight that minimizes the number of unflagged youth (who therefore would not qualify to be part of the analysis) who have a weight. These youth could bias the weights for the flagged youth and, more importantly, bias the weights for the subset of youth who qualify to participate in the analysis.
3. The third step in identifying the most appropriate weight is to calculate the two numbers discussed in the prior paragraph for all weights—the number of flagged youth with no weights and the number of unflagged youth with weights. It is highly recommended that the researcher apply all weights that are reasonable candidates to be selected to the analysis and compare results.

If there is a weight that minimizes both quantities, that is the most appropriate one to use. If weights minimize one quantity but not the other, the researcher will need to make a tradeoff between (a) excluding flagged students who qualify for the analysis and (b) possible biasing the weights by including students who are unflagged (i.e., do not qualify for inclusion) but have weights. In that case, it is generally preferable to select the weight that minimizes the first criterion because the resulting bias is generally small.

Figure 1 below illustrates the process and the issues. The circle with the bold circumference represents all flagged youth—those eligible for inclusion in the analysis. The other circle represents all youth with a particular weight. The intersection, labeled "flagged with weight" is

all individuals who are flagged and have a weight. The portion of the diagram labeled "flagged, no weight" are the eligible students who have no weight. The portion of the diagram labeled "Not flagged, with weight" are those youth who are not eligible for inclusion but have a weight. The researcher needs to select a weight that maximizes the number of youth in the "flagged, with weight" portion of the diagram, and minimizes the number of youth in the "flagged, no weight" and "Not flagged, with weight" portions.

Figure 1. Venn diagram of flagged students and weighted students



For example, a researcher is interested in performing an analysis of male youth with LD who had post-high school employment and who had any wave of Parent/Youth interview/survey and any wave of the General Education Teacher survey. The researcher plans to impute the missing values of any Parent/Youth interview/survey and any General Education Teacher survey as long as the youth has at least one of each of these surveys. For purposes of selecting a weight, the researcher would specify an expanded analysis population including both male and female students and students who had and did not have post-high school employment, with the understanding that when the analysis actually is conducted, the researcher will restrict the analysis to male youth with LD who had (actual or imputed) post-secondary employment. (In general, when selecting the appropriate weight, the researcher should not restrict the sample according to characteristics such as gender, household income, enrollment in special schools, etc. when they can later be used to filter out youth who do not have the desired set of characteristic.) There is no weight created for this combination of instruments. Table 1 shows the numbers of youth with LD in the expanded analysis population with and without weights and the number of youth with LD who are *not* in the expanded analysis population with and without weights.

There are 420 youth with LD in the expanded analysis sample. The researcher would clearly not want to use wt_AnyPYPHS or wt_AnyPYPHSSch as the analysis weight, because using

those weights would result in a loss of 130 or 130 of those youth, respectively. Wt_AnySchTS is superior to wt_Any, wt_AnyPY because there are fewer eligible analysis sample youth with no weight and fewer non-eligible youth with weights. However, if the researcher uses wt_AnySchTS, the weights for the analysis youth will be distorted because 460 youth who are not eligible for the expanded analysis sample also have weights. If the researcher uses wt_AnyPYProg, 30 eligible youth are lost because they have no weight, but there are fewer non-eligible youth with weights to bias the weights for the analysis youth. So the researcher will need to decide between wt_AnySchTS and wt_AnyPYProg, or, preferably, run analyses using both sets of weights and compare results.

Table 1. Example of table and counts used for selecting weights

Weight	In expanded analysis sample; with weight	In expanded analysis sample; no weight	Not in expanded analysis sample, with weight	Not in expanded analysis sample, no weight
Any Parent/Youth survey [wt_Anypy]	420	0	500	190
Any Parent/Youth survey when the youth was out of high school [wt_AnyPYPHS]	290	130	290	410
Data from one or more of the transcript, school program surveys, or general education teacher surveys [wt_AnySchTS]	420	0	460	240
Any wave of Parent/Youth survey and a school program survey [wt_AnyPYProg]	400	30	190	510
Any post high school Parent/Youth survey and one or more of the transcript or school program surveys [wt_AnyPYPHSSch]	290	130	210	490
Any data collection for any wave [wt_Any]	420	10	620	80

A "failsafe" weight is the weight for any student who has data from any source. This weight can always be used if a researcher clearly identifies the portion of the universe that is represented by ineligible students (i.e., the sum of the weights of ineligible students). For example, if analysis requires that a youth have any in-school Parent/Youth interview/survey and the weight for any youth who has data from any source is applied, 92 percent of the total sample weight belongs to such youth—i.e., flagged youth project up to 92 percent of the universe. The researcher can then perform an analysis with this weight, explaining that the remaining 8 percent of youth have insufficient information to include in the analysis. Note that the researcher could sub-select youth

from among those who have an in-school Parent/Youth interview (for example, youth whose parent's education is at a specified level), in which case the analysis projects to the qualifying portion of the 92% of the universe that have an in-school Parent/Youth survey.

A similar approach to the "failsafe" approach might apply to other weights. For example, suppose that a researcher requires that a student have two or more Parent/Youth interviews, and the researcher decides to use the weight for all students who have one or more Parent/Youth interviews. If the sum of the weights for students with two or more Parent/Youth interviews constitute 85 percent of the total of the selected weight, then the researcher can perform the analysis with this weight, explaining that the remaining 15 percent of students have insufficient information to include in the analysis.

Note that imputations were not made by SRI at the time NLTS2 data were finalized and therefore are not included in the database.

See the weights file data dictionary for a listing and description of NLTS2 weights.

Note: All samples sizes included in this unrestricted version of the data dictionary are rounded to the nearest 10, per IES Disclosure Review Board requirements for restricted datasets.

Linking Identifier for Linking Files (or Tables)

All tables are linked by a key called ID, which is a five-digit randomly assigned ID number at the youth level.

Data Source or Type	Links by
Parent Survey	Youth ID [ID]
Parent/Youth Survey	Youth ID [ID]
Direct Assessment	Youth ID [ID]
Alternate Assessment	Youth ID [ID]
Teacher Survey	Youth ID [ID]
Student's School Program Survey	Youth ID [ID]
Student Transcript	Youth ID [ID]
School Characteristics Survey	Youth ID [ID]
Variables created for analysis at the student level using multiple sources and/or waves	Youth ID [ID]

There are varying numbers of records when matching for more than one data source. The following are the results of matching records within each wave

Wave 1 Parent Survey and School Data matches for Parent Survey, School Program Survey, Teacher Survey, School Characteristics Survey, and Transcript data. Students who were in programs that did not have transcripts recorded are not included in these counts.

Number of respondents with all data:

1,870 Parent, Program, Teacher, Characteristics, and Transcript

Number of respondents with 4 sources of data:

110 Parent, Program, Teacher, Characteristics, and [no Transcript]
50 Parent, Program, Teacher, [no Characteristics], and Transcript
150 Parent, [no Program], Teacher, Characteristics, and Transcript
290 [no Parent], Program, Teacher, Characteristics, and Transcript
2,220 Parent, Program, [no Teacher], Characteristics, and Transcript

Number of respondents with 3 sources of data:

520 Parent, Program, [no Teacher], Characteristics, and [no Transcript]
90 Parent, Program, [no Teacher], [no Characteristics], and Transcript
10 Parent, Program, Teacher, [no Characteristics], and [no Transcript]
960 Parent, [no Program], [no Teacher], Characteristics, and Transcript
20 Parent, [no Program], Teacher, Characteristics, and [no Transcript]
20 Parent, [no Program], Teacher, [no Characteristics], and Transcript
320 [no Parent], Program, [no Teacher], Characteristics, and Transcript
20 [no Parent], Program, Teacher, Characteristics, and [no Transcript]
10 [no Parent], Program, Teacher, [no Characteristics], and Transcript

Note: All samples sizes included in this unrestricted version of the data dictionary are rounded to the nearest 10, per IES Disclosure Review Board requirements for restricted datasets.

30 [no Parent], Teacher, [no Program], [no Teacher], Characteristics, and Transcript

Number of respondents with 2 sources of data:

40 Parent, Program, [no Teacher], [no Characteristics], and [no Transcript]
680 Parent, [no Program], [no Teacher], Characteristics, and [no Transcript]
1,780 Parent [no Program], [no Teacher], [no Characteristics], and Transcript
<10 Parent, [no Program], [no Characteristics], Teacher, and [no Transcript]
60 [no Parent], Program, [no Teacher], Characteristics, and [no Transcript]
20 [no Parent], Program, [no Teacher], [no Characteristics], and Transcript
10 [no Parent], Program, Teacher, [no Characteristics], and [no Transcript]
<10 [no Program], Teacher [no Characteristics], and Transcript
90 [no Parent], [no Program], [no Teacher], Characteristics and Transcript

Number of respondents with 1 source of data:

260 Characteristics only
20 Program only
<10 Teacher only
1,040 Transcript only
80 no Wave 1 Parent or School data

Wave 2 Parent/Youth Survey and School Data matches for Parent/Youth Survey, School Program Survey, Teacher Survey, Student Assessment, and Transcript data. The following have:

Number of respondents with all data:

1,040 Parent/Youth, Program, Teacher, Assessment, and Transcript

Number of respondents with 4 sources of data:

10 Parent/Youth, Program, Teacher, Assessment, and [no Transcript]
170 Parent/Youth, Program, Teacher, [no Assessment], and Transcript
1,260 Parent/Youth, Program, [no Teacher], Assessment, and Transcript
60 Parent/Youth, [no Program], Teacher, Assessment, and Transcript
230 [no Parent/Youth], Program, Teacher, Assessment, and Transcript

Number of respondents with 3 sources of data:

10 Parent/Youth, Program, Teacher, [no Assessment], and [no Transcript]
250 Parent/Youth, Program, [no Teacher], Assessment, and [no Transcript]
240 Parent/Youth, Program, [no Teacher], [no Assessment], and Transcript
<10 Parent/Youth, [no Program], Teacher, Assessment, and [no Transcript]
10 Parent/Youth, [no Program], Teacher, [no Assessment], and Transcript
1,840 Parent/Youth, [no Program], [no Teacher], Assessment, and Transcript
<10 [no Parent/Youth], Program, Teacher, Assessment, and [no Transcript]
180 [no Parent/Youth], Program, Teacher, [no Assessment], and Transcript
300 [no Parent/Youth], Program, [no Teacher], Assessment, and Transcript
10 [no Parent/Youth], [no Program], Teacher, Assessment, and Transcript

Number of respondents with 2 sources of data:

70 Parent/Youth, Program, [no Teacher], [no Assessment], and [no Transcript]
<10 Parent/Youth, [no Program], Teacher, [no Assessment], and [no Transcript]
480 Parent/Youth, [no Program], [no Teacher], Assessment, and [no Transcript]
870 Parent/Youth, [no Program], [no Teacher], [no Assessment], and Transcript
10 [no Parent/Youth], Program, Teacher, [no Assessment], and [no Transcript]

Note: All samples sizes included in this unrestricted version of the data dictionary are rounded to the nearest 10, per IES Disclosure Review Board requirements for restricted datasets.

50 [no Parent/Youth], Program, [no Teacher], Assessment, and [no Transcript]

220 [no Parent/Youth], Program, [no Teacher], [no Assessment], and Transcript

<10 [no Parent/Youth], [no Program], Teacher, Assessment, and [no Transcript]

10 [no Parent/Youth], [no Program], Teacher, [no Assessment], and Transcript

600 [no Parent/Youth], [no Program], [no Teacher], Assessment, and Transcript

Number of respondents with 1 source of data:

560 Parent/Youth only

50 Program only

150 Assessment only

1,900 Transcript only

900 no Wave 2 Parent or School data

Wave 3 Parent/Youth Survey and School Data matches for Parent/Youth Survey and Transcript data:

Number of respondents with both sources of data:

4,570 Parent/Youth and Transcript

Number of respondents with 1 source of data:

1,090 Parent/Youth data only

4,360 Transcript only

Wave 4 Parent/Youth Survey and School Data matches for Parent/Youth Survey and Transcript data. The following have:

Number of respondents with both sources of data:

4,480 Parent/Youth and Transcript

Number of respondents with 1 source of data:

1,100 Parent/Youth data only

4,450 Transcript only

Wave 5 Parent/Youth Survey and School Data matches for Parent/Youth Survey and Transcript data. The following have:

Number of respondents with both sources of data:

4,570 Parent/Youth and Transcript

Number of respondents with 1 source of data:

4,360 Transcript only

1,090 Parent/Youth data only

There are 9,580 respondents with some combination of Parent/Youth Survey data. The following are the results of matching records across each Wave for Parent/Youth Survey data:

Number of respondents with five waves of data:

3,730 Wave 1 Parent and Waves 2, 3, 4, and 5 Parent/Youth

Number of respondents with four waves of data:

550 Wave 1 Parent and Waves 2, 3, and 4 Parent/Youth

260 Wave 1 Parent and Waves 2, 3, and 5 Parent/Youth

370 Wave 1 Parent and Waves 2, 4, and 5 Parent/Youth

170 Wave 1 Parent and Waves 3, 4, and 5 Parent/Youth

80 Waves 2, 3, 4, and 5 Parent/Youth

Number of respondents with three waves of data:

540 Wave 1 Parent and Waves 2 and 3 Parent/Youth

Note: All samples sizes included in this unrestricted version of the data dictionary are rounded to the nearest 10, per IES Disclosure Review Board requirements for restricted datasets.

150	Wave 1 Parent and Waves 2 and 4 Parent/Youth
160	Wave 1 Parent and Waves 2 and 5 Parent/Youth
60	Wave 1 Parent and Waves 3 and 4 Parent/Youth
40	Wave 1 Parent and Waves 3 and 5 Parent/Youth
210	Wave 1 Parent and Waves 4 and 5 Parent/Youth
20	Waves 2, 3, and 4 Parent/Youth
10	Waves 2, 3, and 5 Parent/Youth
10	Waves 2, 4, and 5 Parent/Youth
30	Waves 3, 4, and 5 Parent/Youth

Number of respondents with two waves of data:

860	Wave 1 Parent and Wave 2 Parent/Youth
100	Wave 1 Parent and Wave 3 Parent/Youth
130	Wave 1 Parent and Wave 4 Parent/Youth
200	Wave 1 Parent and Wave 5 Parent/Youth
30	Waves 2 and 3 Parent/Youth
10	Waves 2 and 4 Parent/Youth
10	Waves 2 and 5 Parent/Youth
10	Waves 3 and 4 Parent/Youth
<10	Waves 3 and 5 Parent/Youth
30	Waves 4 and 5 Parent/Youth

Number of respondents with one wave of data:

1,690	Wave 1 Parent only
70	Wave 2 Parent/Youth only
10	Wave 3 Parent/Youth only
20	Wave 4 Parent/Youth only
350	Wave 5 Parent/Youth only

The following are the results of matching records across waves and instruments, based on the combinations in the six additional weights (described on page 26).

Number of respondents with data from:

9,580	Any Parent/Youth interview/survey
6,500	Any Parent/Youth interview/survey when the youth was out of high school
8,710	One or more of the transcript, school program surveys, or general education teacher surveys (for example, a student with only a transcript would have a non-zero weight, as would a student who had a transcript and a school program survey)
6,250	Any wave of Parent/Youth interview/survey and a school program survey
5,570	Any post-high school Parent/Youth interview/survey and one or more of the transcript or school program surveys (for example, a student with a post-high school Parent/Youth interview/survey and a school program survey would have a non-zero weight)
10,640	Any data collection for any wave.

Note: All samples sizes included in this unrestricted version of the data dictionary are rounded to the nearest 10, per IES Disclosure Review Board requirements for restricted datasets.

System and Security

Disclosive contact information and sample tracking data are not included in the database. These data are used exclusively for tracking and data collection efforts and are stored in a separate system. Identifiers used for linking files are randomly assigned.

Formats or Value Labels

SAS data are linked to a format library included on the CD-ROM. The format library contains look-up tables of possible values and the value labels for associated categorical and ordinal variables in the SAS data files. When data are accessed in SAS, the system expects to find the associated formats and will issue an error message if these formats are not found. SAS users can use one of the following options:

- To access the formats library in a SAS program, include a "libname" statement with the libname alias "library" and the location (or path) of the library:

  ```
  Libname library [data path] ;
  ```

  The libname statement points to the format library, "Formats.sas7bcat"

- To run programs without the associated formats, exclude the "libname" statement and include an option statement in the SAS program code:

  ```
  option nofmterr ;
  ```

  This option will suppress all format error messages and allow the program to compile.

- To recreate the formats—either to generate the library or create formats for a local session—run the included SAS program, NLTS2_formats.sas. This program contains the program code to generate the formats.

Data were developed in SAS. Using software for data transfer, these SAS data were converted to SPSS. SPSS has a different file structure. In SAS, a single format table is linked to multiple variables and contains all possible values for those variables. When data are transferred from SAS to to SPSS, the values table links with each individual variable. The "one size fits all" SAS format that has all possible values for multiple variables is now attached to a single SPSS variable, and it is quite possible that the single variable does not have all the values listed in that values table seen in the SPSS variable view. SPSS users should refer to appendix C for more information about value labels in SPSS files.

The example below shows what happens in SPSS and SAS using a range format and a discrete format to collapse categories. Variables X and Y have the same values but have different user defined formats: one with a range assignment and one with a discrete value assignment. A frequency distribution was run on these variables in SAS and SPSS:

Note: All samples sizes included in this unrestricted version of the data dictionary are rounded to the nearest 10, per IES Disclosure Review Board requirements for restricted datasets.

User Defined Value Formats with Range of Values Defined

SAS

x	Frequency	Percent	Cumulative Frequency	Cumulative Percent
(1) Mostly As	1	10.00	1	10.00
(2) Mostly As and Bs	1	10.00	2	20.00
(3) Mostly Bs	1	10.00	3	30.00
(4) Mostly Bs and Cs	1	10.00	4	40.00
(5) Mostly Cs	1	10.00	5	50.00
(6) Mostly Cs and Ds	1	10.00	6	60.00
(7-9) Mostly Ds and Fs	4	40.00	10	100.00

SPSS

x

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	(1) Mostly As	1	10.0	10.0	10.0
	(2) Mostly As and Bs	1	10.0	10.0	20.0
	(3) Mostly Bs	1	10.0	10.0	30.0
	(4) Mostly Bs and Cs	1	10.0	10.0	40.0
	(5) Mostly Cs	1	10.0	10.0	50.0
	(6) Mostly Cs and Ds	1	10.0	10.0	60.0
	7.00	1	10.0	10.0	70.0
	8.00	1	10.0	10.0	80.0
	9.00	2	20.0	20.0	100.0
	Total	10	100.0	100.0	

User Defined Value Formats with Discrete Values Defined

SAS

y	Frequency	Percent	Cumulative Frequency	Cumulative Percent
(1) Mostly As	1	10.00	1	10.00
(2) Mostly As and Bs	1	10.00	2	20.00
(3) Mostly Bs	1	10.00	3	30.00
(4) Mostly Bs and Cs	1	10.00	4	40.00
(5) Mostly Cs	1	10.00	5	50.00
(6) Mostly Cs and Ds	1	10.00	6	60.00
(7-9) Mostly Ds and Fs	4	40.00	10	100.00

SPSS

y

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	(1) Mostly As	1	10.0	10.0	10.0
	(2) Mostly As and Bs	1	10.0	10.0	20.0
	(3) Mostly Bs	1	10.0	10.0	30.0
	(4) Mostly Bs and Cs	1	10.0	10.0	40.0
	(5) Mostly Cs	1	10.0	10.0	50.0
	(6) Mostly Cs and Ds	1	10.0	10.0	60.0
	(7-9) Mostly Ds and Fs	1	10.0	10.0	70.0
	(7-9) Mostly Ds and Fs	1	10.0	10.0	80.0
	(7-9) Mostly Ds and Fs	2	20.0	20.0	100.0
	Total	10	100.0	100.0	

SAS users can take advantage of pre-assigned formats that collapse categories and ranges of values, or optionally create new user-defined formats for existing variables or recreate variables with collapsed values. To collapse categories or ranges of values, SPSS users must recode variables into a new variable. Example SPSS programming code for collapsing variables is included in the document SPSS QuickCode.pdf in Appendix C. This SPSS programming code

references the SAS format ranges to create variables with the same collapsed ranges SAS users have available by default.

It is never advisable to rely solely on variable labels and value labels when choosing items to use in analyses. Because there are many sources of data, similar types of items asked of multiple respondents, and complex skip patterns, it is important to refer to the codebook for details about each variable before using these data. In addition, it is highly recommended that users run a frequency distribution on the selected variables and check for values, distribution of values, and n's.

Missing Values

Data can be missing for various reasons in the NLTS2 database. Files associated with each data collection include only those who completed the survey or assessment; the number of respondents varies from file to file. Within a file, respondents may have missing responses to individual items. User defined missing values indicate the reasons items are missing for respondents. In SAS, user defined missing values begin with a period followed by a letter. SPSS does not support the number of discrete missing values that are designated in the NLTS2 database, but SPSS does support using a numeric range of missing values. Missing values in SPSS range from -999 to -900 and in SAS from .a to .z.

Missing Values		
SAS	**SPSS**	**Missing Value**
.a	-999	Not applicable
.b	-988	Not asked this section
.c	-987	Incomplete
.d	-998	Do not know
.g	-986	Sparse data
.i	-993	In secondary school: returned to HS
.m	-997	Mail
.n	-985	Not graded
.p	-996	Partial
.r	-995	Refused
.s	-994	Skipped
.u	-991	Ungraded
.v	-990	Version
.w	-984	Withdrawn
.x	-992	Not in NCES data
.z	-989	Not ascertained

File and Variable Naming Conventions

Names of files or tables reflect the source and the data collection wave. For original data collected by SRI, the prefix to each variable (or column) name reflects the source and data collection wave, if applicable. These variables incorporate the item number from the instrument

Note: All samples sizes included in this unrestricted version of the data dictionary are rounded to the nearest 10, per IES Disclosure Review Board requirements for restricted datasets.

so that an individual data collection instrument is self-documenting. For example, the variable name for data from the Wave 2 Parent Survey (prefix np2), section B, question 1, subitem c would be "np2B1c." Within a given file, sample data, interview preload, created, and multiple source variables may have different prefixes, whereas items coming directly from a single data collection instrument will have a uniform prefix.

File Descriptions

Summary information is listed below for each file by wave and by source. File descriptions include the file name, variable prefix generally used in the variable naming conventions, the name of the weight for the given wave and source, and the number of records in each file. For detailed descriptions and contents of each file, refer to the individual data dictionaries. Data dictionaries are organized by wave and by data collection source.

Note: All samples sizes included in this unrestricted version of the data dictionary are rounded to the nearest 10, per IES Disclosure Review Board requirements for restricted datasets.

NLTS2 Data Collection

A: Wave 1 Data Collection

A1: Parent Survey

File Name: N2W1Parent
Wave: 1
Variable Prefix: np1
Weight: np1Wt
Number of records: 9,230
Links by: ID

A2: Teacher Survey

File Name: N2W1Tchr
Wave: 1
Variable Prefix: nts1
Weight: wt_NTS1
Number of records: 2,590
Links by: ID

A3: Student's School Program Survey

File Name: N2W1Prog
Wave 1
Variable Prefix: npr1
Weight: wt_NPR1
Number of records: 5,640
Links by: ID

A4: School Characteristics Survey

File Name: N2W1SchChar
Wave: 1
Variable Prefix: nsc1
Weight: wt_NSC1
Number of records 7,600
Links by: ID

Note: All samples sizes included in this unrestricted version of the data dictionary are rounded to the nearest 10, per IES Disclosure Review Board requirements for restricted datasets.

A5: Wave 1 Cross-Instrument Data

File Name: N2W1MultInstr
Wave: 1
Variable Prefix: nxm1
Weight: Depending upon analysis, use source weights np1Wt, wt_NTS1, or wt_NPR1 [created specifically for combined Program and Teacher data]
Number of records: 11,270
Links by: ID

B: Wave 2 Data Collection

B1: Parent/Youth Survey

File Name: N2W2ParYouth
Wave: 2
Variable Prefix: np2
Weights: np2Wt [use with parent/guardian response items or parent/guardian/youth combined items]
np2YouthWt [use with youth only response items]
Number of records: 6,860
Links by: ID

B2: Direct Assessment

File Name: N2W2DirAssess
Waves: 1 and 2
Variable Prefix: nda2
Weight: wt_na
Number of records: 5,240
Links by: ID

B3: Alternate Assessment

File Name: N2W2AltAssess
Waves: 1 and 2
Variable Prefix: naa2
Weight: wt_na
Number of records: 1,050
Links by: ID

Note: All samples sizes included in this unrestricted version of the data dictionary are rounded to the nearest 10, per IES Disclosure Review Board requirements for restricted datasets.

B4: Teacher Survey

File Name: N2W2Tchr
Wave: 2
Variable Prefix: nts2
Weight: wt_NTS2
Number of records: 1,740
Links by: ID

B5: Student's School Program Survey

File Name: N2W2Prog
Wave 2
Variable Prefix: npr2
Weight: wt_NPR2
Number of records: 4,080
Links by: ID

C: Wave 3 Data Collection

C1: Parent/Youth Survey

File Name: N2W3ParYouth
Wave: 3
Variable Prefix: np3
Weights: np3Wt [use with parent/guardian response items or parent guardian/youth combined items]
np3YouthWt [use with youth only response items]
Number of records: 5,660
Links by: ID

D: Wave 4 Data Collection

D1: Parent/Youth Survey

File Name: N2W4ParYouth
Wave: 4
Variable Prefix: np4
Weights: np4Wt [use with parent/guardian response items or parent guardian/youth combined items]
np4YouthWt [use with youth only response items]
Number of records: 5,580
Links by: ID

Note: All samples sizes included in this unrestricted version of the data dictionary are rounded to the nearest 10, per IES Disclosure Review Board requirements for restricted datasets.

E: Wave 5 Data Collection and Cross-Wave Data

E1: Parent/Youth Survey

File Name: N2W5ParYouth

Wave: 5

Variable Prefix: np5

Weights: np5Wt [use with parent/guardian response items or parent guardian/youth combined items]
np5YouthWt [use with youth only response items]

Number of records: 5,320

Links by: ID

E2: Transcript Data Files

Wave: 5

Weight: Tr_Wt

Links by: ID

E2a: Transcript Overall

File Name: n2Trans_Overall

Variable Prefix: nta

Number of records

In file: 8,210

Per respondent: One record per student with any transcript data for students in programs with transcripts

E2b: Transcript By Year

File Name: n2Trans_ByYear

Variable Prefix: nty

Number of records

In file: 31,740

Per respondent: Multiple records per student: one record per school year for each school year student attended

Multiple records per student: one record per school year

Note: All samples sizes included in this unrestricted version of the data dictionary are rounded to the nearest 10, per IES Disclosure Review Board requirements for restricted datasets.

E2c: Transcript By Course

File Name: n2Trans_ByCourse

Variable Prefix: ntc

Number of records

In file: 339,610

Per respondent: Multiple records per student: one record per course for each course reported on student's transcript

E2d: Transcript Coursetaking Summary By Grade Level

File Name: n2Trans_ByGrade

Variable Prefix: ntg

Number of records

In file: 28,730

Per respondent: Multiple records per student: one record per grade for each grade level reported in student's course level data

E2e: Transcript Coursetaking Summary

File Name: n2Trans_Summary

Variable Prefix: nts

Number of records

In file: 6,760

Per respondent: One record per student of those who left secondary school and have course level data for grade levels attended

E3: Cross-Instrument Data

File Name: n2MultInstr

Waves: 1-5

Variable Prefix: w1rp, w2rp, w3rp, w4rp, w5rp

Number of records: 11,270

Links by: ID

E4: Sample Data

File Name: n2Sample

Wave: All

Variable Prefix: [none]

Weight [none]

Number of records 11,270

Links by: ID

Note: All samples sizes included in this unrestricted version of the data dictionary are rounded to the nearest 10, per IES Disclosure Review Board requirements for restricted datasets.

E5: Weights

File Name: n2Weights

Waves: 1-5

Variable Prefix: [varies by type of weight]

Weights: Weights calculated for each data collection source within a wave and weights calculated for cross-wave and/or cross-source data

Cross wave/cross data collection source weights are included for analysis of data combined from multiple waves and sources. (Note: it is advisable that imputation of data from missing sources be considered when using these weights). Use:

- wt_Any for any data from any wave and any source
- wt_AnyPY for any Parent/Youth data from any wave
- wt_AnyPYPHS for any Parent/Youth data collected at the time youth was out of secondary school
- wt_AnySchTS for one or more of the following: School Program Survey, Teacher Survey, or transcript data
- wt_AnyPYProg for any Parent/Youth data and, in addition, any School Program Survey
- wt_AnyPYPHSSch for any Parent/Youth data collected at the time youth was out of secondary school and one or more of the following: any School Program Survey or transcript

Number of records: 11,270

Links by: ID