



Washington State
Healthy Youth Survey

*Data Analysis &
Technical Assistance
Manual*

2010 Healthy Youth Survey Data
January 2013

WASHINGTON STATE DEPARTMENT OF HEALTH

Healthy Youth Survey Data Analysis & Technical Assistance Manual



Prepared by Susan Richardson, Lillian Bensley and Vivian Hawkins
Washington State Department of Health

For more information contact:

Washington State Department of Health
TEL: 877-497-7111 (Toll Free)
Email: healthy.youth@doh.wa.gov

DOH Pub 210-088

Throughout this manual: STATA commands are in **grey** and STATA output is in **black**

Table of Contents

Introduction	4
Purpose	4
Audience	4
Uses	4
Manual Layout	4
Issues in Analyzing Healthy Youth Survey Data	6
1. HYS Overview	10
Survey History	10
Survey Questionnaires	13
Survey Implementation Schedule	16
Sampling	16
Participation Rates	17
Confidence Intervals	20
Bias Analysis	20
2. Getting Access to HYS Data	22
Data Sharing and Human Research Review Requirements	22
Data Sharing Agreements	24
3. Getting to Know Your HYS Data	25
Demographic Variables	26
Substance Use Variables	30
Calculated/Computed Variables	34
Risk and Protective Factors	39
Content Changes Over Time	39
4. Getting to Know STATA	40
5. HYS Data Analysis in STATA	43
Opening your Dataset	44
General Set Up for Survey Analysis	45
Analysis by Grade	50
Frequencies and Summary of Statistics	51
Creating New Variables	53
Labeling New Variables	56
General Rules on Creating Dichotomous Variables	57

Two-Way Tables or Crosstabs	58
Additional Options with Svy	59
Additional Tips for Formatting Data	61
Stratified Analysis and Subpopulations	63
6. HYS Data Analysis – Quick Example	67
Set Up for Survey Analysis	68
Data Analysis Example	70
7. Comparing State and Local Data	72
Appending	73
Comparing Local vs. the Rest of the State	74
8. Comparing Years of Data	75
Appending	75
Analysis Stratified by Year	76
When to Combine Multiple Years of Data	78
Methods for Combining Years	78
Year Standardized Estimates	79
9. Combining Grade Levels	82
When to Combine Grades	82
Methods for Combining Grades	82
Grade-Adjusted Estimates	83
Synthetic High School Estimates	86
10. Adding Additional Data	88
Merging	88
11. Checking Findings and Significance Online	91
AskHYS.net Website	91
Online Tool for Determining Statistical Significance	94
12. Additional Resources	95
Web Resources	95
Previously Used Computed/Calculated Variables	96
Enrollments by Year and Coding for Synthetic High School Weights	98

Introduction

Purpose

- Establish standard methods for analysis of Healthy Youth Survey (HYS) for simple frequency and crosstab analyses
- Support STATA programming – the concepts can be translated into other software languages by users, we will focus on STATA, Version 10

Audience

People who conduct or request analysis of HYS data:

- Department of Health (DOH) epidemiology/research staff
- Department of Social and Health Services (DSHS)/Division of Behavioral and Health Recovery (DBHR) or other state agency research
- Local Health Jurisdiction staff
- Other community partners
- Researchers (University of Washington (UW) or others)
- Graduate students (projects)

Uses

This manual was developed to be used by a variety of people:

- Experienced or novice STATA users new to the HYS
- People familiar with the HYS but new to STATA

While this manual provides basic information about analyzing the HYS, it is by no means exhaustive. Nor does it present the only way or the best way to run analyses. As STATA users know, there are multiple ways to program to achieve the same results.

Manual Layout

This manual is accompanied by examples of STATA coding and tables and charts. In this manual STATA coding and output are formatted as:

STATA coding is highlighted in grey

STATA output is in black boxes

This manual includes references to other sections of this document or to outside websites. References to outside websites do not imply endorsement by DOH.

Throughout this manual: STATA commands are in grey and STATA output is in black

The manual is divided up into the following sections:

1. HYS Overview
 - provides a brief overview of the survey, its history and goals
2. Getting Access to HYS Data
 - describes our data sharing agreements and terms of use
3. Getting to Know your Data
 - describes common HYS variables including demographic, 30 day and lifetime substance use, calculated and computed variables, and risk and protective factors. Also provides variable coding.
4. Getting to Know STATA
 - a table with commonly used STATA commands.
5. HYS Data Analysis in STATA
 - describes how to set up STATA for different types of data, how to explore your data, transform it and run some simple analyses
6. HYS Data Analysis – Quick Example
 - provides an example of how to run crosstab analyses in STATA using state data, county sample, census or mixed data, and ESD data.
7. Comparing State and Local Data
 - describes how to combine state and local data and compare local data to the rest of the state sample.
8. Comparing Years of Data
 - describes how to combine years of data and compare one year to another.
9. Combining Grade Levels
 - describes how to add compare multiple years of data.
10. Adding Additional Data
 - describes how to add additional data to your HYS dataset by merging.
11. Checking Findings with the HYS Website
 - describes the information available on the DOH HYS website and how to use it to verify your analysis results.
12. Additional Resources
 - Includes STATA and statistical web resources, coding for variables computed in other years, and weighting for other years.

Issues in Analyzing Healthy Youth Survey Data

The Healthy Youth Survey is a large-scale effort and involves a number of complexities which affect data analysis. These issues are discussed throughout this manual and are also summarized below. They include:

- Complex sampling designs and survey designs that vary between geographic areas
- Comparisons of state and county data
- Multiple forms of the questionnaire
- Surveying particular grades
- Response rates and valid survey rates, which are estimated based on available data before final enrollments become available
- Cell size

Sampling Designs

The Healthy Youth Survey is intended to provide information about students in public schools at a variety of geographic levels: state, county, Educational Service District (ESD), district, and school (or in the case of small schools, groups of schools). The design for these different geographic levels varies. For small groups, such as schools, school districts, and small counties, a census design is used in which all students in the grades of interest in all schools in that area are asked to participate. For larger counties and for the state as a whole, in order to increase efficiency, we use a complex sampling design in which we select random samples of schools and then recruit all students in the grades of interest in participating schools. In the absence of drawing a sample, we assume a census design for the purpose of analysis.

State level

At the state level, in order to efficiently provide information that is representative of students in public schools statewide, we select three simple random samples of public schools in the state containing grades 6, 8, and 10/12, and recruit those schools for the state sample. All of the students in these sampled schools in the surveyed grades are asked to participate. This “clustered” sampling design reduces student to student variability because students in the same school may tend to answer survey questions in similar ways; that is, the data are correlated within schools. We adjust for the clustered design by using a statistical program developed to analyze data from complex sampling designs. Since the sample is drawn by randomly selecting schools within grades, the grade/school combination (schgrd) is the primary sampling unit (PSU). (On non-identified data sets, schgrd is replaced by a sequential variable called “psu” that is converted from schgrd to remove identifying information.)

Using a statistical analysis that incorporates the design used and designating the PSU is necessary in order to obtain correct standard errors, confidence intervals, and significance tests. Using an analysis that adjusts for the clustered sampling design

compensates for the reduced variability due to intra-correlation within schools and provides error estimates that should approximate what would have been obtained with a simple random sample. Not accounting for PSUs will generally underestimate the variability in the sample and give you lower standard errors and narrower confidence intervals.

Local levels

To produce local results, schools not selected for the state sample are also invited to participate in the survey. Most local data assume a census design, in which all students in the grades of interest in that area would ideally participate. In order to use these data to generalize beyond the particular students surveyed (e.g., to intervening years or to students who may have not been surveyed) these data can be analyzed and confidence intervals obtained by using a random sample design. In these cases, the PSU is the individual student so you do not need to set a PSU and you can use any statistical program.

In large counties, where there will be a gain in efficiency by drawing a sample instead of a census design, county level samples are drawn. The criteria for drawing a sample at a particular grade in a county is that there need to be at least 30 schools in the county containing that grade. County samples are drawn by beginning with schools selected for the state sample in that county and adding an additional random sample of schools. When analyzing counties with county samples, the school building is designated as the PSU to compensate for the clustering effect (just like in the state sample).

Combined levels

In order to combine data from geographic areas that used different sampling designs (such as ESDs which include both sampled and census counties, or comparisons between a census county and the state sample) more complicated approaches may be necessary. Also, although the individual samples are designed to be self-weighting, combining data using different designs may require weighting the data. See the instructions below for setting up your data for analysis and how to designate the PSU depending on your specific data.

NOTE: For more information on sampling design see the Sampling section. For more information on data analysis depending on your sampling design see the General Set Up for Survey Analysis section

Comparisons of County and State Data

Schools in the state sample are also in a county, and so there is overlap between county and state data. Thus, when comparing a county to the state, there are two ways to do the comparison:

1. Compare the county to the *entire* state sample, including schools in the state sample that are also in the county.
2. Compare the county to the *rest of* the state sample with all of the schools from the county removed from the state sample.

For most counties, the number of schools that are in the state sample is generally too small to raise a concern and both options will give similar results. But technically, to conduct statistical comparisons, it is necessary to have the county and state respondents independent of each other. Thus we recommend the following when conducting comparisons between the county and the state:

- To determine statistically significant differences, remove the county schools from the state sample dataset prior to comparing.
- To report state point estimates and confidence intervals, use the full state sample (not the results with the county schools removed). This way your results will not contradict the previously published state results.

NOTE: For more information see the Comparing State and Local Data section.

Multiple Forms of the Questionnaire

In order to include a large number of questions on the surveys, we use four versions of the questionnaires: one version for grade 6, and three versions (A and B, or A and NS), given in alternating order, for grades 8, 10 and 12. Form B is exactly the same as Form NS, except it includes 4 additional questions on sexual behavior at the end of the survey.

Questions that are only on one version cannot be crossed with questions that are only on another version, I.e., you cannot cross a question only on Form A with a question only on Form B. Also, some items are on an optional “tear-off” sheet that schools can remove prior to administration, and these items have much smaller numbers of responses than other items. The sexual behavior questions may have an even lower number of responses. They are on the optional “tear-off” or Form B, so schools may remove them or they can order Form NS (which does not include them) when they register.

These factors also mean that while there are a relatively large number of participants in the HYS, the number available for detailed breakdowns may be much smaller. For some analyses, it may be necessary to combine data from two years to obtain adequate numbers, or even this may not provide adequate numbers.

NOTE: More detail on these issues is provided in the Survey Questionnaire section.

Surveying Particular Grades

The Healthy Youth Survey is conducted in grades 6, 8, 10 and 12. It is highly recommended that analysis be limited to a single grade. However there are situations in which combining grades may be desirable, for example when comparing to the high school estimates from the YRBS, or if there are very small numbers that cannot be reported.

NOTE: More detail on this is provided in the Stratified Analysis and Subpopulations section.

Survey Participation

Participation rates for the Healthy Youth Survey are calculated by the number of valid surveys returned divided by the total enrollment. For census designs the total enrollment for the region (i.e., the school, district, or small county) is used and for sample designs the total enrollment of schools selected for the sample (whether or not they participated) is used. Adequate participation rates are necessary to help ensure that the results are representative of the larger region.

There are a number of factors that may influence participation rates, including non-participating schools, participating schools not surveying all students in the grade, student absences, students opting out of taking the survey, and the loss of surveys during the data cleaning process.

NOTE: More information about the response rates and about analyses conducted to examine possible sources of bias in the data are available in the Participation Rates section.

Cell Size

To report results, you must have at least 5 observations per cell when running cross tabulations of state level data, or at least 10 observations per cell when running sub-state level cross tabulations.



HYS Overview

This section provides a brief overview of the survey, its history and goals.

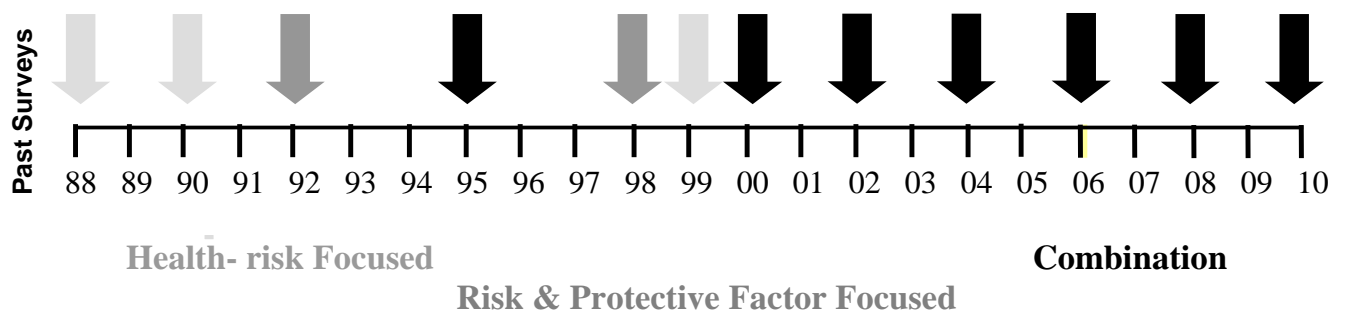
Survey History

The first “Healthy Youth Survey” to assess student risk/protective factors and health behaviors was administered to Washington students in October 2002 and it is currently scheduled for administration every two years, in the Fall of even-numbered years. This document provides a brief description of the survey’s purpose and implementation, to help provide a common understanding for community and school stakeholders.

Nationally, trends in youth behaviors and risk/protective factors have been measured using federally developed school-based surveys such as the Monitoring the Future Survey (MTF), and the Centers for Disease Control and Prevention’s Youth Risk Behavior Survey (YRBS) and Youth Tobacco Survey (YTS).

State agencies have organized twelve statewide youth surveys between 1988 and 2010. The most recent survey in 2010 was administered to over 210,000 students in 1,049 schools, in 235 (of 295 total) school districts, and in all 39 of Washington’s counties.

History of Washington’s Youth Survey Efforts



Past statewide Washington survey implementation and content have changed with each administration. For example, prior to HYS, four of the previous seven surveys were given to youth during fall months, and three were given during spring months. The surveys in 1988, 1990, and 1999 had a health-risk focus, whereas surveys in 1992 and 1998 were centered on risk and protective factors. More recent versions of the survey – 1995, 2000, 2002, 2004, 2006, 2008 and 2010 – are a combination of both. The years of administration were also not systematic (that is, there was no pattern for which years the surveys were given). The lack of consistent survey attributes meant that surveys were not necessarily comparable to each other over time, and school personnel being asked to participate in each state survey had to learn what was uniquely expected or included in each survey process.

During recent years, interest in youth surveys and need for data for planning and evaluation of science-based programs to support youth have both increased. School administration and staff were receiving requests to participate in the various state surveys, national surveys, research studies, and community-generated or school system-generated surveys.

Simultaneously, beginning in 1997 Washington began to implement required student achievement testing as part of evaluating educational systems. The Washington Assessment of Student Learning (WASL) test was required for administration to students in grades 4, 7, and 10. Implementation of this test disrupted several days of instruction for schools in the spring of each year.

State Superintendent of Public Instruction Terry Bergeson determined in 1998 that state agencies must cooperate to administer only one survey of youth behaviors every two years. In response, staff from OSPI, DBHR, DOH, the Department of Commerce and the Governor’s Family Policy Council formed the Joint Survey Planning Committee [JSPC].

Common Goals

The JSPC first identified issues of interest to each agency and to local constituents. These included:

- Describing school, community, family, and peer-individual risk and protective factors (similar to the “Communities that Care” model developed by the University of Washington Social Development Research Group – including Dr. Hawkins and Dr. Catalano)
- Describing youth health habits, risks, and outcomes
- Gathering state-level data in a consistent way (with predictable timing and using comparable measures over time)
- Supporting local-level data collection and use for planning/assessment and evaluation of programs to serve youth.

Agreement about Survey Features

After agreeing on common goals, agencies negotiated specific features of the survey – to be called the “Healthy Youth Survey” – necessary to achieve these goals. Agreed features of the survey are as follows:

1. Only one statewide school-based survey of youth will be administered, supported by all state agencies. State agencies in the JSPC agreed to not conduct independent surveys of schools to gather youth data. This agreement should increase efficiency of surveys that are conducted, and reduce the burden on schools for surveys. Agencies understood that this would mean challenges in coming to agreement on content for a unified survey.
2. A simple random sample of schools will be recruited at the state level, and county samples will be provided (as appropriate). Methods used to identify a sample of schools to be included in state surveys had changed over time. These changes can have some impact on results, and also complicate year-to-year comparisons of data. Identification of a simple sampling plan makes the survey easier to manage and analyze. The disadvantage of this method is that few schools in any particular area would be included in the state sample, but the JSPC agreed that local schools would be provided some way to “piggyback” (voluntarily participate) to gather local-level data, and county samples could be drawn for counties that are large enough to do so.
3. The survey will be consistently administered in the fall of even years (2002, 2004, etc.). This predictable timeline will avoid conflict with WASL testing, allow school and communities to have data available for spring grant writing/needs assessment activities, and help school administrators to plan ahead for participation. Gathering of data in the fall does make comparison to some national surveys (YRBS, YTS) more difficult, because those surveys are conducted in spring months, when youth are older and more likely to engage in risky behaviors.
4. The survey will be given to 6th, 8th, 10th and 12th graders. Data collection of these grades on a two-year cycle will enable communities and state agencies to watch “cohorts” of youth over time. In other words, the 6th graders who take the Fall 2006 survey will participate as 8th graders in the Fall 2008 survey, and participate as 10th graders in the Fall 2010 survey, and so on. In comparison to national surveys such as the YRBS, which is given to 9-12th graders, this method will collect more data from younger youth, which is important for early prevention efforts.
5. The survey will be given to youth using survey booklets with a one-page tear-off answer sheet. In comparison to past school surveys, which were given as scannable booklets, having a separate scannable answer sheet will dramatically increase the speed of scanning and delivering results. It will also decrease the cost of printing. This layout will make it easier to provide the survey in different languages. It is possible that this change will increase the number of mistakes that youth might make as they “bubble” their

answers on a separate page from the questions, and might also increase the time it takes for youth to complete any question – the JSPC investigated this administration change in a small HYS pilot prior to the first administration in 2002, and used results to identify a maximum number of questions that most students could complete during a class period using a separate answer sheet from that pilot. We also found that the number of illogical answers (either due to mistakes or to students purposely “drawing patterns” on answer sheets rather than answering questions) was not excessive and could still be managed using logic checks during normal quality control screening.

6. The survey will be given to 8th, 10th, and 12th grade youth as a two-form “interleaved” administration. To manage the length of the survey with the breadth of information desired by agencies and stakeholders, there will be a “Form A” and a “Form B” for the survey. Alternately seated students will receive “Form A” and “Form B”, but it will not be obvious to youth sitting next to each other that they have different versions. All youth will have the same “core” questions in their surveys. Youth who complete “Form A” will go on to answer additional questions about risk/protective factors (similar to past WSSAHB surveys) while youth who complete “Form B” will answer additional questions about health risks and outcomes (similar to past YRBS). See Figure 1 for an illustration of the survey layout. All sixth graders will have a single version (“form C”) that includes similar items to A and B, to be negotiated among the agencies, but is shorter and in some cases includes simplified wording to assure that younger students can successfully complete it.

Survey Questionnaires

The “core” items for the survey include about 30 questions to describe:

- Student demographics
- 30-day use and/or lifetime use of alcohol, tobacco, and other drugs
- Key violence-related questions (weapon-carrying, perceived safety)
- School-specific asset questions (attachment to school, opportunities for involvement)
- Depression.

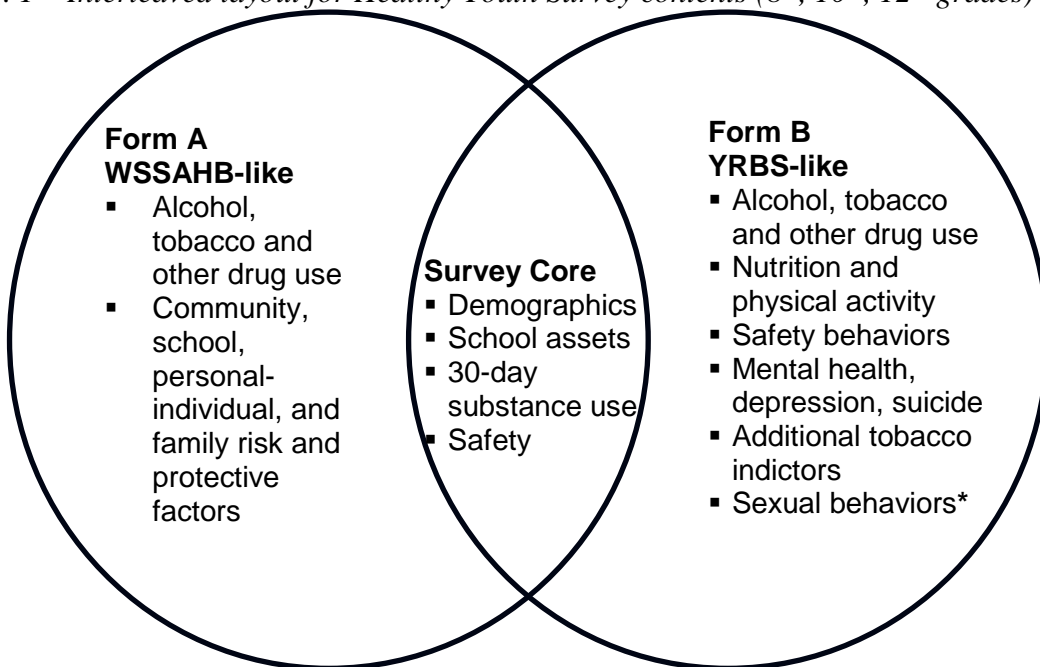
The questions for “Form A” have been identified by a working group of OSPI, DBHR, and Commerce and constituents. The questions for “forms B/NS” have been identified by a working group of DOH and constituents.

The 6th grade survey is a single version, with fewer questions. Questions are consistent with the longer form A and forms B/NS questionnaires. These differences are because 6th grade youth do not have reading skills to complete a longer survey, because some questions applicable to older youth are not appropriate for younger youth, and because there are more small buildings for 6th graders than for older grades where giving results would be impacted by having only half the youth take a particular version.

NOTE: Each survey has a “Tear-off” section that includes some more sensitive forms that schools can decide to include or not.

The survey forms for Healthy Youth Survey 2010 are available for viewing or downloading as a PDF on the DOH website (see left hand side of main page): <http://www.doh.wa.gov/DataandStatisticalReports/HealthBehaviors/HealthyYouthSurvey.aspx>

Fig. 1 – Interleaved layout for Healthy Youth Survey contents (8th, 10th, 12th grades)



Core Survey Items

- Demographics
- Alcohol, tobacco, other drugs
- Key violence-related items
- School-specific Risk and Protective Factor items
- Depression

35 questions on forms A and B/NS, and 18 questions on forms A, B/NS and C

Tear-Off Section Items

Optional section at the end of the survey that is perforated so districts or schools can tear off the questions.

Examples of tear-off questions:

- Family risk and protective factors
- Physical abuse and dating violence
- Asthma
- Pregnancy/STD prevention education
- Sexual behavior (only on Form B)

Need for Community Partnerships

The JSPC has worked to develop this survey design for Washington State, and has considered and attempted to resolve as many issues as possible for the state as a whole. Having done this, there is still a need for strong community partnerships to support the implementation and use of the survey, to make it successful. Issues for discussion at the local level may include:

- Supporting participation -- In some areas of the state, for a variety of reasons, school administrators and staff are reluctant to take the state survey. Local stakeholders who want to support gathering local-level data may wish to identify these areas, and create plans for targeted recruitment efforts. *The final decision to participate in the survey is made by the school administrator*; however, local recruitment efforts can focus on communicating needs for data, creating partnerships to support youth including through assessment-driven planning, demystifying the current survey efforts, and providing support for survey participation.
- Small schools – to receive school building results, it is necessary to have at least 15 valid responses in a grade (to protect the confidentiality of the students). Some small schools may not have enough students in a grade to be eligible for their own results, but their results are aggregated into higher level results (district, county, or ESD– depending on any sampling). If there are similar small buildings, they have the opportunity to join together and form a “consortium.” To receive consortium results, the schools must register separately and then submit a combined consortium registration form (which must receive approval from DOH). Neither school will receive its own results, but the combined consortium results will be available to both schools. To receive consortium results, the combined schools still have to have at least 15 valid responses per grade. More than two schools can be combined if desired.
- Effective participation – A school building, district, or county-wide buildings and districts, may agree to participate, but ineffectively administer the survey. For example, a survey might be distributed to only part of the students, or a date for the survey might be selected when significant groups of youth are missing (such as when the band members go on a trip). Local partners might choose to congratulate schools on their desire to participate, and also encourage schools to achieve a goal of at least 70% participation among youth in any grade group. Examining past survey participation rates may be helpful.
- Use of data – Community members should discuss how to *use* and talk about their data prior to receiving their reports. This includes discussion about how to manage media messages, and/or respond to media questions about the data. This planning may be very useful during recruitment activities, as school administrators may have fears and questions about how their data (or their region’s data) might be portrayed in the media.

Washington's nine Educational Service Districts (ESDs) are central points of recruitment activity for the state. The ESD school-based tobacco program coordinators have facilitated local stakeholder processes to support recruitment planning for the survey within ESD regions throughout the state.

Survey Implementation Schedule

Example for 2010 Administration:

- **Dec 2009:** state and county samples identified, Human Research Review Board approval obtained
- **Jan 2010:** survey content finalized; recruitment letters sent to Washington school administrators
- **Feb 2010-June 30, 2010:** recruitment of schools to participate
- **June 30, 2010:** last day for schools to sign up for the survey
- **October 18-22, 2010:** schools administer survey to youth
- **March 2011:** school district superintendents receive building/district reports, local health jurisdictions receive county results, ESD superintendents receive ESD reports, state results available

For the 2010 administration, funds were available to support “no-cost piggybacking”, i.e., allowing any non-sampled schools to register for the survey at no cost. This funding is not guaranteed for future HYS administrations.

Participating schools must agree to have a survey coordinator who will attend a short training, send and post the approved parent and student notification, and administer the survey according to directions.

Sampling

A simple random sample of all schools in the public school system is drawn, with the following restrictions: schools must contain at least one of grades 6, 8, 10 or 12, and there must be at least 15 students in that grade (based on the most recent enrollment figures from OSPI). Within the participating schools, all students in the surveyed grades are asked to participate.

- Three samples are drawn: one for 6th grade schools, one for 8th grade schools and one 10th and 12th grade schools combined (since the grades are often in the same school).
- Non-sampled schools are also invited to participate in the survey; participation allows these schools to obtain their own school results and to contribute to district-, county-, and ESD-level results.
- County samples are drawn for counties with more than 30 schools in a grade. In 2010, all grades (6, 8, 10, 12) were sampled in King, Pierce, and Snohomish, grades 6 and 8 in Clark and Spokane, and grade 6 in Thurston had county samples drawn.

The responses from these sampled schools were used to produce county-level estimates.

- For all other counties, the responses from all schools are used to produce the county-level estimates, whether they are in the state sample or a piggyback school.

NOTE: For information on sampling, see technical notes at:

<http://www.doh.wa.gov/DataandStatisticalReports/HealthBehaviors/HealthyYouthSurvey.aspx>

<http://www.askhys.net>

Survey Participation Rates

Calculating response rates for the Healthy Youth Survey is complicated by the loss of data both to non-response and during cleaning, the various levels of aggregation, and the availability of enrollment figures. Thus, both the numerators and denominators for the response rates need explanation.

Loss of data to non-response and during cleaning

Reasons for data being unavailable included 1) refusals to participate by some schools, 2) students being absent, refusing to participate, or being away from their school during survey administration because of involvement in programs such as “Running Start” which allow them to take classes at junior colleges, and 3) cases discarded during cleaning based on an algorithm that includes the amount of missing and inconsistent responses, responses to a question asking about fictitious drug use, and responses to a question asking about honesty of responding.

Levels of aggregation

Response rates for local data were calculated by dividing the number of valid surveys in the sampled schools by the total enrollment in schools selected for the sample. Although issues affecting data lost to non-participation and data discarded during cleaning may be different, the vast majority of unavailable data was due to non-participation in the survey, and only about 4% of data collected is discarded during cleaning. Thus, these figures (actually the valid survey rates) provide estimates of the response rates.

In 2010, at the state level we calculated both response rates (calculated by dividing the number of participants in the sampled schools by the total enrollment in schools selected for the sample) and valid survey rates (calculated by dividing the number of valid surveys in the sampled schools by the total enrollment in schools selected in the sample). We also calculated response rates for the local reports with the enrollment data available at the time of the release of the data, which was 2008-2009 enrollment data. See Tables below for information about the state response rates and participation from HYS 2002, 2004, 2006, 2008 and 2010.

HYS Student Response Rates for State Sample by Year

Grade	2002	2004	2006	2008	2010
Grade 6	61%	68%	78%	76%	83%
Grade 8	65%	73%	70%	77%	77%
Grade 10	44%	58%	63%	60%	67%
Grade 12	40%	49%	51%	50%	53%
Total	50%	61%	65%	66%	72%

Number of HYS Participants (with Valid Surveys) by Year, State Sample and Piggybacked Schools

Grade	2002		2004		2006	
	Sample	Piggyback	Sample	Piggyback	Sample	Piggyback
Grade 6	7,952	32,641	8,825	46,031	7,862	46,222
Grade 8	7,473	32,792	8,912	47,970	8,466	45,996
Grade 10	5,127	26,884	8,514	41,465	8,059	36,582
Grade 12	4,133	20,333	6,280	30,315	5,876	26,032
Total	24,685	112,650	32,531	165,781	30,263	154,832
Total (Sample & Piggyback)	137,335		185,095		198,312	

Grade	2008		2010	
	Sample	Piggyback	Sample	Piggyback
Grade 6	9,068	48,566	11,549	46,166
Grade 8	8,730	50,687	9,723	48,431
Grade 10	6,907	46,181	6,889	45,997
Grade 12	5,641	35,071	5,908	37,390
Total	30,346	180,505	34,069	177,984
Total (Sample & Piggyback)	210,851		212,053	

Availability of enrollment figures

The denominators used for calculating response rates and valid survey rates are drawn from OSPI October enrollment figures (available online at the Office of the Superintendent for Public Instruction website). The enrollment figures are reported by schools and compiled by OSPI, and to date the final results have not been available when the Healthy Youth Survey results are reported in the spring of the following year. In order to provide the “best available” estimates of response rates with the reports, these are calculated using the previous year’s enrollment figures. When the final enrollments become available, we re-calculate the state response rates, and recommend that local response rates be re-calculated as well.

Importance of Participation Rates

Participation or response rates are determined by the number of valid surveys returned divided by the total enrollment (or estimated enrollment before final enrollment figures become available). In general, the following guidance may be used when using county-level Healthy Youth Survey data. If the response rates are:

- 70% or greater: The HYS results are probably representative.
- 40-69%: The HYS results may be representative of students but further examination of other data (by school or district) to identify any important differences between participants and non-participants should be completed before generalizing results to the county.
- Less than 40%: Response rates less than 40% are quite low, and these HYS results should not be interpreted as representative of the county.

If groups of students did not take the survey, there may be limitations even if there is a high response rate. Data for grades with less than a 70% response rate should be interpreted cautiously.

NOTE: For information on participation rates, technical notes at:

<http://www.doh.wa.gov/DataandStatisticalReports/HealthBehaviors/HealthyYouthSurvey.aspx>

Validity, Reliability and Generalizability

Validity is the degree to which the results are likely to be true, believable and free of bias and can be generalized to a larger population. A survey item is valid if it accurately measures the concept it is intended to measure. A number of methods are used to help ensure validity, including:

- Sampling
- Using items from established youth surveys such as the YRBS and YTS
- Piloting new untested questions with youth
- Data cleaning

Only “valid” surveys are included in the final dataset. The contractor uses a series of quality controls to remove data that were incomplete, obviously inaccurate, or internally inconsistent. On average about 4% of the returned surveys are culled during the process and are removed from the final dataset. Quality control checks include looking for:

- Inconsistent answers
- Evidence of faking high level of substance use
- Dishonesty
- Wrong grade

Reliability is the extent to which a survey measure, procedure or instrument yields the same result on repeated trials. A survey item is reliable if it consistently produces the same results under the same circumstances. We try to maximize reliability by:

- Using standardized administration procedures (e.g., coordinator training, teacher training, written instructions, teacher stays in room but at desk, single class period to avoid discussion, absent students do not make up).
- Providing a safe and confidential environment.
- Informing students about the importance of survey.
- Keeping student responses confidential (no student name or other identifying information and students place own answer sheet in envelope).

Confidence Intervals

Confidence intervals are used with the survey data to give an estimate of how accurately you can generalize from samples, such as the state sample, to a larger population, such as students in public schools in Washington, assuming that the data are not biased. Specifically, the 95% confidence interval gives the range that should contain the true population value 95% of the time.

NOTE: For information on confidence intervals, see:

<http://www.doh.wa.gov/DataandStatisticalReports/HealthBehaviors/HealthyYouthSurvey.aspx>
<http://www.askhys.net>

Bias Analysis

Survey responses are often used to estimate the frequency of behaviors or other characteristics in a population larger than those who actually completed the survey. Thus, while only a portion of public school students took the Healthy Youth Survey in 2010, we would like to use their responses to characterize all 6th, 8th, 10th and 12th graders in Washington. This is only possible if those who participated in the Healthy Youth Survey are not different in their behaviors from those who did not participate. If they are different, we say that the survey is biased and we are then limited in our ability to generalize the results to all students. Bias represents systematic error and is different from the random fluctuation that is measured by confidence intervals. Comparisons could be done using information from sources such as the census, school achievement test results, or other demographic information.

A bias analysis for HYS 2010 HYS has been completed and will be posted at <http://www.doh.wa.gov/DataandStatisticalReports/HealthBehaviors/HealthyYouthSurvey>.

The bias analysis found that the results from the 2010 Healthy Youth Survey state sample can be generalized to students attending non-alternative public schools in Washington State. However, due to a low proportion of grade 8 schools administering the sexual behavior questions and differences between schools that did and did not administer the sexual behavior questions, caution should be taken when generalizing grade 8 sexual behavior results to Washington State students.

That the 2010 Healthy Youth Survey can be generalized to students attending non-alternative schools was in agreement with the findings of the 2004 bias analysis, which found that the results of the 2004 Healthy Youth Survey can be generalized to all public school students in 6th, 8th, 10th and 12th grades, except to students in alternative schools. The sexual behavior questions were new in 2010.

NOTE: For information on bias analyses and the results of the 2010 HYS bias analysis once completed, see:
<http://www.doh.wa.gov/DataandStatisticalReports/HealthBehaviors/HealthyYouthSurvey>

2

Getting Access to HYS Data

This section describes HYS data sharing agreements and terms of use.

Data Sharing and Human Research Review Requirements

The ability to share and report data that contains information about geographic levels lower than statewide is limited by protections of confidentiality for participants and by issues of identifiability for schools and school districts. Data sharing agreements provide information about these requirements, as well as other issues important to data users. This information is explained below and a sample data sharing agreement is available on request. This information is current for the HYS 2010.

Protections of Confidentiality for Participants

Importance of anonymity. Prior to participation, all survey participants are informed “Your answers to these questions are *anonymous*. This means that no one will see your answers or know which answer sheet you completed.” Thus, data sharing procedures are designed to assure anonymity. These procedures are part of the human research review process and are included in the approval by the Washington State Institutional Review Board (WSIRB).

Availability of data with geographic identifiers. Outside of the state agencies participating in the HYS, access to data files containing individual level data (e.g., SPSS, SAS or STATA files) and geographic identifiers is very limited. Because local health jurisdictions (LHJs) have a long history of ability to handle confidential data and of sharing data with DOH, they have access to the data with a data sharing agreement. Other local organizations wishing information about that geographic area are referred first to the LHJ; and DOH acts as backup to the LHJ. Researchers who wish access to the individual level data with geographic identifiers must submit an Exempt Determination Request to the WSIRB. Although educational institutions such as schools and school districts are important participants in the HYS, educational institutions that might have access to students and information about students drawn from student records do not have access to identifiable

data because information from the HYS, in combination with additional information available to the educational institutions, might make the students identifiable.

Availability of data without geographic identifiers. Statewide data that does not contain geographic identifiers (i.e. school, school district, ESD, or county identifiers) cannot be used to identify individual students. Thus, a non-identified statewide data set (from which all geographic identifiers have been removed) is available to legitimate researchers with a data sharing agreement. Interactive access to frequencies and crosstabs based on state sample data for 2002-2010 is available at www.askhys.net. The website includes both a data query system and topic specific fact sheets. HYS reports are available at the state, county and ESD level and with permission from the district superintendents, at the district or school level.

Reporting data while retaining anonymity. LHJs and researchers, prior to receiving HYS data, must sign a data sharing agreement stating that they will comply with procedures approved by the WSIRB. These include reporting requirements to protect individual identifiability. **These requirements state that for data identified by a geographic level, less than statewide, frequencies will only be reported where there are at least 15 valid surveys and crosstabs other than grade level will only be reported where there are at least 10 cases per cell.** At the state level, frequencies in crosstabs can be reported if there are at least 5 cases per cell. They also agree to comply with reporting requirements regarding identifiability of schools, described below.

Identifiability of Schools and School Districts

School and school district level information. The HYS planning committee considers that schools and school districts are the “owners” of their data reports, subject to any state and federal laws pertaining to public access to information. Consistent with this, at the time of registering for participation, schools may “opt out” from receiving a school-level report of results, in which case the report will not be generated. Individuals desiring reports of school or school district results are referred to the school or school district.

Reporting data identifiable by school or school district. If a data user wishes to report data in such a way that the results are identifiable by school or school district, he or she must obtain written permission from the principal or superintendent. Otherwise, data from at least three schools and three school districts must be combined for reporting purposes.

Data Sharing Agreements

Data sharing agreement. Prior to receiving individual level data, LHJs or researchers must sign a data sharing agreement, which includes the data sharing agreement *per se* and an Attachment (A). The agreement must be signed by the individual with authority to sign for the organization. Attachment A must be signed by each of the data users working with the data.

Statutory authority for this data sharing is based on Interlocal Cooperation Act, RCW 39.34, which allows agencies to jointly share their powers and contract with one another, provided the use of the data is for a legally authorized activity and not used in a manner which exceeds the requesting department's jurisdiction. In the data sharing agreement, the receiving agency agrees to (1) not release the data file without the agreement of the agency providing the data; (2) not use the data to identify individual students or report the data in a way that individual students can be identified and (3) not report the data in ways that identify schools or school districts, unless schools agree in writing and students cannot be identified. It also includes provisions for receiving, storing and destroying the data file. A sample data sharing agreement is available on request.

Receiving the data. Data are sent by a secure means. This generally means that the data are uploaded to a secure file transfer site.. The code for school identifiers (if needed) is provided separately. Data are available in SPSS, SAS, or STATA; other programs may also be available.

For more information

More information about data sharing requirements is available by contacting Lillian.Bensley@doh.wa.gov.

More information about the WSIRB is available at <http://www.dshs.wa.gov/rda/hrrs>

3

Getting to Know Your HYS Data

This section describes common variables in the 2010 Healthy Youth Survey data set. It includes information on:

- Demographic variables
- 30 day and lifetime substance use variables
- Calculated and computed variables, including how to code them in STATA
- And risk and protective factors

Most variables consist of a letter such as c, d, f, h, etc. followed by a number. The letter prefixes give you an idea about the variable topic:

C - school climate

D - alcohol, tobacco and other drugs

F - family risk and protective factors

G - demographics

H - health

L - quality of life

M - community risk and protective factors

P - peer and individual risk and protective factors

S - school risk and protective factors

Other computed variables are usually acronyms such as bmi or yqols. Computed risk and protective factor scales consist of the word risk followed by a number.

For a detailed description of HYS variables see the Crosswalk of HYS variables from 2002, 2004, 2006, 2008, and 2010 available on AskHYS in the QxQ online data analysis section:

<http://www.AskHYS.net>

Demographic Variables

coname and conum

Depending on the type of data set you have, you may or may not have these variables. Each county can be identified with either the coname and conum variables. Coname is a string variable that identifies the county name, “Adams County”. Conum is a unique two digit numeric code that represents each of the 39 counties in alphabetical order starting with Adams (conum==1) and ending with Yakima (conum==39).

Adams=1, Asotin=2, Benton=3, Chelan=4, Clallam=5, Clark=6, Columbia=7, Cowlitz=8, Douglas=9, Ferry=10, Franklin=11, Garfield=12, Grant=13, Grays Harbor=14, Island=15, Jefferson=16, King=17, Kitsap=18, Kittitas=19, Klickitat=20, Lewis=21, Lincoln=22, Mason=23, Okanogan=24, Pacific=25, Pend Oreille=26, Pierce=27, San Juan=28, Skagit=29, Skamania=30, Snohomish=31, Spokane=32, Stevens=33, Thurston=34, Wahkiakum=35, Walla Walla=36, Whatcom=37, Whitman=38, Yakima=39.

distname, distnum, and codis

Depending on the type of data set you have, you may or may not have these variables. District level data should never be reported unless you have the written approval from the school district.

Distname is a string variable that identifies the school district name, “Almira School District.” Distnum is a three digit numeric code for the district. These codes are developed by OSPI (information is available on the OSPI website). The distnum variable is only unique within a county. Codis is a unique five digit numeric variable for each county – district combination. Codis should be used instead of distnum unless you only have data from a single county.

schname, schnum, schgrd and psu

Again, depending on your data set you may or may not have these variables and school building data should only be reported with written permission from the school district superintendent.

Schname is a string variable that identifies the school building name. Schnum is a unique four digit numeric code for the school building. These codes are also developed by OSPI. Most schools have codes between 1500 and 4999. Private schools have numbers between 8000 and 8999. Numbers between 9000 and 9999 are special cases and are not official OSPI codes.

School codes are associated with physical school buildings and buildings may open, close, move or change their grade levels over time, so it is important to verify that your school numbers, grades and names match when comparing data over time.

Schgrd is a six digit numeric code that combines both the school building code and then the grade level of the respondent.

Psu is actually the same type of variable as schgrd, it contains a unique identifier for each school building code and grade level of the respondent. Psu is used in de-identified datasets, that is, those datasets that don't identify the name or number of the school building.

grade and hdrgrade

When conducting analysis by grade always use the grade variable, never use hdrgrade. The grade variable has the proper four grade response options; 6th, 8th, 10th and 12th. The hdrgrade variable has more options that are associated with the types of other grades in the school building.

Age g01, g02

In the HYS dataset there are two different variables for age, g01 is asked on forms A and B for 8th, 10th and 12th graders, while g02 has less response options and is asked on form C for 6th graders.

Gender g05

The variable for gender is g05, females are response option 1 and males are response option 2.

formtype

The HYS has four survey forms. All 6th graders take Form C. Half of 8th, 10th and 12th graders take Form A and half take Form B/NS. Some variables cannot be cross tabulated because they are on different surveys (i.e., one variable is on Form A and the other is on Form B/NS). If you run a crosstab and STATA says there are "no observations" it could mean that you are trying to cross variables on different surveys. Formtype can be useful if you want to investigate which form your variable is on or if you want to restrict your analysis to include only respondents from one of the forms.

Race/Ethnicity g06, g06a, g06b, g06c, g06d, g06e, g06f, g06g

In the HYS dataset, there is a calculated race/ ethnicity variable that includes the following response options, g06:

1. Asian or Asian American
2. American Indian or Alaska Native
3. Black or African American
4. Hispanic or Latino/ Latina
5. Native Hawaiian or other Pacific Islander
6. White or Caucasian
7. Other
8. More than one race/ethnicity marked

The HYS race/ethnicity question asks youth to mark any race/ethnicity that applies. In g06, respondents who only selected one race/ethnicity are counted in that particular

response option. Youth who checked multiple races/ethnicities, are placed into an additional 8th category: More than One Race/ethnicity marked. I.e., if a respondent only selected “Asian” then they are counted as “Asian”, but if they selected “Asian” and “Black” they would be counted as “More than one race/ethnicity”.

To recode g06 to only include the main six race/ethnicity categories:

```
gen race=g06
recode race 1=5 2=4 3=3 4=2 5=6 6 6=1 7=. 8=.
lab def newrace 1"White" 2"Hispanic" 3"Black" 4"Indian" 5"Asian" 6"Pacific
Islander"
lab val race newrace
```

Or to combine Asian and Pacific Islanders together for 5 race/ethnicity categories:

```
gen race=g06
recode race 1=5 2=4 3=3 4=2 5=5 6=1 7=. 8=.
lab def newrace 1"White" 2"Hispanic" 3"Black" 4"Indian" 5"API"
lab val race newrace
```

There are also individual variables for each race/ethnic response option:

- g06a: Asian or Asian American
- g06b: American Indian or Alaska Native
- g06c: Black or African American
- g06d: Hispanic or Latino/ Latina
- g06e: Native Hawaiian or other Pacific Islander
- g06f: White or Caucasian
- g06g: Other

You want to choose an individual race variable if you are looking at one particular race and need to capture all of the youth who checked a certain race, you should use the individual race variables (g06a-g06g). I.e., if a respondent only selected “Asian” and “Black” they would be included in both variables any “Asian” responses in g06a and any “Black” response in g06c.

For example, in 2008 in the state sample, looking at variable g06b, a total of 1,833 youth checked American Indian or Alaska Native as a response option. In the rolled up g06 variable, there are only 1,076 American Indian youth listed. That is because 757 of those American Indian youth also checked another race and are included as “More than one race/ethnicity marked” in g06.

If you wanted to recode race to be non-Hispanic White, Hispanic, non-Hispanic Black, non-Hispanic American Indian or Alaskan Native, and non-Hispanic Asian or Pacific Islander :

```
gen race=.
replace race=1 if g06==6
replace race=3 if g06==3
replace race=4 if g06==2
replace race=5 if (g06==1 | g06==5)
replace race=2 if g06d==1
lab def newrace 1"White non-H" 2"Hispanic" 3"Black non-H" 4"Indian non-H" 5"API
non_H"
lab val race newrace
```

Warning: In 2006, the g06a and g06b variables were switched from coding in 2002 and 2004:

- In 2002 and 2004 g06a was American Indians/Alaskan Natives, but in 2006, 2008 and 2010 g06b was American Indians/Alaskan Natives
- In 2002 and 2004 g06b was Asian or Asian American, but in 2006, 2008 and 2010 g06a was Asians or Asian Americans.

This is only a problem if you are combining more than one year of data and trying to compare race/ethnicity results by year. Most datasets were fixed, but you should double check before making comparisons.

Language Spoken at Home, g07_06, g08

In the HYS dataset there are two different variables for language spoken at home, g07_06 is asked on Forms A and B/NS for 8th, 10th and 12th graders. Three new response options; Chinese, Korean and Japanese were added to g07_06 in 2006. In 2002 and 2004 the variable was g07. The variable g08 is asked on form C for 6th graders. It only includes three response options; English, Spanish or Other and has not changed over time.

Parental Education Status, g17, g18

In 2006 the variable for maternal education was g17 and paternal education was g18. The question wording and the response options changed from 2004 to 2006. In 2004 and 2002, maternal education was g10 and paternal education was g09. Currently the two survey questions are, "How far did your mother/father get in school?". In 2004 and 2002, the two survey questions were "What is the highest degree or diploma your mother/father earned?". The current question also includes a "does not apply" option.

Some youth do not know their parents level of education, so these questions have always had a large number of "don't know", "doesn't apply" and missing values. For example, for maternal education in 2010 about 30% of 8th graders, 18% of 10th graders and 11% of 12th grades responded don't know/doesn't apply/or left it blank. In 2004 and 2002 about 50% of 8th graders, 27% of 10th graders, and 16% of 12th grades responded "don't know" or left it blank. The numbers of "missing values" are even a bit larger for paternal education. Due to the large number of missing values, these questions should be used with caution, especially for 8th graders.

The maternal education variable has been used as a proxy measure for low socioeconomic status. See the section on Socioeconomic Status (SES) under Computed Variables.

Substance Use Variables

Some of the 30 day and lifetime substance use variables are created from recoded variables or by combinations of variables. The 30 days use questions ask about the use of a substance in the past 30 days and are available in with all of the original responses or in a collapsed version with no use in the past and any use. Many of the lifetime substance use variables are recoded from questions that ask about the age of first use.

The following are lists of the 30 day and lifetime use variables from 2010. Substance use questions rotate on and off the survey. For more information of variables including which survey form they are on and the survey item number, see: HYS Question Crosswalk available on <http://www.AskHYS.net> in the QxQ data analysis section.

30 Day Substance Use Variables in 2010

For each of the substance use questions there are two variables:

- One includes all of the responses (i.e. d14 is the 30 day use variable for cigarettes with the response options 0 days, 1-2 days, 3-9 days 10-29 days, All 30 days)
- Another one which includes just the collapsed response options “yes” for use on any days and “no” for use on 0 days.
 - Cigarettes: d14 or collapsed none/any d14use. On all forms.
 - Chewing tobacco: d15 or collapsed none/any d15use. On all forms.
 - Cigars: d16 or collapsed none/any d16use. Only on form B/NS.
 - Alcohol: d20 or collapsed none/any d20use. On all forms.
 - Marijuana: d21 or collapsed none/any d21use. On all forms.
 - Illegal drug not including alcohol, tobacco or marijuana: d63 or collapsed none/any d63use. On all forms. This was a new variable in 2004. In 2002, 30 day drug use questions were asked in a different order and are not comparable.
 - Illegal drug not including alcohol or tobacco: d68 or collapsed none/any d68use. This is a combination of d63 and d21. On all forms.
 - Pain killers: d75 or collapsed to none/any d75use. Only on forms A and B/NS.
 - Candy flavored tobacco: d84 or collapsed none/any d84use. Only on form B/NS.

Lifetime Substance Use Variables in 2010

- Cigarette, just a puff: d01 - asked as age (p19), but recoded for lifetime. Only on form A.
- Alcohol, sip: d05 - asked as age (p21), but recoded for lifetime on form A and B. On form C p21 - asked as ever yes/no. On all forms.
- Marijuana: d06 - asked as age (p17), but recoded for lifetime on form A and B. On form C p18 - asked as ever yes/no. On all forms.
- Steroids (without prescription): d07 - asked as ever yes/no. On form A.
- Cocaine: d08 - asked as ever yes/no. On form A.
- Inhalants: d11 - asked as ever yes/no. On form A and C.
- Other illegal drugs: d12 - asked as ever yes/no. Only on form C.
- Methamphetamines: d10 - asked as ever yes/no. On form A.
- Heroin: d89 - asked as ever yes/no. On form A.

Age of First Substance Use in 2010

Variables that ask the age of first use for substances can be used to calculate the average age of first :

- Cigarette, just a puff: p19 on form A.
- Alcohol, sip: p20 on forms A and B.
- Alcohol, began regular drinking: p22 on form A.
- Marijuana: p17 on forms A and B.

Prior to running the mean age, you need to recode the respondents who have not used the substances to missing and change the other response options to match the age level they represent. To calculate the age of first sip of alcohol by grade:

```
gen agesip=p20
recode agesip 1=. 2=10 3=11 4=12 5=13 6=14 7=15 8=16 9=17
svy:mean agesip, over(grade)
```

Binge Drinking

Binge drinking in the past two weeks was asked on all forms in 2010, d61. A collapsed yes/no variable is computed - d61bool.

Levels of Alcohol Use

The computed levels of alcohol use variable, cdv, is on all forms in 2010. The cdv variable combines past 30 day alcohol drinking and binge drinking to break drinking down into the following levels:

- No drinking .
- Experimental drinking - 1-2 days drinking and no binge drinking.
- Problem drinking - 3-5 days drinking and/or one binge.
- Heavy drinking - 6 or more days drinking and/or two or more binges.

Warning: In 2006 the levels of alcohol use variable, `cdv`, was calculated incorrectly. In 2006, the binge drinking question was only asked on form A, so the levels of alcohol use should only include respondents who answered form A. To fix the problem, use the following STATA coding:

```
*Fixing 2006 cdv
replace cdv=. if formtype~="A" & year==2006
```

Usual Sources of Alcohol

HYS ask youth about where they usually get their alcohol. Youth were asked to check all sources that applied, so there are multiple variables – `d76a`, `d76b`, `d76c`, `d76d`, `d76e`, `d76f`, `d76g`, `d76j`, `d76k` and `d76l`. In 2010 the response options were changed – `d76h` “on the internet” and `d76i` “stole it from a store” were replaced by `d76k` “older brother or sister” and `d76l` “family celebration/ceremony/party”. A variable with the total number of respondents to the question is `d76_n` is also available.

Often we want to recode this variable to only look at the youth who actually got alcohol. The question’s first response option is “I did not get alcohol in the past 30 days”, so the recommended method for recoding is to create a new variable for each of the sources and replace the “did not get” respondents to missing.

```
gen boughtstore=d76b
replace boughtstore=. if d76a==1
```

This variable could also be recoded by restricting it to only include current alcohol users. The method is not recommended because this question does not mention “using” it only mentions “getting alcohol”.

Usual Sources of Tobacco

Question `d56` ask youth about where they usually get their tobacco. Often we want to recode this variable to only look at the youth who actually used or got tobacco. The question’s first response option is “I did not use tobacco in the past 30 days”, so the recommended method for recoding it is to set that response to missing.

```
gen tobsource=d56
recode tobsource 1=. 2=2 3=3 4=4 5=5 6=6 7=7 8=8
lab def tobsource 2"store" 3"vending" 4"gave money" 5"bummed" 6"older person"
7"stole" 8"other"
lab val tobsource tobsource
```

This variable could also be recoded by restricting it to only include current tobacco users. The method is not recommended because there are youth who say “I did not use tobacco in the past 30 days” in the usual source question, but also say that they used a tobacco product in the past 30 days (`d14`, `d15`, `d16`, `d17`, `d18`, `d19`). There are also youth who did not use a tobacco product in the past 30 days, but responded that they usually get their tobacco from one of the options. Unfortunately, it is difficult to reconcile these differences, as youth may have used tobacco but did not obtain it, or they may have obtained it but not used it in the past 30 days.

Susceptibility to Smoking

The measure of susceptibility to smoking was developed by researchers in California to identify youth who have not made strong commitments to remaining smoke-free. This measure has been found to be predictive of progression to smoking within a longitudinal study of youth behaviors.

In 2010, both variables `stu` and `d72` are available. Use `d72` to match the results used in the frequency reports. If you are using an older multi-year dataset, the `stu` variable is probably included (not `d72`).

You can calculate susceptibility, `d72`, by coding:

```
* Susceptibility to smoking uptake - All respondents
gen d72=.
replace d72=1 if (d29==1 & d30==1)
replace d72=2 if (d29==2 | d29==3 | d29==4 | d30==2 | d30==3 | d30==4)
lab def d72 1 "yes, not susceptible" 2 "no, susceptible"
lab val d72 d72
lab var d72 "Made firm commitment to not smoke cigarettes"
```

Susceptibility is often only calculated for those youth who are not currently smoking¹ and calculated by coding:

```
* Susceptibility to smoking uptake - Among NON-Smokers
gen nonsmoker=d14use
recode nonsmoker 1=0 2=1
svy:tab d72 grade, subpop(nonsmoker) col se obs
```

You can calculate the `stu` variable for susceptibility, from the older multi-year dataset (2008 and before), by coding:

```
* Susceptibility to smoking uptake - All respondents
gen stu=.
replace stu=0 if (d29==1 & d30==1)
replace stu=1 if (d29==2 | d29==3 | d29==4 | d30==2 | d30==3 | d30==4)
lab def stu 0 "not susceptible" 1 "susceptible"
```

Any Tobacco Use

In 2010, youth were asked about their past 30 day use of a number of tobacco products: cigarettes, smokeless, cigars, and candy-flavored tobacco. Most youth who use tobacco are multi-product users. It is possible to combine all types of tobacco for a single "any tobacco use" variable, but its usefulness can be limited by the number of respondents.

The cigarette and smokeless tobacco questions are core items on forms A, B/NS and C. The cigar and candy-flavored tobacco questions are only on form B/NS. The calculated

¹ Source: Pierce JP, Gilpin EA, Farkas AJ, Merritt RK. "Validation of susceptibility as a predictor of which adolescents take up smoking in the United States" *Health Psychology* 1996;15(5):355-361

variable should be restricted to only include respondents who took survey form B and who answered all of the 30 day tobacco questions.

If wanting to create an “any tobacco use” variable to compare with 2008, candy-flavored tobacco d84 should be left out (not asked in 2008).

```
*Any tobacco use(cigarette, smokeless, and cigars)
gen anytob=.
replace anytob=1 if(d14use==1|d15use==1|d16use==1)
replace anytob=0 if(d14use==2&d15use==2&d16use==2)
replace anytob=. if(d14use==.|d15use==.|d16use==.)
replace anytob=. if grade==6 | (formtype~="B" | formtype~="NS")
lab def anyuse 1"used any" 0"no use"
lab val anytob anyuse
```

Other Calculated/Computed Variables

There are a number of computed variables in the HYS; some of these were not provided for earlier years of the survey. We are providing the computations so that you can create these variables for datasets where they do not exist, or simply so that you understand where the computed variables come from.

Asthma – recode for “current asthma” in 2010

In 2010, there were two primary variables used to describe asthma prevalence: ever being told by a doctor or nurse you have asthma (lifetime asthma) h22, and do you still have asthma h86. This matches the national Youth Risk Behavioral Survey questions to calculate current asthma. Prior to 2008, HYS used different questions so a comparison of current asthma over time is not available.

For more discussion on this topic, refer to “The Burden of Asthma in Washington State 2005”, pages 33-35, available at: http://www.doh.wa.gov/portals/1/Documents/Pubs/345-201_TheBurdenofAsthmaInWashingtonState.pdf

```
* Lifetime asthma
gen lifeasthma=h22
recode lifeasthma 1=1 2= 3=
lab def yesno 1 “yes” 2 “no”
lab val lifeasthma yesno
```

```
* Current asthma
gen currentasthma=.
replace currentasthma=1 if (h22==1 & h86==2)
replace currentasthma=2 if (h22==2 | h22==3 | h86==1 | h86==3 | h86==4)
lab val currentasthma yesno
lab var currentasthma "Current asthma, diagnosed by a dr. and still have-NEW in 2008"
```

Physical Activity – Recode for Meeting Physical Activity Recommendations

The CDC has recently provided new and more complicated guidelines for physical activity recommendations for youth. Prior to 2005, however, the broad physical activity recommendation has been to achieve at least 30 minutes of moderate activity five or more times per week, or 20 minutes of vigorous activity three or more times per week. In 2005, the recommendation was changed so that youth are encouraged to participate in at least 60 minutes of moderate intensity physical activity most days of the week, preferably daily.

More information about physical activity recommendations is available at <http://www.cdc.gov/nccdphp/dnpa/physical/recommendations/index.htm>

In 2006 a new question (h63) was added to HYS to measure the new USDA recommendation to “Engage in at least 60 minutes of physical activity on most, preferable all, says of the week.” This measure was met if youth reported they were physically active for at least 60 minutes a day at least 5 days a week.

```
*physical activity standard
tab h63, missing
gen pa_60days=.
recode pa60_5days 1=0 2=0 3=0 4=0 5=0 6=1 7=1 8=1
lab def pa5days 1"5+ days" 0"< 5 days"
lab val pa60_5days pa5days
```

In 2010, the older questions about vigorous and moderate activity were not asked.

Socioeconomic Status (SES)

Socioeconomic status - a measure of an individual or family’s relative economic and social ranking - is an important social determinant of health; however, we recognize that youth are not accurately able to report on family income. Maternal education (the level of education that has been completed by the student’s mother) is a proxy measure for family SES that has been described in the literature. Thus we can use this as a “best guess” for the student’s SES. Maternal education can be stratified in a variety of ways; we recommend stratifying as “lower SES” if a mother has no post-high school education and “moderate - higher SES” if a mother has had any post-high school education.

In 2006, the response options for parent’s level of education were changed to more closely match the question for the youth. “How far in school do you think you will go?”. This created a new variable for mother’s level of education (g17). It is important to note that while the change in response options resulted in a decrease in the proportion of youth who reported that they do not know what level of education their mother has had, the percent unknown or missing is still high for this question (20% overall in 2008). Therefore it is important to use this measure cautiously and be aware of the impact that a large proportion of missing data may impact comparisons to previous years (variable g10).

For more information on the parental education variables, see the section on Demographic Variables.

```
gen lowsese=g17
recode lowsese 1=1 2=1 3=0 4=0 5=0 6=. 7=.
lab def lowsese 1"low lowsese" 0"higher"
lab val lowsese lowsese
```

“Obesity” from Body Mass Index h01

Obesity is calculated using BMI based on students’ self-reported height and weight. Height is converted to centimeters and weight to kilograms, then BMI is computed using the standard formula:

$$\text{BMI}=(\text{weight in kilograms})/(\text{height in centimeters squared})$$

The cutpoints for obesity and overweight are based on age and gender specific growth charts developed by the CDC. Individuals in the top 5 percent for BMI based on age- and gender-specific growth charts are considered obese. Those in the top 15 percent, but not the top 5 percent, are considered overweight.

Quality of Life Scale

The Youth Quality of Life Instrument-Surveillance Version (YQOL-S) is a 13-item questionnaire designed to assess quality of life among adolescents. The instrument contains five contextual items that are potentially verifiable, and eight perceptual items that are things known only to the adolescent. Poorer quality of life as measured by the YQOL-S has been shown to be associated with increased health-risk behaviors among adolescents. The scale was developed by the Seattle Quality of Life Group and colleagues at the University of Washington.

The 2010 Healthy Youth Survey included six of the eight perceptual items. The two items not included were (I enjoy life and I feel life is worthwhile). In preliminary analyses by the Seattle Quality of Life Group, this subset of questions was collapsed into a quality of life scale was also associated with health-risk.

Note: There is already a pre-calculated summary variable for YQOL in the HYS dataset (l13) which categorized youth quality of life into: low, medium low, medium high and high. More information on how this is done is listed below.

The subset of the YQOL-S included on Form B/NS of the 2010 Healthy Youth Survey follows. The following are some statements that you might make about yourself. With 0 being “not at all true,” and 10 being “completely true,” please fill in the number on the scale that best describes how closely the statement applies to you.

- I feel I am getting along with my parents or guardians. (L02)
- I look forward to the future. (L03)
- I feel good about myself. (L04)
- I am satisfied with the way my life is now. (L05)
- I feel alone in my life. (L06)

The final statement (L07), had a slightly different format, with 0 being “much worse than others” and 10 being “much better than others.”

- Compared with others my age, my life is...

These items are available as single items. They also have been collapsed into an YQOL scale. To create this scale, the Washington State Department of Health followed the methods of the Seattle Quality of Life Group to create a continuous variable on a scale of one to 100 as follows:

- Recoded L06 (I feel alone in life) so that a higher score reflects feeling less alone in life.
- Transformed all items to t-scores so that they are on a scale of 100 (see formula below).
- Obtained the mean of the items as long as at least five items were answered.
- If fewer than five items were answered, the items must be used individually.

$$tscore = \frac{actual\ raw\ score - lowest\ possible\ raw\ score}{possible\ raw\ score\ range} * 100$$

There is a computed variable in the HYS dataset (yqols) that calculates a youth quality of life scale from 0 to 100. The Washington State Department of Health then created a categorical scale from the continuous scale scores in yqols and based on the HYS 2002 results, to approximate quartiles for each grade. Because many students were coded to the same continuous scale score, it was not possible to create exact quartiles. The coding for the recoded variable, called l13, follows. 1 could be considered low, 2 medium low, 3 medium high and 4 high. This item is most useful in combination with other items to examine the relationship between quality of life and health-related behaviors.

Healthy Youth Survey 2002

12th Grade, 1851 responses				10th Grade, 2207 responses				8th Grade, 3092 responses			
YQOLS	N	%		YQOLS	N	%		YQOLS	N	%	
1	0 to <60	[401]	21.7%	1	0 to <60	[538]	24.4%	1	0 to <60	[692]	22.4%
2	60 to <80	[549]	29.7%	2	60 to <80	[668]	30.3%	2	60 to <80	[829]	26.8%
3	80 to <90	[400]	21.6%	3	80 to <90	[440]	19.9%	3	80 to <90	[605]	19.6%
4	90+	[501]	27.1%	4	90+	[561]	25.4%	4	90+	[966]	29.9%

Links:

For background on the YQOL questions and answers to frequently asked questions, see the Seattle Quality of Life Group website: <http://depts.washington.edu/yqol>

Homeless Screener

In 2008, a question was added to try to identify homeless youth, based on the definition used in the McKinney-Vento act, a complicated legal definition. This screener was changed for HYS 2010, but can be used in 2008 as a surrogate measure for homeless youth.

```
*Homeless screener_2010
gen homeless= f31_10
recode homeless 1=0 2/8=1
lab def homeless 0 "Not homeless" 1 "Homeless"
lab val homeless homeless
lab var homeless "Homeless screener-20108"
```

Sexual Behavior

Age of first sexual intercourse, the number of sexual intercourse partners and condom use during last sexual intercourse are often only reported for those who have had sexual intercourse. These questions on sexual behavior were first asked in 2010, and appear only on the Form B tear-off.

```
* Among those who ever had sex, age of first sexual intercourse
gen agesex=h97
replace agesex=. if h97==1
lab def agesex 2 "11 or younger" 3 "12 years old" 4 "13" 5 "14" 6 "15" 7 "16" 8
"17 or older"
lab val agesex agesex
```

```
* Among those who ever had sex, number of sexual intercourse partners
gen sexpartners=h98
replace sexpartners=. if h98==1
lab def sexpartners 2 "1 person" 3 "2 people" 4 "3 people" 5 "4 people" 6 "5
people" 7 "6 or more people"
```

```
* Among those who ever had sex, those who used a condom last time they had
sexual intercourse
gen condom=h99
replace condom=. if h99==1
lab def yesno 1 "yes" 2 "no"
```


Risk and Protective Factors

Risk factors are characteristics of individuals, families, and communities that make us more vulnerable to ill health. Protective factors are characteristics that "protect" and thus significantly reduce the likelihood of disease, injury, or disability. Health-related risk and protective factors are commonly grouped into three general categories including lifestyle and behavior; environmental exposure, encompassing both the physical and social environments; and biologic and genetic characteristics. Risk and protective factors are often measured as different ends of the same continuum. For example, wearing seatbelts protects against motor vehicle-related injury and death; not using a seatbelt increases risk for these outcomes.

The risk and protective factors in the Healthy Youth Survey focus on lifestyle and behaviors and the social environment. The social environment includes the school, peer, community and home environments, as well as individual assets. The survey includes some factors directly related to health, but most of the risk and protective factors are associated with intermediary behaviors, such as drug and tobacco use, violence and staying in school. Many of these factors have been compiled into scales following the research of Hawkins and Catalano at the Social Development Research Group (SDRG), University of Washington.

The Hawkins and Catalano theoretical framework of risk and protective factors includes twenty-five factors, the scales for which are part of a survey called Communities That Care (CTC). The presence of multiple risk factors predicts an increased likelihood that an individual will engage in substance use, while the presence of protective factors helps to buffer the effect of risk factors and increase resilience.

For a detailed summary of the history of Risk and Protective Factors Scales used in the HYS see: <http://www.hys.wa.gov/Reporting/Default.aspx>

Content Changes Over Time

Several Healthy Youth Survey questions have changed over time. A crosswalk of survey questions from 2002, 2004, 2006, 2008 and 2010 highlights key changes that have occurred. HYS Question Crosswalk available on <http://www.AskHYS.net> in the QxQ data analysis section.

4

Getting to Know STATA

This section includes a table that provides a brief overview of some useful STATA commands.

For more information on the specific commands and the output they generate see Data Analysis sections 4 and 5 Or in STATA type help and the command , or use the Help drop down on your STATA tool bar and select STATA command.

Command	Example	Results
<i>For retrieving and saving data</i>		
use	use "C:\My Documents\2008 HYS.dta"	Opens the STATA file
save	save "C:\My Documents\new 08 HYS data.dta"	Saves a modified STATA data file
keep	keep d14 d36 grade g05, or keep if conum==1	Keeps only specific variables, or specified response options
drop	drop d14, or drop if conum==2	Drops specific variables, or Specified response options
<i>For variable exploration</i>		
codebook	codebook c01	Describes the variable c01. Includes the question, the data type (numeric or string), the number of values, the number of missing, and the response options and labels.
summarize	summarize c01	The number of observations, the mean, the standard deviation, the minimum value and the max value
summarize, detail	summarize c01, detail	Also includes the percentiles, variance, skewness and kurtosis
histogram	histogram c01	Plots a histogram of the variable responses

Command	Example	Results
Creating and transforming variables		
gen	gen year==2004 gen bully=c01	Creates a new variable, or Creates a new variable based on an original variable
recode	recode bully 1=0 2=1 3=1 4=1 5=1	Recodes the variable response options, in this example recodes the response options to be not bullied vs. bullied
replace	gen bully=. replace bully==0 if c01==1 replace bully==1 if (c01==2 c01==3 c01==4 c01==5)	In this example the gen command creates a new variable and the replace commands describe the new variable response options. Replace can also be used to create more complex recodes that combine more than one original variable
For labeling variables		
lab var	lab var bully "bullied, none vs. any"	Labels the variable with a description of the variable
lab def	lab def noneany 0"none" 1"any"	Creates new response option labels that can be applied to a variable
lab val	lab val bully noneany	Applies the response option label
Setup commands for analysis		
svyset	gen fakewt==1 svyset [pweight=fakewt]	Creates a new variable with a weight of 1 Designates the weighting. In this example the newly created fakewt variable is used, so the weight for all responses is equal to 1. Use for analysis of a census county.
	svyset [pweight=fakewt], psu(schgrd) or (schgnoid)	Sets the weight as 1 and the primary sampling unit as the school building/grade. Use for analysis of the state sample or analysis of a county with a county sample.
Updating STATA		
update all		Install official updates to STATA and provides new programs or commands. Use this often ☺

Command	Example	Results
For computing frequencies		
tab	tab c01 grade	Runs a crosstab of the two variables. Tab does not calculate percentages but just provides the number of observations for each cross
svy:tab	svy:tab c01 grade, col se ci obs	Can be used once the data is set up with the svyset command. Svy:tab runs crosstabs of two variables and provides a percentage by row or column and can include additional information such as the standard error (se), 95% confidence intervals (ci) and the number of observations (obs) if designated
For adding additional datasets		
merge	merge (schgrd) using "C:\My Documents\2004 school demo.dta"	Adds additional data to the respondents. In this example we are adding school building information based on the schgrd, possibly school type or enrollment, or free and reduced lunch rates. Remember if you have a de-identified dataset you will have to use the variable schgnoid.
append	append using "C:\My Documents\2002 HYS data.dta"	Adds additional respondents. In this example we are adding an additional year of data.
A few more useful commands		
if	svy:tab h01 grade if g05==1	Limits the analysis to females. "If" at the end of a command means the command is to use only the data specified. When doing CI, use "if" with caution as it can affect CI. Subpop is preferable.
&	keep if (conum==1 & grade==6)	And
	keep if (conum==1 conum==2)	Or
~=	keep if conum~=1	Does not equal. In this example all counties would be kept, except conum 1 would be dropped
*		Use in "do files" for notes, * before any statement will not run in STATA
///		Use in "do files" if statements are too long to fit on a page. /// at the end of a statement will make it continue to the next line

5

HYS Data Analysis in STATA

This section describes how to set up STATA for different types of data, how to explore your data, transform it and run some simple analyses.

If you are using the state sample data, you should be able to reproduce the outputs in this section. This section is formatted so that STATA commands are highlighted in grey and STATA outputs are highlighted in black boxes

This section covers the following topics:

- Opening your dataset
- Do files
- General set up for survey analysis – state, county, ESD, district and building
- Analysis by Grade
- Frequencies and summaries of statistics
- Creating new variables
- Labeling new variables
- Dichotomizing variables
- Two-way tables and crosstabs
- More options for using “svy”
- Additional tips for formatting
- Stratified analysis and subpopulations

For a table of commonly used STATA commands see the previous section [Getting to Know STATA](#) . For short examples of STATA coding see the [Data Analysis – Quick Examples](#).

Results presented in this section are from the 2010 HYS data.

Opening your Dataset

To open your HYS dataset, you need to tell STATA how much memory you want to devote based on the size of your dataset. This varies somewhat from computer to computer, but in general if you are using the state sample dataset or a smaller dataset, then devoting 100 megabytes (100m) should be fine. If you are using the complete dataset, you'll need 200 megabytes (200m) or more. Always start small and increase the memory as necessary – STATA will tell you if you need more.


You can also set the memory permanently: `set mem 200m, perm`

After your memory is set, open your file by typing “use” and then the file pathway in quotes (see syntax below). Or use the STATA drop down menus by selecting File – Open – then find the dataset you want to open and double click on it:


```
clear
set mem 200m
*use "C:hys10 final state dataset.dta"
    *insert the file path to your state sample data
```


Sometimes using data saved on a compact disc causes STATA to run slowly. To speed things up, save the data file on your hard drive and run it from there.

“Do Files”

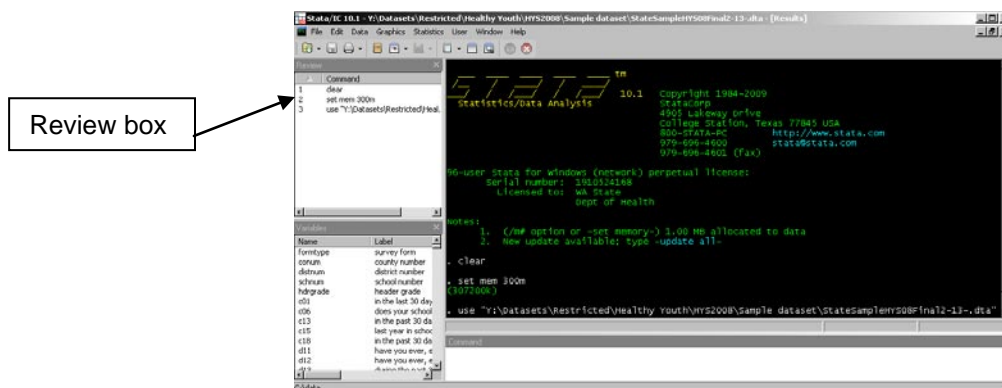
To open a “do file” click the “New do-file Editor” icon:  on the tool bar, select Do File Editor from the Window drop down menu, or hit Cntrl 8.

Once you have a blank do file open, you can begin writing your commands or open an existing do file by selecting Open from the do file - File drop down menu. “Do files” are handy because you can keep a record of your analysis. They also make it easy to change commands and rerun analysis.

To run individual lines or sections of commands in your “do file”, just highlight them and hit the icon  that looks like a page with text with an arrow.

To run the complete do file hit the icon that looks like a blank page with an arrow .

Or you can right click, select all and copy commands you have typed into the review box in STATA (usually in the upper left) and paste them into a do file.



Throughout this manual: STATA commands are in grey and STATA output is in black

General Set Up for Survey Analysis

Prior to survey analysis you must provide STATA with set up commands to account for weighting, primary sampling units and strata.

The set up options you use will depend on the type of data you are using and which type of analysis you are conducting. Below are some examples of types of analysis that would influence set up options:

- State sample analysis
- County sample analysis
- County census analysis
- County “mixed sampling” analysis
- ESD analysis
- District analysis
- Building analysis

State Sample Analysis

The state sample was drawn by simple random sample, so there is no weighting or strata required. For survey analysis STATA requires a weight, so you will need to create a fake weight (fakewt) that is equal to 1. The state sample was drawn at the school building level, so the primary sampling unit is the school building (schgrd), or schgnoid if you the dataset you have has the school buildings de-identified.

Set up command example:

```
gen fakewt=1
svyset [pweight=fakewt], psu(schgrd)
keep if staterec==1
```

County Analysis

County results should *only* be reported for grades that had at least a 40% participation rate and participation from two or more districts (with at least 15 valid responses per district). The easiest way to determine what you can report is to look on AskHYS.net. If a county has a grade-level report posted, then you can report those results.

Information from counties without enough districts participating can only be reported with written permission from the school district superintendent. In 2010, reports were produced for counties with only one district (Columbia, Garfield, and Wahkiakum) because the districts gave DOH written permission. Franklin County did not receive results because they have multiple districts, but only one participated. Their participating district gave DOH permission to combine their results with Benton County, so there are combined Benton-Franklin results.

NOTE: The county participation rates vary by year, for more information on which counties received reports, go to www.AskHYS.net and select “HYS Results -

Frequency Reports”. For combined county reports, like Benton-Franklin, see the reports under the “State” dropdown menu.

To exclude specific counties or grades, use the “drop” command. For 2010, the following counties and grades should be dropped:

```
drop if conum==5 & (grade==6 | grade==10 | grade==12)
drop if conum==11
drop if conum==16 & grade==12
drop if conum==30 & (grade==6 | grade==8)
drop if conum==33 & (grade==6 | grade==10 | grade==12)
```

County sample analysis

Random county samples were drawn for counties with more than 30 schools in a grade. The following table describes the county samples from 2002, 2004, 2006, 2008 and 2010.

Sampled Counties by Year	2002	2004	2006	2008	2010
Clark	Grades 6 and 8	-	-	Grades 6 and 8	Grades 6 and 8
King	All grades	All grades	All grades	All grades	All grades
Kitsap	Grade 6	Grade 6	Grade 6	-	-
Pierce	All grades	All grades	All grades	All grades	All grades
Snohomish	All grades	All grades	All grades	All grades	All grades
Spokane	All grades	Grade 6	Grades 6 and 8	Grade 6	Grades 6 and 8
Thurston	Grade 6	-	-	Grade 6	Grade 6

In 2010, county samples were drawn for all grades in King, Pierce, and Snohomish counties. To analyze data from these one of these counties, use the same set up as the state sample.

Set up command example:

```
keep if conum==17
    *i.e., conum==17 is King County
keep if corec==1
gen fakewt=1
svyset [pweight=fakewt], psu(schgrd)
```

County census analysis

For almost all other counties, all schools in the county are included (a census), so the primary sampling unit is the individual student. You do not need to set a psu.

Set up command example:

```
gen fakewt=1
keep if conum==1
    *i.e., conum==1 is Adams County
keep if corec==1
svyset [pweight=fakewt]
keep if corec==1
```


County with “mixed sampling” analysis

In 2010, three counties were a mix of sampling and census. Samples were only drawn for Clark 6th and 8th grade, Spokane 6th and 8th grade, and Thurston 6th grade. The rest of the grades were census. County sampling changes from year to year, for previous years see the table of Sampled Counties by Year.

This scenario deserves special attention because it is dependent on the grades being analyzed. If you are just analyzing the 6th and/or 8th grade in Clark or Spokane (or only 6th grade in Thurston), then use the set up for county sample analysis. If you are trying to look at other grades in the county, you need to create a new variable for your primary sampling unit. The new variable needs to simultaneously take into account 1) the primary sampling unit for grade six as the school building and 2) the primary sampling unit for the other grades as the individual student.

Set up command for Spokane example:

```
keep if conum==32
keep if corec==1
gen fakewt=1
gen id = _n
gen psu=id +10000
replace psu=schgrd if (grade==6 | grade==8)
svyset [pweight=fakewt], psu(psu)
```

All or multiple county analysis

The following commands can be used if you are running analysis on all counties, some sampled and some census. You need to have a complete state census data set to run all counties.

You will also need to create a new primary sampling unit variable that takes into account the different sampling schemes, school building for counties and grades with samples and individual students for census counties.

For 2010, the following code is needed to create a psu to account for county sampling and set up data for analyzing data from multiple counties:

```
keep if corec==1
gen fakewt=1
gen id=_n
gen psu=id +10000
replace psu=schgrd if conum==6 & (grade==6 | grade==8)
replace psu=schgrd if conum==17
replace psu=schgrd if conum==27
replace psu=schgrd if conum==31
replace psu=schgrd if conum==32 & (grade==6 | grade==8)
replace psu=schgrd if conum==34 & grade==6
svyset [pweight=fakewt], psu(psu)
```

The command “gen id=_n” creates a unique identifier for each respondent. When we create our new “psu” variable we add 10,000 to the “id” variable to make sure the new “psu” variable is also unique. Then we replace the individual “id” with the school identifier “schgrd” (or schgnoid in some datasets) in the counties that were sampled.

ESD Analysis

ESDs are made up of counties or sections of counties. Some ESDs are made up of counties with samples and some with census. To account for the different sampling schemes, a weight needs to be used that takes the enrollment of schools in the sampled counties. The different sampling schemes also affect the primary sampling units, so a new primary sampling unit variable needs to be created. Also because county is another layer of sampling, it needs to be accounted for by being designated as strata.

As with counties, you should make sure that the ESD had a 40% response rate and enough school participate to receive results. In 2010, all ESD grade levels received results, so none of them need to be dropped.

ESD without sampled counties analysis

In 2010, ESDs 105, 114, 123 and 171 did not have any sampled counties, so no special weight or psu needs to be applied.

Set up command example for ESD with no sampled counties:

```
keep if esdum==105
keep if esdrec==1
gen fakewt=1
svyset [pweight=fakewt], strata(conum)
```

ESD with sampled counties analysis

In 2010, ESD 101, 112, 113, 121 and 189 had some counties with samples. To analyze data from these one of these ESDs we need to apply weighting and a psu that takes into account the different county sampling.

Set up command examples for ESDs with some sampled counties:

```
keep if esdum==101
keep if esdrec==1
gen id=_n
gen esdpsu=id +10000
replace esdpsu=schgrd if conum==32 & (grade==6 | grade==8)
svyset [pweight=esdwt], psu(esdpsu) strata(conum)
```

```
keep if esdum==121
keep if esdrec==1
gen id=_n
gen esdpsu=id +10000
replace esdpsu=schgrd if conum==17
replace esdpsu=schgrd if conum==27
svyset [pweight=esdwt], psu(esdpsu) strata(conum)
```

All or multiple ESD analysis

Similar to the counties, the following code is needed to create a psu to account for county sampling and set up data for analyzing data from multiple counties in 2010:

```
keep if esdrec==1
gen id = _n
gen esdpsu=id + 10000
replace esdpsu=schgrd if conum==6 & (grade==6 | grade==8)
replace esdpsu=schgrd if conum==17
replace esdpsu=schgrd if conum==27
replace esdpsu=schgrd if conum==31
replace esdpsu=schgrd if conum==32 & (grade==6 | grade==8)
replace esdpsu=schgrd if conum==34 & grade==6
svyset [pweight=esdwt], psu(esdpsu) strata(conum)
```

District and Building Analysis

For district analysis all school buildings are to be included because all buildings were eligible to participate, so the primary sampling unit is the student. The variable distnum is not a unique number, i.e., more than one district have the distnum 100. District numbers are only unique within counties, so for district analysis always use the codis variable (a number that includes the county number and the district number).

Set up command example for district:

```
keep if codis==15204
    *i.e., codis=15204 is Coupeville School District in Island County
keep if distrec==1
gen fakewt=1
svyset [pweight=fakewt]
```

For building analysis all students were eligible, so students are the primary sampling unit.

Set up command example for building:

```
keep if schnum==4460
    *i.e., schnum=4460 is Beaver Lake Middle School in Issaquah
gen fakewt=1
svyset [pweight=fakewt]
```

Analysis by Grade

The variable for a student's grade level is "grade". Do not use the variable "hdrgrade". See the Getting to Know Your Data - grade and hdrgrade.

We recommend that all analyses be done stratified by grade, due to the sampling procedure by grade and since responses are often variable according to the student grade level and that you use the "svy" option in STATA. "Svy" is a prefix used with STATA commands when you are analyzing survey data. "Svy" takes your weights, psu, strata, etc. into account when you are running estimation commands.

NOTE: There may be some exceptions to this, see the Combining Grade Levels section.

You may choose to look at grade differently depending on the types of analysis you are doing and the variables you are looking at. Some variables such as the ones that measure substance abuse vary greatly by grade level, others such as the prevalence of asthma are more stable across grade levels.

To simply look at the results for one variable by grade do with "svy:tab":

```
svy:tab d14use grade, col obs per
```

30-day use: cigarettes		grade				
	6	8	10	12	Total	
Use	1.729 193	6.572 624	12.66 860	19.59 1145	8.476 2822	
No use	98.27 1.1e+04	93.43 8871	87.34 5933	80.41 4700	91.52 3.0e+04	

Interpretation: 2% of 6th graders, 7% of 8th graders, 13% of 10th graders and 20% of 12th graders smoked cigarettes in the past 30 days in 2010.

Frequencies and Summaries of Statistics

Even before you use your set up commands you can run basic frequencies using the “tab” command.

Example in STATA using variable d14, 30 day current cigarette use:

```
tab d14
```

```
during the past 30 days, on how many days did you: smoke
cigarettes?
```

	Freq.	Percent	Cum.
None	30,472	91.52	91.52
1-2 days	1,013	3.04	94.57
3-5 days	423	1.27	95.84
6-9 days	307	0.92	96.76
10-29 days	428	1.29	98.04
All 30 days	651	1.96	100.00

For initial variable exploration, you can use the summarize command to find out the number of observations, mean, standard deviation, min and max type:

```
summarize d14
```

Variable	Obs	Mean	Std. Dev.	Min	Max
d14	33294	1.232685	.8970091	1	6

For more information including the percentile breakdowns, variance, skewness and kurtosis:

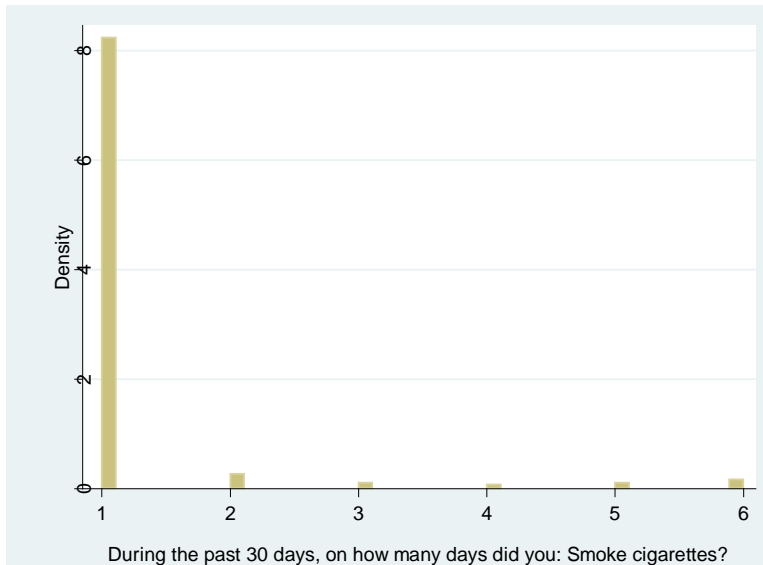
```
summarize d14, detail
```

```
during the past 30 days, on how many days did you: smoke
cigarettes?
```

		Percentiles		Smallest	
1%	1	1			
5%	1	1			
10%	1	1	Obs	33294	
25%	1	1	Sum of Wgt.	33294	
50%	1		Mean	1.232685	
			Largest	Std. Dev.	.8970091
75%	1	6			
90%	1	6	Variance	.8046254	
95%	3	6	Skewness	4.258751	
99%	6	6	Kurtosis	20.64794	

Using histograms can also be helpful in getting a quick view of the distribution:

`histogram d14`



You can also explore your variables by demographics such as grade to find out the number of observations for each category. Example, current smoking use by grade:

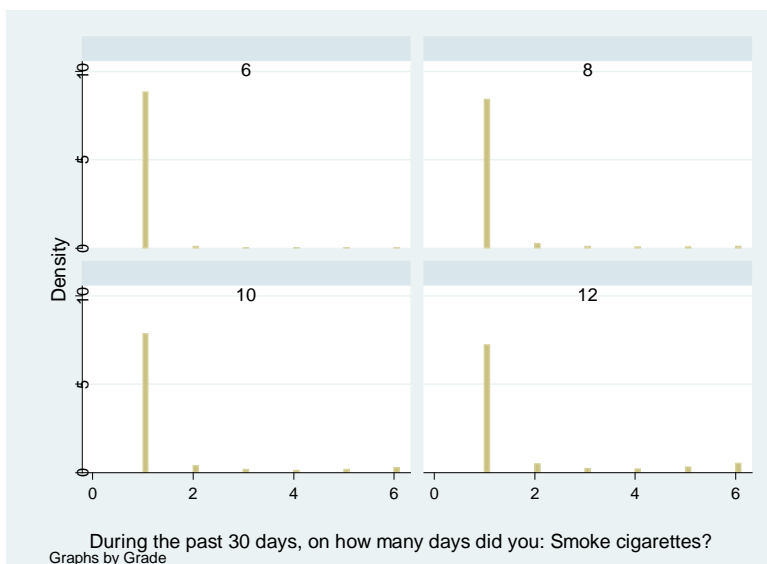
`tab d14use grade`

30-day use: cigarettes		grade				
	6	8	10	12	Total	
Use	193	624	860	1,145	2,822	
No use	10,968	8,871	5,933	4,700	30,472	

Notice that the proportion of 12th graders who smoked cigarettes in the past 30 days is much higher than 6th graders.

You can also get a visual look with a histogram:

`histogram d14, by(grade)`



Throughout this manual: STATA commands are in grey and STATA output is in black

Creating New Variables

There are many ways to create new variables in STATA - below are a few commands you can use.

Generating

The command for creating a new variable is “gen”. Below are a few examples of how you can use the “gen” command:

- `gen cig30=d14` ~ creates a new variable that is the same as the original variable
- `gen cigchew30 = d14use + d15use` ~ creates a variable that adds the responses from one variable to another for each respondent
- `gen new=.` ~creates a variable with all missing values
- `tab grade, gen(gradecat)` ~ creates a new dummy variable for each of the original variable response options - with “gradecat” as the prefix followed by the numbers “1,2,3, etc.” depending on the number of response options. In this case “gradecat1”, “gradecat2”, etc.

NOTE: For more information on “gen”, type the command “help gen” in STATA

Recoding

Often during analysis you want to collapse or drop response options. The simplest way to do this is to create a new variable using the “gen” command and reorder the response options using the “recode” command. It is usually a good idea to create a new variable before recoding because you may want to go back and use the original response options sometime during your analysis or recode the variable in a different way.

Before recoding, look at the numerical values assigned to each response option using the “codebook” command:

```
codebook d14
```

```
d14  During the past 30 days, on how many days did you: smoke
      cigarettes?

      type:  numeric      (byte)
      label:  LABC

      range:      [1,6]      units:      1
      unique values:  6      missing .:  775/34069

      tabulation:  Freq.  Numeric      Label
                   30472  1      None
                   1013  2      1-2 days
                   423   3      3-5 days
                   307   4      6-9 days
                   428   5      10-29 days
                   651   6      All 30 days
                   775   .
```

Now you know that the variable has 6 response options. If you wanted to recode the 30 day smoking response options into none or any, you need to change 1 the “none” response to 0 and all of the other responses to 1 “any”. After recoding your new variable, run a “tab” to make sure your new response options are the way you want them.

```
gen cig30 = d14
recode cig30 1=0 2=1 3=1 4=1 5=1 6=1
tab cig30 grade
```

cig30	grade				Total
	6	8	10	12	
0	10,968	8,871	5,933	4,700	30,472
1	193	624	860	1,145	2,822

You can also recode the above variable like this:

```
recode cigthirty 1=0 2/6=1
```

After recoding it is always a good idea to check your new results to make sure they make sense when compared to your pre-collapsed variable results. In this case you can check your recode by using the pre-collapsed variable d14use.

```
tab d14use cig30
```

NOTE: For more information on “recode”, type the command “help recode” in STATA

Replacing

To combine more than one variable and to do more complex recoding, you can use the “replace” command. For example, to calculate if someone has either seen a doctor or a dentist in the past 24 months, you need to combine two different variables, h24 visiting a doctor and h25 visiting a dentist.

Before starting to replace, it’s always a good idea to run the codebook command on any variables that you will be using to make sure you know which numeric value is given to each response option.

```
codebook h24 h25
```

```
h24  when was the last time you saw a doctor or health care
      provider for a check-up or exam...

      tabulation: Freq. Numeric  Label
                  6578  1      During the past 12 months
                  1839  2      Between 12 and 24 months ago
                   653  3      More than 24 months ago
                   384  4      Never
                  1429  5      Not sure
                  23186  .
```



```
h25    when was the last time you saw a dentist for a check-up,  
       exam, teeth cleaning...
```

```
tabulation: Freq. Numeric Label  
            8239  1      During the past 12 months  
            1088  2      Between 12 and 24 months ago  
             610  3      More than 24 months ago  
             173  4      Never  
             771  5      Not sure  
            23188 .
```

If you wanted to determine who visited both a doctor and a dentist, you can create a new variable “anyvisit” with all values designated as missing. To do this type “gen anyvisit=.” This ensures that you will only add in the respondents you want.

```
gen visitboth=.
```

For those who visited both a doctor and a dentist in the past 12 months, we want respondents that answered “during the past 12 months” for both of the questions. The following symbols are needed to tell STATA what to do:

Use “=” to assign the numeric value to the response option for our new variable

Use “==” to designate which variable response options you are using

Use “&” to symbolize the word “and”

Below is an example of how you would use the symbols mentioned above to tell STATA the conditions for designating non-exposed as zero:

```
replace visitboth=1 if (h24==1 & h25==1)
```

For those who didn’t visit either a doctor or a dentist in the past 12 months, we want respondents who answered “between 12 and 24 months ago” or “more than 24 months ago” or “never”. The following symbol is needed to tell STATA what to do

Use “|” to symbolize the word “or” (use shift and hit “\”)

Below is an example of how you would use this symbol above to tell STATA the multiple conditions for designating exposed as the value of one:

```
replace visitboth=0 if (h24==2 | h24==3 | h24==4 | h25==2 | h25==3 |  
h25==4)
```

To make sure we exclude respondents that didn’t answer both questions, we need to tell STATA to set them to missing:

```
replace visitboth=. if (h24==. & h25==.)  
tab visitboth grade
```

```
gen visitboth=.  
(34069 missing values generated)  
replace visitboth=1 if (h24==1 & h25==1)  
(5607 real changes made)  
replace visitboth=0 if (h24==2 | h24==3 | h24==4 | h25==2 |  
h25==3 | h25==4)  
(3846 real changes made)  
replace visitboth=. if (h24==. & h25==.)  
(0 real changes made)
```

```
tab visitboth grade
```

		grade			
visitboth		8	10	12	Total
	0	1,443	1,192	1,211	3,846
	1	2,433	1,747	1,427	5,607

Notice that there are no results for 6th grade because these questions were not asked of 6th graders.

Recoding can be tricky because it is not just one sided coding, you need to include exactly the respondents you want and drop the respondents you don't want.

Labeling New Variables

Once you have created a new variable or recoded response options, you may want to create labels for them. Use the following commands to create labels:

- "lab var" or "label variable" ~ adds a description to your variable
- "lab def" or "label define" ~ creates response option labels (once you create a response option label, you can reuse it over and over with other variables)
- "lab val" or "label value" ~ applies response option labels to your variable

```
lab var visitboth "visited both a doctor and a dentist in the past year"  
lab def visit 1"both" 0"one or none"  
lab val visitboth visit  
tab visitboth
```

```
lab var visitboth "visited both a doctor and a dentist in the past year"  
lab def visit 1"both" 0"one or none"  
lab val visitboth visit  
tab visitboth  
visited both a doctor and a dentist in the past year
```

	Freq.	Percent	Cum.
One or none	3,846	40.69	40.69
both	5,607	59.31	100.00

NOTE: For more information on labeling, type the command "help label" in STATA

General Rules on Creating Dichotomous Variables

When creating dichotomous (yes/no) variables, HYS uses the guidelines the CDC uses for its' Youth Risk Behavior Survey since many of the questions in HYS come from the YRBS. Thus we can compare ourselves to national data.

When calculating dichotomous variables , in general the numerator is the percent saying "Yes". The denominator is either all students or a subset of students who have indicated in the current survey they participate in a selected activity or behavior. Students must have provided valid data to be included in any dichotomous variable calculations. Therefore students with missing responses are not included. Some examples are included below.

Question: During the past 12 months, did you ever feel so sad or hopeless almost every day for two weeks or more in a row that you stopped doing some usual activities?

A. Yes / B. No

Summary text: Percentage of students who felt so sad or hopeless almost every day for two weeks or more in a row that they stopped doing some usual activities during the past 12 months

Numerator: Students who answered A

Denominator: Students who answered A or B

Question: Has a doctor or nurse ever told you that you have asthma?

A. Yes / B. No / C. Not sure

Summary text: Percentage of students who had ever been told by a doctor or nurse that they had asthma

Numerator: Students who answered A

Denominator: Students who answered A, B, or C

Question: When you rode a bicycle during the past 12 months, how often did you wear a helmet?

A. I did not ride a bicycle during the past 12 months

B. Never wore a helmet

C. Rarely wore a helmet

D. Sometimes wore a helmet

E. Most of the time wore a helmet

F. Always wore a helmet

Summary text: Among students who rode a bicycle during the past 12 months, the percentage who never or rarely wore a bicycle helmet

Numerator: Students who answered B or C

Denominator: Students who answered B, C, D, E, or F

NOTE: So if a question has responses Yes/ No/Not Sure, the Not Sures are included in the denominator. This is different from the BFSSS survey , a telephone survey of adults that allows the caller to keep probing for a Yes/No response.

Two-Way Tables or Crosstabs

“Svy” is a prefix used with STATA commands when you are analyzing survey data. “Svy” takes your weights, psus, strata, etc. into account when you are running estimation commands. “Svy:tab” is a tabulation command. It also provides you with a test of independence.

Example of crosstab using variables:

d28: Have you ever smoked cigarettes every day for 30 days? (yes/no)

g05: Are you: male or female?

```
svy:tab d05 g05
```

```
Have you ever had more than a sip or two of beer wine or hard
liquor?

```

	gender		
	female	male	Total
No	.2926	.2669	.5595
Yes	.219	.2215	.4405
Total	.5116	.4884	1

Key: cell proportions
Pearson:
Uncorrected chi2(1) = 21.2382
Design-based F(1, 211) = 12.8299 P = 0.0004

Interpretation: There are four cells in the two-way table. The results in the four cells add up to 100%:

female no (29%) + female yes (2%) + male no (27%) + male yes (22%) = 100%

The key below the total row of the table reminds you that the results are displayed as cell proportions.

Underneath the key, the output gives you the results of a Pearson correlation test. If the P (p value) is less than 0.05, then one of the cells is significantly different than the others at a 95% confidence level.

Additional Options with “Svy”

There are a number of additional options that can be added to a “svy:tab” to change the way the data is displayed or to provide you with more information. To use the additional options type a comma (,) after the variables.

Col and Row

“col”: gives you column percents. In this example, results are displayed for females no/yes in the first column and for males no/yes in the second column. **Each column adds up to 100%.**

```
svy:tab d05 g05, col
```

```
Have you ever had more than a sip or two of beer wine or hard
liquor?
      gender
      female   male   Total
No      .5719   .5465   .5595
Yes     .4281   .4535   .4405
Total   1       1       1
```

Interpretation: 43% of females and 45% of males have ever drunk alcohol in their lifetime.

“row”: gives you row percents. In this example, results are displayed for no female/male in the first row and yes female/male in the second row. **Each row adds up to 100%.**

```
svy:tab d05 g05, row
```

```
Have you ever had more than a sip or two of beer wine or hard
liquor?
      gender
      female   male   Total
No      .523    .477    1
Yes     .4972   .5028   1
Total   .5116   .4884   1
```

Interpretation: Of those students who ever drank alcohol in their lifetime 50% were female and 50% were male.

NOTE: Remember if “col” or “row” are not specified, the cells in the entire table add up to 100%

Obs

Adding “obs” at the end of the “svy:tab” command will give you the number of observations in each cell, each column, each row, and the total observations.

SE and CI

You can also add options at the end of “svy:tab” to give you the standard error (se) and 95% confidence intervals (ci):

Percentages

The “per” or “percent” command allows you display the point estimates as percentage points.

```
svy:tab d05 g05, col obs se ci per
```

```
Have you ever had more than a sip or two of beer wine or hard
liquor?

              gender
              female   male   Total
No           57.19     54.65   55.95
             (1.754)   (1.443) (1.57)
             [53.7,60.61] [51.8,57.48] [52.84,59.02]
             9535      8698    1.8e+04

Yes          42.81     45.35   44.05
             (1.754)   (1.443) (1.57)
             [39.39,46.3] [42.52,48.2] [40.98,47.16]
             7138      7217    1.4e+04
```

Interpretation:

Among females, who ever drank alcohol in their lifetime:

- 42.8% = point estimate
- $\pm 3.4\%$ = symmetric 95% confidence interval
(calculated by multiplying the standard error $1.754 * 1.96 = 3.4$)
- [39, 46] = non- symmetric 95 percent confidence interval
- 7138 respondents

Among males, who ever drank alcohol in their lifetime:

- 45.4% = point estimate
- $\pm 2.8\%$ = symmetric 95% confidence interval
(calculated by multiplying the standard error $1.443 * 1.96 = 2.8$)
- [43, 48] non- symmetric 95 percent confidence interval
- 7217 respondents

Additional Tips for Formatting Data

The following commands can be used to format your output into a more understandable and readable format:

Widening table columns

Use `stubwidth` and `cellwidth` to change the size of your columns so that all of the label text can be displayed:

```
svy:tab s01 g05, row ci stubwidth (20) cellwidth (15)
```

```
how often do you feel the schoolwork you are assigned is meaningful
and important?

```

	gender		Total
	female	male	
Almost always	.5504 [.536, .5647]	.4496 [.4353, .464]	1
Often	.5079 [.4944, .5214]	.4921 [.4786, .5056]	1
Sometimes	.5017 [.4906, .5129]	.4983 [.4871, .5094]	1
Seldom	.4547 [.4353, .4742]	.5453 [.5258, .5647]	1
Never	.3755 [.3469, .405]	.6245 [.595, .6531]	1

Rounding

The “`format`” command allows you to specify the display format for variables. When used as below, the number after the period allows you to indicate how many decimal points you want to show (thus 0 means to round the results).

```
svy:tab grade g05, per row ci format (%4.0f)
```

```


```

	gender		Total
grade	female	male	
6	50 [49,51]	50 [49,51]	100
8	50 [50,51]	50 [49,50]	100
10	52 [51,53]	48 [47,49]	100
12	50 [48,51]	50 [49,52]	100

If you need the exact number of observations use the `format` command to tell STATA how many numbers you would like it to display before and after the decimal point.

The rounding used to produce HYS results has changed over time. For consistency, it is recommended that analysis be produced to the 100th and rounded to the 10th or a whole number. E.g., use `format(%3.2f)` to produce the results 52.45% and report as 52.5% or 52%.

Removing Scientific Notation

Rounding can also be useful if you have large numbers of observations and your results come out in scientific notation.

```
svy:tab grade g05, row per obs format (%9.3f)
```

grade	gender		Total
	female	male	
6	50.321	49.679	100.000
	5793.000	5719.000	11512.000
8	50.490	49.510	100.000
	4893.000	4798.000	9691.000
10	52.335	47.665	100.000
	3598.000	3277.000	6875.000
12	49.593	50.407	100.000
	2926.000	2974.000	5900.000

In this example of the option “`format(%9.3f)`”, the 9 tells STATA to display up to 9 digits before the decimal point and the .3 tells it to display 3 digits after the decimal point. You can see how this effects both the point estimate (in the previous example when format was not specified, 4 digits were displayed after the decimal point) and how it effects the observations. Play with the numbers in the format command to get your ideal display.

Vertical Alignment

The “`vert`” or “`vertical`” command will display your confidence intervals (ci) on top of each other and without the bracket and comma. This can be useful if you are coping your results into Excel.

```
svy:tab grade g05, row ci per vert
```

grade	gender		Total
	female	male	
6	50.32	49.68	100
	49.37	48.73	
	51.27	50.63	
8	50.49	49.51	100
	49.56	48.58	
	51.42	50.44	

For more information on format, type the command “`help format`”, or see the Additional Tips for Formatting Data section in this manual.

Stratified Analysis and Subpopulations

Often you want to look at crosstab results among specific subpopulations, i.e. among certain grade level, races, etc. One simple way is to use “drop” or “keep” commands to limit your dataset to only the subgroup you are interested in. For example if you are only looking at results among 8th grade students:

```
keep if grade==8 will remove students from all of the other grades.  
drop if grade==8 will remove 8th grade students, but keep 6th, 10th and 12th graders.
```

NOTE: Make sure you do not save over your data file after using a keep or drop command. Doing so will overwrite your file and you will lose the records that were there previously.

If you are only looking at results among students who have smoked cigarettes in the past 30 days:

```
keep if d14use==1 will only keep the current smokers in your dataset.
```

Another option is to use the subpop command. Any binary variable that is coded as 0, 1 can be used as a subpopulation. Examples for making subpop variables:

creates a subpop of only current smokers

```
gen smoke=d14use  
recode smoke=1=1 2=0
```

creates a subpop of only Black-African Americans

```
gen black=g06  
recode black 1=0 2=0 3=1 4=0 5=0 6=0 7=0 8=0
```

creates a subpop of only 8th grade students (ok to use replace since there are no missing respondents in the grade variable, but check any the number of missing for any other variable as missing values will be coded 0)

```
tab grade, missing  
gen eight=1 if grade==8  
replace eight=0 if grade~=8
```

You can also create new combined variables for subpops, for example, creates a subpop of only 8th grade Black-African American students:

```
gen black8=g06  
recode black8 1=0 2=0 3=1 4=0 5=0 6=0 7=0 8=0  
replace black8=. if grade ~=8
```

The best way to create subpops is to create “dummy” variables. This command will generate a new variable for each response option:

```
tab grade, gen(gradecat)
```

Creates four new variables:

- gradecat1 (for 6th grade)
- gradecat2 (for 8th grade)
- gradecat3 (for 10th grade)
- gradecat4 (for 12th grade)

NOTE: Four dummy variables will be created if you still have all four grades left in your dataset. If for example you have dropped grade six, then you will only get three dummy variables and gradecat1 will be 8th grade.

Once you have your subpop variables created, you can use them with svy:tab. i.e., to look at smoking in the home by current smoking status among 8th graders:

```
svy:tab d14use d49, subpop(gradecat2) row per
```

30-day use cigarettes	does anyone who lives with you now smoke cigarettes?(secondary)		
	no	yes	Total
Use	40.24	59.76	100
No use	70.91	29.09	100
Total	68.96	31.04	100
Key: row percentages			
Pearson:			
Uncorrected	chi2(1)	=	257.4423
Design-based	F(1, 120)	=	99.6559 P = 0.0000

Interpretation: Among 8th graders who smoke, 67% live with someone who smokes. Among 8th graders who do not smoke, 31% live with a smoker. The p-value is 0.000, so 8th grader who smoke are more likely to live with someone who smokes compared to 8th graders who don't smoke.

You could also look at this the other way, i.e., to look at current smoking status by smoking in the home among 8th graders:

```
svy:tab d49 d14use, subpop(gradecat2) row per
```

Does anyone who lives with you now smoke cigarettes? secondary)	30-day use cigarettes		
	yes	no	Total
No	3.696	96.3	100
Yes	12.2	87.8	100
Total	6.334	93.67	100
Key: row percentages			
Pearson:			
Uncorrected	chi2(1)	=	257.4423
Design-based	F(1, 120)	=	99.6559 P = 0.0000

Interpretation: Among 8th graders who live with a smoker, 14% smoke cigarettes. Among 8th graders who do not live with a smoker, 4% smoke cigarettes. The p-value is 0.000, so 8th grader who live with a smoker are more likely to smoke compared to 8th graders who do not live with a smoker.

Another way to conduct stratified analyses is to use the “over” command. The “over” command in STATA 9 that replaces the “by” command used in previous versions. The variable or variables in parentheses after the over command define your subpopulations, i.e., to look at current smoking by grade and gender. Since you are looking at the mean, make sure that the response you are interested in is equal to 1 and the other responses are equal to 0.

```
recode d14use 1=1 2=0
svy:mean d14use, over(grade g05)
```

```
Over: grade g05
_subpop_1: 6 female
_subpop_2: 6 male
_subpop_3: 8 female
_subpop_4: 8 male
_subpop_5: 10 female
_subpop_6: 10 male
_subpop_7: 12 female
_subpop_8: 12 male
```

Over	Mean	Std. Err.	[95% Conf. Interval]	
d14use				
_subpop_1	.0138102	.002176	.0095207	.0180997
_subpop_2	.0207991	.0022957	.0162736	.0253246
_subpop_3	.0703174	.0065421	.0574212	.0832135
_subpop_4	.0604561	.0039968	.0525772	.068335
_subpop_5	.1226204	.0105658	.1017923	.1434485
_subpop_6	.1309227	.0102785	.1106609	.1511844
_subpop_7	.1637366	.0145536	.1350475	.1924257
_subpop_8	.2273656	.0136101	.2005363	.2541948

Interpretation: The key lets you know that _subpop_1 represents 6th grade females, so current smoking for 6th grade females is 1.4%. Current smoking for 6th grade males is 2.1%.

You can also use this with a continuous variable like BMI to get a mean by grade and gender:

```
svy:mean bmi, over(grade g05)
```

```
Over: grade g05
_subpop_1: 8 female
_subpop_2: 8 male
_subpop_3: 10 female
_subpop_4: 10 male
_subpop_5: 12 female
_subpop_6: 12 male

      Over      Mean   Std. Err.   [95% Conf. Interval]
bmi
_subpop_1  21.48329   .1700908   21.14652   21.82006
_subpop_2  21.70731   .1954243   21.32038   22.09423
_subpop_3  22.40279   .2319733   21.9435    22.86208
_subpop_4  22.87559   .2061686   22.46739   23.28378
_subpop_5  23.04815   .1527548   22.7457    23.35059
_subpop_6  24.51104   .2443785   24.02718   24.99489
```

Interpretation: The key lets you know that `_subpop_1` represents 8th grade females, so the mean BMI for 8th grade females is 21.5. The mean BMI for 8th grade males is 21.7.

6

HYS Data Analysis – Quick Example

This section provides a few examples of how to run crosstab analyses in STATA with:

STATA set up commands for analysis:

- State sample data
- State census data
- County sample, census, or mixed sampling data
- ESD level analysis

STATA commands for simple crosstabs:

- One variable by grade
- One variable by grade and gender
- One recoded variable by race and grade
- Two variables by grade
- Two variables by race
- Two variables by grade and gender

For more in depth information on STATA coding including examples of coding and output, see the [Data Analysis – Detailed State Sample Examples](#) section. For a quick reference of commonly used STATA commands see the next section [Useful STATA Commands and Options](#).

Set Up for Survey Analysis

The following STATA commands to set up for each different level of analysis. For more information on set up commands see the section General Set Up for Survey Analysis.

NOTE: Some datasets do not have the variable “schgrd”, but instead have the variable “schgnoid”. If your dataset has “schgnoid”, use it in place of schgrd in these STATA commands.

State Sample Data

```
gen fakewt=1
svyset [pweight=fakewt], psu(schgrd)
keep if staterec==1
```

State Census Data

```
gen fakewt=1
svyset [pweight=fakewt]
```

County Sample Data ~ for Counties without Samples (Census)

In 2008, these counties included: Adams, Asotin, Benton, Chelan, Clallam, Columbia, Cowlitz, Douglas, Ferry, Franklin, Garfield, Grant, Grays Harbor, Island, Jefferson, Kitsap, Kittitas, Klickitat, Lewis, Lincoln, Mason, Okanogan, Pacific, Pend Oreille, San Juan, Skagit, Skamania, Stevens, Wahkiakum, Walla Walla, Whatcom, Whitman, Yakima.

```
keep if conum==X*
*Insert your county number (conum) for X (see Demographic variables, conum)
keep if corec==1
gen fakewt=1
svyset [pweight=fakewt]
```

County Sample Data ~ for Sampled Counties

In 2008, these counties included: Pierce, King and Snohomish (plus Clark 6th and 8th, Thurston 6th only, and Spokane 6th grade if only analyzing those grades – see Mixed Sampling below).

```
keep if conum==X*
*Insert your county number (conum) for X (see Demographic variables, conum)
keep if corec==1
gen fakewt=1
svyset [pweight=fakewt], psu(schgrd)
```

County Sample Data ~ for Counties with Mixed Sampling

In 2010, counties with mixed sampling (some grades census and some grades sampled) included:

Clark: 6th and 8th were sampled, 10th and 12th grade were census

Spokane: : 6th and 8th were sampled, 10th and 12th grade were census

Thurston: 6th grade was sampled, 8th, 10th and 12th grades were census

Clark Example:

```
keep if conum==6
keep if corec==1
gen fakewt=1
gen psu=schoolid +10000
replace psu=schgrd if (grade==6 | grade==8)
svyset [pweight=fakewt], psu(psu)
```

Spokane Example:

```
keep if conum==32
keep if corec==1
gen fakewt=1
gen psu=schoolid +10000
replace psu=schgrd if (grade==6 | grade==8)
svyset [pweight=fakewt], psu(psu)
```

Thurston Example:

```
keep if conum==34
keep if corec==1
gen fakewt=1
gen psu=schoolid +10000
replace psu=schgrd if grade==6
svyset [pweight=fakewt], psu(psu)
```

Regional ESD Data

ESD regions are made up of counties or parts of counties, some which are sampled and some which are census. The following coding will set up analysis of any ESD.

```
keep if esdnum==X*
*Insert your ESD number (esdnum) for X
keep if esdrec==1
gen id = _n
gen psu=id + 10000
replace psu=schgrd if conum==6 & (grade==6 | grade==8)
replace psu=schgrd if (conum==17)
replace psu=schgrd if (conum==27)
replace psu=schgrd if (conum==31)
replace psu=schgrd if conum==32 & (grade==6 | grade==8)
replace psu=schgrd if (conum==34 & grade==6)
svyset [pweight=esdwt], psu(psu), strata(conum)
```

Data Analysis Example

*Current marijuana use by grade

```
svy:tab d21use grade, col per se ci obs format(%3.2f)
```

Cross one variable, d21use (current marijuana – already coded as no use or any use) by grade.

Formatting Options:

- col for column percentages
- per for results displayed in %
- se for standard error (to convert se to 95% ci, you need to *1.96)
- ci for confidence intervals
- obs for “n”
- format(%3.2f) designates the

*Current marijuana use by grade and gender

```
tab g05, gen(gender)  
rename gender1 girl  
rename gender2 boy
```

Generate binary (0,1) GENDER dummy variables for subpopulations.

```
svy:tab d21use grade, subpop(girl) col per se ci obs format(%3.2f)  
svy:tab d21use grade, subpop(boy) col per se ci obs format(%3.2f)
```

*or

```
tab grade, gen (gradecat)  
svy:tab d21use g05, subpop(gradecat1) col per se ci obs format(%3.2f)  
svy:tab d21use g05, subpop(gradecat2) col per se ci obs format(%3.2f)  
svy:tab d21use g05, subpop(gradecat3) col per se ci obs format(%3.2f)  
svy:tab d21use g05, subpop(gradecat4) col per se ci obs format(%3.2f)
```

You can also generate binary (0,1) GRADE subpopulations depending on how you want your output to look

*Excess pop drinking by 5 race codes (API Asian and Pacific Islander together)

```
codebook g06  
gen race=g06  
recode race 1=4 2=5 3=3 4=2 5=4 6=1 7=. 8=.  
lab def race 1"white" 2"hispanic" 3"black" 4"api" 5"indian"  
lab val race race  
lab var race "5 category race group"
```

Define and attach the new response option labels. Provide the new variable with a description.

```
codebook h09  
gen sodaex=h09  
recode sodaex 1=0 2=0 3=1 4=1 5=1  
lab def soda 0"1 or less" 1"2 or more"  
lab val sodaex soda  
lab var sodaex "excess soda drinking, 2 or more per day"
```

Create a new variable and recode

```
svy:tab sodaex race, subpop(gradecat1) col per se ci obs format(%3.2f)  
svy:tab sodaex race, subpop(gradecat2) col per se ci obs format(%3.2f)  
svy:tab sodaex race, subpop(gradecat3) col per se ci obs format(%3.2f)  
svy:tab sodaex race, subpop(gradecat4) col per se ci obs format(%3.2f)
```


*Current marijuana use by excess pop drinking

```
svy:tab sodaex d2luse, subpop(gradecat1) row per se ci obs format(%3.2f)
svy:tab sodaex d2luse, subpop(gradecat2) row per se ci obs format(%3.2f)
svy:tab sodaex d2luse, subpop(gradecat3) row per se ci obs format(%3.2f)
svy:tab sodaex d2luse, subpop(gradecat4) row per se ci obs format(%3.2f)
```

*Current marijuana use by excess pop drinking among white students

```
gen white6=race
recode white6 1=1 2=0 3=0 4=0 5=0
replace white6=0 if (grade==8 | grade==10 | grade==12)

gen white8=race
recode white8 1=1 2=0 3=0 4=0 5=0
replace white8=0 if (grade==6 | grade==10 | grade==12)

gen white10=race
recode white10 1=1 2=0 3=0 4=0 5=0
replace white10=0 if (grade==6 | grade==8 | grade==12)

gen white12=race
recode white12 1=1 2=0 3=0 4=0 5=0
replace white12=0 if (grade==6 | grade==8 | grade==10)
```

Generates binary (0,1)
combined RACE and
GRADE subpopulations

```
svy:tab sodaex d2luse, subpop(white6) row per se ci obs format(%3.2f)
svy:tab sodaex d2luse, subpop(white8) row per se ci obs format(%3.2f)
svy:tab sodaex d2luse, subpop(white10) row per se ci obs format(%3.2f)
svy:tab sodaex d2luse, subpop(white12) row per se ci obs format(%3.2f)
```

*Current marijuana use by excess pop drinking among boys

```
gen boy6=g05
recode boy6 1=0 2=1
replace boy6=0 if (grade==8 | grade==10 | grade==12)

gen boy8=g05
recode boy8 1=0 2=1
replace boy8=0 if (grade==6 | grade==10 | grade==12)

gen boy10=g05
recode boy10 1=0 2=1
replace boy10=0 if (grade==6 | grade==8 | grade==12)

gen boy12=g05
recode boy12 1=0 2=1
replace boy12=0 if (grade==6 | grade==8 | grade==10)
```

Generate binary (0,1) combined
GENDER and GRADE

```
svy:tab sodaex d2luse, subpop(boy6) row per se ci obs format(%3.2f)
svy:tab sodaex d2luse, subpop(boy8) row per se ci obs format(%3.2f)
svy:tab sodaex d2luse, subpop(boy10) row per se ci obs format(%3.2f)
svy:tab sodaex d2luse, subpop(boy12) row per se ci obs format(%3.2f)
```

NOTE: Use caution with crosstabs of variables with low prevalence, or when you are using small subpopulations to NOT report results if there are less than 5 observations per cell when running state level data, or less than 10 observations per cell when running sub-state level analysis.

7

Comparing State and Local Data

This section describes how to compare local data to the state. How you compare data depends on the type of data you have and the types of comparisons you want to make.

The easiest way to compare state and local data is to use the HYS Reports of Results. Reports of Results were generated by the HYS survey contractor, RMC Research Corporation, for school buildings, districts, education service district regions, and counties that participated in the survey at the minimum level. Reports of results for the state sample, state sample subpopulations (gender and race), and counties are available on the Department of Health's Healthy Youth Survey website. Building, district, ESD and county reports also include the state sample results so comparisons can be made by using the confidence intervals to determine differences (if confidence intervals do not overlap then the difference is statistically significant). This is always a good first step, even if you go on to run your comparisons in STATA. You can confirm your results with the produced reports.

Now that you have these results, you can also use an "Excel Tool for Determining Statistical Significance", available at: <http://www.AskHYS.net/Training>.

You can also do formal comparisons for statistical testing with STATA. How you make comparisons depends on the data sets you have and the comparison you want to run. We recommend that when are determining statistically significant differences between your local data and the state that you use compare yourself to the rest of the state sample (i.e., the state sample minus your local results).

When you report percentage point estimates and confidence intervals for the state sample, you may want to use the full state sample results, so that you do not contradict previously published state results. If you do this, you should note in your methods or under your results where your results came from and how your comparisons were conducted.

Appending

Use the “append” command if you want to add datasets with similar variables. For example if you wanted to combine your local and state sample Healthy Youth Survey results you can use the “append” command. Appending simply adds the additional data respondents below the originals respondents matching up the responses to the variable names.

Note: STATA defines your original data (the one you open first) as the “master data” and the new data you are appending on as your “using data”.

Data Preparation:

You need to create a new variable that will differentiate the respondents from each datasets. Open your 2010 state sample dataset and create a new variable for location:

```
use "C:2010 state.dta"  
gen location=0
```

If you are comparing your results to the rest of the state, you also need to drop any of your local schools from the state sample. Double check your conum variable to see if it was dropped and then save the file under a new name.

```
drop if conum==X*
```

```
*Insert your county number (conum) for X (see Demographic variables, conum)  
tab conum
```

Be careful to save your new dataset under a different name. Don’t save over your original state sample dataset.

```
save "C:2010 state location.dta"
```

Open your 2010 local dataset and create the same location variable with a different value:

```
use "C:2010 local.dta"  
gen location=1  
keep if corec==1  
save "C:2010 local location.dta"
```

Sometimes it is useful to only include the variables that you will need for your analysis. Use the “drop” or “keep” command to get rid of any unnecessary variables in both dataset before you append. This can speed up analysis and decrease the chance that STATA may become confused during the append.

Append the data:

Open your new 2010 state dataset and append using:

```
use "C: 2010 state location.dta"  
append using "C:2010 local location.dta"
```

Label your new location variable:

```
lab var location "state and local identifier"  
lab def location 0"state" 1"county"  
lab val location location
```

Append Investigation:

It is important to verify that your append came out correctly. To make sure that all of the data is there run a tab by location to see if you have the same number of respondents as you did in both of the original data sets (you should not have any missing data:

```
tab location, missing
```

If everything looks good, save your new combined dataset with a new file name:

```
save "C: 2010 state and local combo.dta"
```

Comparing Local vs. the Rest of the State Sample

Now that you have a combined state and local dataset, you need to open it and set it up for survey analysis:

```
use "C: 2010 state and local combo.dta"
```

If your local data is a sample (such as King, Pierce, Snohomish counties) then set up for analysis with:

```
gen fakewt=1  
svyset [pweight=fakewt], psu(schgrd)
```

If your local data is census data (most counties, all districts and schools) then set up for analysis with:

```
gen fakewt=1  
gen id = _n  
gen psu = id + 5000  
replace psu = schgrd if location==0  
svyset [pweight=fakewt], psu(psu)
```

This creates a psu with individual responses for your local census and groups school building responses for the state sample.

If your local data has "mixed" sampling (Clark, Spokane and Thurston) then set up for analysis like this example for Clark:

```
gen fakewt=1  
gen id = _n  
gen psu = id + 5000  
replace psu = schgrd if location==0  
replace psu = schgrd if location==1 & (grade==6 | grade==8)  
svyset [pweight=fakewt], psu(psu)
```

This creates a psu with individual responses for your local census grades 10 and 12, local sample in grades 6 and 8, and groups school building responses for the state sample.

Now you are ready to run a svy:tab by your group variable. You will need to first create a subpopulation to run your variable by a specific grade:

```
tab grade, gen(gradecat)  
rename gradecat2 eight  
svy:tab d14use location, subpop(eight) col se obs
```

8

Comparing Years of Data

This section describes how to combine multiple years of HYS data. It includes information about how to use the append command. Append allows you to add more respondents to your data.

Note: STATA defines your original data as the “master data” and the new data you are appending on as your “using data”.

Appending

Use the “append” command if you want to add datasets with similar variables. For example if you wanted to combine your 2008 and 2010 Healthy Youth Survey results you can use the “append” command. Appending simply adds the additional data respondents below the originals respondents matching up the responses to the variable names.

Data Preparation:

You need to create a new variable that will differentiate the respondents from each datasets. Open your 2010 dataset and create a new variable for year:

```
use "C:2010 data.dta"  
gen year=2010  
save "C:2010 data year.dta"
```

Open your 2008 dataset and create the year variable:

```
use "C:2008 data.dta"  
gen year=2008  
save "C:2008 data year.dta"
```

Sometimes it is useful to only include the variables that you will need for your analysis. Use the “drop” or “keep” command to get rid of any unnecessary variables in both dataset before you append. This can speed up analysis and decrease the chance that STATA may become confused during the append.

Append the data:

Open your 2010 dataset and append using:

```
use "C:2010 data year.dta"  
append using "C:2010 data year.dta"
```

Append Investigation:

It is important to verify that your append came out correctly. In general, the 2010 variables will stay in the same order and any variables that were unique to the 2008 data will now be at the bottom of your variable list. To make sure that all of the data is there run a tab by year to see if you have the same number of respondents as you did in both of the original data sets:

```
tab year, missing
```

You should not have any unexpected missing data. You may also want to run some frequencies to verify that you are getting the same results as you were before your append.

If everything looks good, save your new combined dataset with a new file name:

```
save "C:2008 and 2010 combo.dta"
```

Analysis Stratified by Year

At this time we are not recommending that you use STATA to determine significant trends over time, only to determine changes from a single survey administration to another, i.e., a change from 2008 to 2010. For trend analysis, we recommend that you have at least 5 data points and use the regression analysis program Joinpoint.

Joinpoint is available at : <http://srab.cancer.gov/joinpoint>

Now that you have a combined state and local dataset, you need to open it and set it up for survey analysis. The following is a comparison of current alcohol use for 8th and 10th graders from 2006 to 2010 using the state sample:

```
use "C:2008 and 2010 combo.dta"
```

If you are comparing 2008 and 2010 state sample data, or local sample data (such as King, Pierce, Snohomish counties) then set up for analysis with:

```
gen fakewt=1  
svyset [pweight=fakewt], psu(schgrd)
```

If you are comparing 2008 and 2010 local census data (most counties, all districts and schools) then set up for analysis with:

```
gen fakewt=1  
svyset [pweight=fakewt]
```

If you are comparing local census data that is sampled one year and not the other, or in one grade or not the other, see Chapter 5 HYS Data Analysis in STATA – the section General Set Up for Survey Analysis and the sub-section on County with mixed sampling analysis.

Then you will need to create subpopulations to run your variable by a specific grade:

```
tab grade, gen(gradecat)
rename gradecat1 six
rename gradecat2 eight
rename gradecat3 ten
rename gradecat4 twelve
svy:tab d20use year, subpop(eight) col se obs per
```

30-day alcohol use:		year		Total
	2008	2010		
yes	16.11 (.7937)	14.41 (.6574)		5.21 (.5324)
	1362	1363		2725
no	83.89 (.7937)	85.59 (.6574)		84.79 (.5324)
	7094	8096		1.5e+04
Total	100	100		100
	8456	9459		1.8e+04

Key: column percentages
(linearized standard errors of column percentages)
number of observations

Pearson:

Uncorrected	chi2(1)	=	34.8233	
Design-based	F(1, 374)	=	2.8813	P = 0.0904

- Interpretation: 8th grade current alcohol use was 16.1% in 2008 and 14.4% in 2010. A significant change did not occur from 2008 to 2010 (p-value is 0.09, NOT less than 0.05).

```
svy:tab d20use year, subpop(ten) col se obs per
```

30-day alcohol use:		year		Total
	2008	2010		
yes	31.68 (.8335)	27.65 (.9518)		29.66 (.7064)
	2143	1873		4016
no	68.32 (.8335)	72.35 (.9518)		70.34 (.7064)
	4622	4902		9524
Total	100	100		100
	6765	6775		1.4e+04

Key: column percentages (linearized standard errors of column percentages) number of observations

Pearson:

Uncorrected	chi2(1)	=	121.8556	
Design-based	F(1, 374)	=	10.5132	P = 0.0013

- Interpretation: 10th grade current alcohol use was 31.7% in 2008 and 27.7% in 2010. There was a significantly decrease from 2008 to 2010 (p-value is 0.001, less than 0.05).

When to Combine Multiple Years of Data

We generally recommend that all analyses be done stratified by year. However, under certain conditions, you may want to consider combining years of data. Some possible reasons to combine years include:

1. If your crosstabs don't meet the minimum cell requirements (5 per cell for state and 10 per cell for local).
2. If you have small number of respondents, like in smaller counties, or when analyzing non-core items located toward the end of the survey form.
3. If you want to analyze variables that are only applicable to a small group, such as trying to find out how many students with current asthma visited an emergency room in the past year.

Methods for Combining Years

We recommend the following decision rules for year-standardization when you are considering using year-combined single grade estimates:

1. Crude: If there is no substantial difference in a factor across years, you could report a "crude" estimate and note that the results are from multiple years. If you run a factor by year in STATA, the Total column will give you this "crude" estimate, or you can run an analysis without year as a variable or subpopulation (i.e., not stratifying by grade). In this case, you want to set up your analysis to include "year" as strata, i.e., for state sample analysis:

```
svyset [pweight=fakewt], psu(schgrd) strata(year)
```

2. Average: If there is a significant difference in a factor by year, but the purpose of your analysis is to simply express the burden of a condition, then you can use an average of the year specific results. Averaging gives equal weight to the results for each year, instead of giving equal weight to each respondent.

For example, if you wanted to estimate the percent of 10th graders who seriously considered suicide, run the factor by year then add the estimates for each year together and divide by the number of years, e.g., in 2006 and 2008 the statewide average seriously considering suicide was: 16.2%, calculated by $(15.1 + 17.3)/2$. Unfortunately this method does not give you a confidence interval.

3. Standardized: If there is significant difference in a factor by year and the purpose of your analysis is to present an assessment of underlying factors that may lead to a condition, then it would be appropriate to use a year-standardized estimate.

For example, if you are displaying the percent of youth smokers by gender who say that tobacco is easy to get and want to illustrate that it is different for males and females to inform planning, then you should use a year-standardized estimate.

Year-Standardized Estimates

Year-standardization ensures that each year contributes equally to the overall percent estimate. This can be especially useful if the number of respondents differ by year, e.g., there was greater participation in 2008 than in 2006. To generate a year-standardized estimate, you must weight the data.

Steps for Creating Year-Standardized Estimates

Using the 2006 and 2008 state samples, here is the methodology for weighting the data to create year-standardized estimates (i.e., combining both years 2006 and 2008).

According to the 2008-2009 and the 2010-2011 OSPI enrollment data for the state (available on their website: <http://www.k12.wa.us> under the Data and Reports section), there are:

	2008	2010	Combined
6 th graders	77,313	78,639	155,952
8 th graders	78,999	78,576	157,575
10 th graders	85,359	82,072	167,431
12 th graders	80,013	84,319	164,332

Looking at the number of valid respondents in the 2008 and 2010 state sample:

```
tab grade year
```

	2008	2010
6 th graders:	9,068	11,549
8 th graders:	8,730	9,723
10 th graders:	6,907	6,889
12 th graders:	5,641	5,908

For each grade, the enrollments for each year are added together to produce a combined enrollment number for the years. Then for each grade and year, the combined enrollment for that grade is divided by the total number of respondents for that specific grade and year.

```
gen yearwt=.
replace yearwt = 155952/9068 if (grade==6 & year==2008)
replace yearwt = 155952/11549 if (grade==6 & year==2010)
replace yearwt = 157575/8730 if (grade==8 & year==2008)
replace yearwt = 157575/9723 if (grade==8 & year==2010)
replace yearwt = 167431/6907 if (grade==10 & year==2008)
replace yearwt = 167431/6889 if (grade==10 & year==2010)
replace yearwt = 164332/5641 if (grade==12 & year==2008)
replace yearwt = 164332/5908 if (grade==12 & year==2010)
```

Year-Standardized Example

The following is an example looking at the prevalence of diabetes among 10th graders by soda drinking in the 2010 state sample:

```
gen diabetes=h77
recode diabetes 1=0 2=1 3=0
lab def yesno 1"yes" 0"no"
lab val diabetes yesno
```

```

tab grade, gen(gr)
rename gr3 ten
gen fakewt=1
svyset [pweight=fakewt], psu(schgrd)
svy:tab diabetes h09, subpop(ten) col se obs per format(%3.2f)

```

how many sodas or pops did you drink yesterday? (do not count diet soda.)						
diabetes	none	1	2	3	4 or more	Total
No	95.76 (0.56) 1944.00	96.20 (0.72) 709.00	95.15 (1.70) 216.00	95.12 (2.28) 78.00	87.23 (6.41) 41.00	95.68 (0.59) 2988.00
Yes	4.24 (0.56) 86.00	3.80 (0.72) 28.00	4.85 (1.70) 11.00	4.88 (2.28) 4.00	12.77 (6.41) 6.00	4.32 (0.59) 135.00

- Looking at 10th graders, it appears that diabetes is higher among those who drink more soda, but notice that the number of respondents is too small to report for those drinking 3 sodas (n=4), and those drinking 4 or more sodas is small (n=6).
- You could consider combining grades, it just depends on how you want to present the results. In this case, it is probably be fine to combine grades because the 2010 prevalence of diabetes by grade isn't significantly different (p-value=0.40) – see the next section on Combining Grade Levels. If you were analyzing the use of a substance, you might not want to consider combining grades because the prevalence usually increases by grade – so it would be better to combine years.

Using a combined 2008 and 2010 dataset (see the previous Appending section), we can create the “yearwt” that we calculated previously by combining the enrollments from each year by grade and dividing them by the number of respondents for each year and grade.

```

use "C:2008 and 2010 combo.dta"
gen yearwt=.
replace yearwt = 155952/9068 if (grade==6 & year==2008)
replace yearwt = 155952/11549 if (grade==6 & year==2010)
replace yearwt = 157575/8730 if (grade==8 & year==2008)
replace yearwt = 157575/9723 if (grade==8 & year==2010)
replace yearwt = 167431/6907 if (grade==10 & year==2008)
replace yearwt = 167431/6889 if (grade==10 & year==2010)
replace yearwt = 164332/5641 if (grade==12 & year==2008)
replace yearwt = 164332/5908 if (grade==12 & year==2010)
gen diabetes=h77
recode diabetes 1=0 2=1 3=0
lab def yesno 1"yes" 0"no"
lab val diabetes yesno
tab grade, gen(gr)
rename gr2 eight
svyset [pweight=yearwt], psu(schgrd) strata(year)

```

```
svy:tab diabetes h09, subpop(ten) col se obs per format(%3.2f)
```

```

how many sodas or pops did you drink yesterday?
diabetes

```

	none	1	2	3	4 or more	Total
No	95.77 (0.39) 3690.00	95.74 (0.50) 1438.00	5.36 (1.06) 473.00	90.87 (2.05) 169.00	89.91 (3.12) 107.00	95.47 (0.36) 5877.00
Yes	4.23 (0.39) 163.00	4.26 (0.50) 64.00	4.64 (1.06) 23.00	9.13 (2.05) 17.00	10.09 (3.12) 12.00	4.53 (0.36) 279.00

- Notice that the number of respondents who had diabetes and drank multiple sodas is now reportable with two years of data.

If you just wanted to know if 10th graders with diabetes were more likely to drink 4 or more sodas for both years combined, you could collapse the soda drinking variable:

```

gen soda4=h09
recode soda4 1=0 2=0 3=0 4=0 5=1
lab def soda4 1"4+ sodas" 0">4 sodas "
lab val soda4 soda4
svy:tab diabetes soda4, subpop(ten) col se obs per format(%3.2f)

```

```

diabetes

```

	>4 sodas	4+ sodas	Total
No	95.58 (0.33) 5770.00	89.91 (3.12) 107.00	95.47 (0.36) 5877.00
Yes	4.42 (0.33) 267.00	10.09 (3.12) 12.00	4.53 (0.36) 279.00

Pearson:

Uncorrected	chi2(1)	= 26.8594	
Design-based	F(1, 330)	= 8.3178	P = 0.0042

- Interpretation: In 2008 and 2010 combined, 10th graders who reported drinking 4 or more sodas yesterday were more likely to have diabetes (10.1%) compared to youth who drank less soda (4.4%), (p -value = 0.004).

9

Combining Grade Levels

This section describes when it is acceptable to combine grade levels and how to create grade-adjusted and high school estimates.

When to Combine Grades

We generally recommend that all analyses be done stratified by grade, see Analysis by Grade in section 5. However, under certain conditions it may be desirable to combine the results from different grade levels. Some possible reasons to combine grades include:

1. If your crosstabs don't meet the minimum cell requirements (5 per cell for state and 10 per cell for local).
2. If you have small number of respondents, like in smaller counties, or when analyzing non-core items located toward the end of the survey form.
3. If you want to analyze variables that are only applicable to a small group, such as trying to find out how many students with current asthma visited an emergency room in the past year.
4. If you need to produce a high school estimate for comparison with the YRBS.

Methods for Combining Grades

We recommend the following decision rules for grade-adjustment when you are considering using grade-combined estimates for a single year of data:

4. Crude: If there is no substantial difference in a factor across grades, you could report a "crude" estimate and note that the results are from multiple grades. If you run a factor by grade in STATA, the Total column will give you this "crude" estimate, or you can run an analysis without grade as a variable or subpopulation (i.e., not stratifying by grade).
5. Average: If there is a significant difference in a factor by grade, but the purpose of your analysis is to simply express the burden of a condition, then you can use an

average of the grade specific results. Averaging gives equal weight to the results for each grade, instead of giving equal weight to each respondent.

For example, if you wanted to estimate the percent of youth who seriously considered suicide, run the factor by grade then add the estimates for each grade together and divide by the number of grades, e.g., in 2010 the statewide average seriously considering suicide was: 15.4%, calculated by $(14.6 + 17.6 + 13.9)/3$. Unfortunately this method does not give you a confidence interval.

6. Adjusted: If there is significant difference in a factor by grade and the purpose of your analysis is to present an assessment of underlying factors that may lead to a condition, then it would be appropriate to use a grade-adjusted estimate.

For example, if you are displaying the percent of youth smokers by gender who say that tobacco is easy to get and want to illustrate that it is different for males and females to inform planning, then you should use an grade-adjusted estimate.

Grade-Adjusted Estimates

Grade-adjusted estimates ensure that each grade group contributes equally to the overall percent estimate, instead of giving equal “weight” to each respondent. They can be especially useful if the number of respondents differ by grade, e.g., there are more 8th grade respondents than 10th grade respondents. To generate a grade-adjusted estimate, you must weight the data.

This is similar to “age-adjusted” analyses often used in Healthy People 2010 or other national measures where population demographics change over time and may influence the factor you are trying to measure.

Steps for Creating Grade-Adjusted Estimates

Using the 2010 state sample, here is the methodology for weighting the data to create grade-adjusted estimates (i.e., combining all grades together 6, 8, 10 and 12):

According to the 2010-2011 OSPI enrollment data for the state (available on their website: <http://www.k12.wa.us> under the Data and Reports section), there are:

6th enroll:	78,639
8th enroll:	78,576
10th enroll:	82,072
12th enroll:	84,319
Total:	323,606

Looking at the number of valid respondents in the 2010 state sample, there are:

<code>tab grade</code>	
6th graders:	11,549
8th graders:	9,723
10th graders:	6,889
12th graders:	5,908

The enrollments for each grade are added together to produce a combined enrollment number for the grades. Then for each grade, the combined enrollment is divided by the total number of respondents for that specific grade:

```
gen gradewt=.
replace gradewt = 323606/11549 if grade==6
replace gradewt = 323606/9723 if grade==8
replace gradewt = 323606/6889 if grade==10
replace gradewt = 323606/5908 if grade==12
svyset [pweight=gradewt], psu(schgrd)
```

Grade-Adjusted Example

The following is an example looking at the prevalence of diabetes by heavy soda drinking by grade in the 2010 state sample:

```
gen diabetes=h77
recode diabetes 1=0 2=1 3=0
lab def yesno 1"yes" 0"no"
lab val diabetes yesno

gen soda4=h09
recode soda4 1=0 2=0 3=0 4=0 5=1
lab def soda4 1"4+ sodas" 0">4 sodas"
lab val soda4 soda4

tab grade, gen(gr)
gen fakewt=1
svyset [pweight=fakewt], psu(schgrd)
svy:tab diabetes soda4, subpop(gr2) col se obs per format(%3.2f)
svy:tab diabetes soda4, subpop(gr3) col se obs per format(%3.2f)
svy:tab diabetes soda4, subpop(gr4) col se obs per format(%3.2f)
```

svy:tab diabetes soda4, subpop(gr2) col se obs per format(%3.2f)				
diabetes	>4 sodas	4+ sodas	Total	
No	96.53 (0.36) 3671.00	89.42 (3.00) 93.00	96.34 (0.36) 3764.00	
Yes	3.47 (0.36) 132.00	10.58 (3.00) 11.00	3.66 (0.36) 143.00	

svy:tab diabetes soda4, subpop(gr3) col se obs per format(%3.2f)				
diabetes	>4 sodas	4+ sodas	Total	
No	95.81 (0.52) 2947.00	87.23 (6.37) 41.00	95.68 (0.59) 2988.00	
Yes	4.19 (0.52) 129.00	12.77 (6.37) 6.00	4.32 (0.59) 135.00	

svy:tab diabetes soda4, subpop(gr4) col se obs per format(%3.2f)				
diabetes	>4 sodas	4+ sodas	Total	
No	96.64 (0.38) 2617.00	88.24 (3.77) 60.00	96.43 (0.38) 2677.00	
Yes	3.36 (0.38) 91.00	11.76 (3.77) 8.00	3.57 (0.38) 99.00	

Throughout this manual: STATA commands are in grey and STATA output is in black

- Looking at the results for each grade, it appears that diabetes is higher among those who drink more soda, but that the number of respondents is pretty small for those with diabetes and drinking 4 or more sodas; 11 8th graders, 6 10th grades and 8 12th graders.
- You can also see that the prevalence of diabetes among heavy soda drinkers doesn't vary too much by grade; 10.6% of 8th graders, 12.8% of 10th graders, and 11.8% of 12th graders.

To calculate grade-adjusted estimates we can create the “gradewt” that we calculated previously by combining the grade enrollments and dividing it by the number of respondents for each grade.

```
use "C:2010 data year.dta"
gen gradewt=.
replace gradewt = 323606/11549 if grade==6
replace gradewt = 323606/9723 if grade==8
replace gradewt = 323606/6889 if grade==10
replace gradewt = 323606/5908 if grade==12
svyset [pweight=gradewt], psu(schgrd)
gen diabetes=h77
recode diabetes 1=0 2=1 3=0
lab def yesno 1"yes" 0"no"
lab val diabetes yesno
gen soda4=h09
recode soda4 1=0 2=0 3=0 4=0 5=1
lab def soda4 1"4+ sodas" 0">4 sodas "
lab val soda4 soda4
svy:tab diabetes soda4, col se obs per format(%3.2f)
```

diabetes	>4 sodas	4+ sodas	Total
No	96.32 (0.26) 9235.00	88.44 (2.39) 194.00	96.15 (0.27) 9429.00
Yes	3.68 (0.26) 352.00	11.56 (2.39) 25.00	3.85 (0.27) 377.00
Pearson:			
Uncorrected	chi2(1)	= 35.2283	
Design-based	F(1, 168)	= 32.9369	P = 0.0000

- Interpretation: In 2010, 8th, 10th and 12th graders combined who reported drinking 4 or more sodas yesterday were more likely to have diabetes (11.6%) compared to youth who drank less soda (3.7%), (p-value = 0.000).

Synthetic High School Estimates

The Centers for Disease Control and Prevention's Youth Risk Behavior Survey (YRBS) measures health behaviors of students in grades 9, 10, 11 and 12. They report "high school" estimates that combine all four grades. YRBS high school estimates are often used for setting bench marks, like the Healthy People 2010. In order to compare HYS results to national measures, we can create a synthetic high school estimate by following the steps for grade-adjusted weighting (described above) and applying an additional weight for the non-surveyed grades 9th and 11th. According to the 2010-2011 OSPI enrollment data for the state, there are:

Grade	Enrolled	% High School
9th	84,498	0.2556
10th	82,072	0.2483
11th	79,636	0.2409
12th	84,319	0.2551
Total	330,525	1.0000

To create a weight for each grade, we include the proportion that each grade contributes towards the high school enrollment and a proportion that takes into account how much the grade level should contribute to the overall estimate. For example, $\frac{1}{2}$ of the 8th grade estimate and $\frac{1}{2}$ of the 10th grade estimate should be used to create a 9th grade estimate.

Grade	Grade Weight Formula	Reasoning
8th	$0.2556 * 0.5 = \mathbf{0.1278}$	contributes to $\frac{1}{2}$ of 9 th grade
10th	$0.2556 * 0.5 = 0.1278$ $0.2483 * 1 = 0.2483$ $0.2409 * 0.5 = 0.1205$ $\mathbf{0.4966} = 0.1278 + 0.2483 + 0.1205$	contributes to $\frac{1}{2}$ of 9 th grade for the 10 th grade contributes to $\frac{1}{2}$ of 11 th grade
12th	$0.2409 * 0.5 = 0.1205$ $0.2551 * 1 = 0.2551$ $\mathbf{0.3756} = 0.1205 + 0.2551$	contributes to $\frac{1}{2}$ of 11 th grade for the 12 th grade
Total = (gr8*0.1278) + (gr10*0.4966) + (gr12*0.3756)		

The coding for generating 2010 synthetic high school estimates is:

```
gen hswt=.
replace hswt=.1278*100 if grade==8
replace hswt=.49655*100 if grade==10
replace hswt=.37555*100 if grade==12
```


Or if you want STATA to do the math for you, you can use the following formula:

```
gen hswt=.
replace hswt=(84498/330525*100*.5) if grade==8
replace hswt=((84498/330525*100*.5)+(82072/330525*100*1) ///
+(79636/330525*100*.5)) if grade==10
replace hswt=((79636/330525*100*.5)+(84319/330525*100*1)) if grade==12
svyset, clear
svyset [pweight=hswt], psu(schgrd)
svy:tab d14use grade, col se obs per format(%3.2f)
```

cigarettes	grade			Total
	8	10	12	
Use	6.57 (0.47) 624.00	12.66 (0.89) 860.00	19.59 (1.24) 1145.00	13.81 (0.68) 2629.00
No use	93.43 (0.47) 8871.00	87.34 (0.89) 5933.00	80.41 (1.24) 4700.00	86.19 (0.68) 19504.00

- Notice that the weighted synthetic high school estimate for current cigarette smoking is 13.8% ±1.3.

10

Adding Additional Data

This section describes how to add more data onto your HYS data. It includes information about how to use the merge command. Merge allows you to add additional data to your original data by joining a common variable.

Note: STATA defines your original data as the “master data” and the new data you are appending on as your “using data”.

Merging

Merging is used when the data you want to add has at least one variable in common with your original data set, like school building number, or county number.

For example if you wanted to conduct analysis of the state sample data according to the four classifications for urban/rural and you had a dataset with those classification by school building number, you could add the classification to your HYS data with merge.

Data Preparation:

Keep your merge simple, don't include unnecessary variables. Sometimes both dataset have the same (duplicate) variables. Duplicates can confuse STATA and cause problems with your merge. Only keep the duplicate variables that you need to make a proper merge. If you want to keep other duplicate variables, rename them so they will be distinct variables in your new dataset.

You also need to make sure that the variable in your new data is in the same format as your HYS data. For example the variable schgrd in the HYS dataset is numeric. If you are going to merge your new data with schgrd, you need to make sure that the schgrd variable in your new data is also numeric. If the schgrd variable in your new data is a string change it to numeric using the encode command:

```
encode schgrd, gen(school)
drop schgrd
rename school schgrd
```

Sort Using Data:

Your “using data” should be the dataset that you are adding on to the HYS dataset. Prior to merging, you need to sort your new dataset by your merge variable(s).

```
sort schgrd
```

After sorting, save your dataset with a new name. This is now referred to as your “using dataset”.

```
save "C:new using data.dta"
```

Sort Master Data:

Open your HYS dataset (your “master dataset”) and sort it by your merge variable(s).

```
sort schgrd
```

Merge the Data:

Once your master data is sorted you can merge on your new dataset:

```
merge (schgrd) using "C:new using data.dta"
```

Merge Investigation:

While merging, STATA will try to tell you if something looks wrong, look for any error messages. Error messages are usually in red font. Messages in green font are usually just letting you know that there were commonalities between the two datasets, like the same variable labels.

After the merge you will have a new variable “_merge”. You can use this variable to check the results of your merge. The response options provided for _merge are 1, 2 and 3:

```
tab _merge
```

```
1 = the using data did not have a match  
2 = the master data did not have a match  
3 = matched
```



Depending on the dataset you are adding, you may or may not be expecting all of the data to match. I.e., when we check our merge of the urban rural classifications to HYS we get mostly 3’s (matches), but we get some 2’s (non-matches in the master) This is OK because we know that the urban rural classifications includes all schools in the state and our HYS data only includes schools that participated in the survey. We expect that the schools that did not participate should not match. So in this case we would simply get rid of the non-matched data by dropping them:

```
drop if _merge==2
```

If we get some 1s with this same merge we would need to do more investigation. We expect that every school in our HYS dataset should have an urban rural classification. We can find out the names of the schools that didn’t match by:

```
tab schname if _merge==1
```

Then we would want to check our urban rural classification to see if that school existed. If not, we would need to figure out why ~ maybe the school is new or it changed its location in the past year, etc. If this is the case, you can update your urban rural dataset and then re-merge.

You can also look at your actual data to see what happened in your merge by looking in the “data browser” by click on the toolbar icon  or in the “data editor” by clicking on the toolbar icon . In either the browser or the editor you can sort by the `_merge` variable to see exactly which data are not matching up. The data browser opens faster than the data editor, but the data editor allows you to add and delete data. When you exit the data editor it will ask you if you want to preserve your changes – only keep changes you are sure you want.

Once you’re satisfied with your merge you can get rid of the “_merge” variable:

```
drop _merge
```

NOTE: You cannot merge on additional data until you drop the “_merge” variable or rename it.

For some reason, it usually takes most of us multiple attempts to get our merges correct. So don’t worry if you it takes you a few tries, and always investigate your merge to make sure it did what you wanted it to.

Checking Findings and Significance Online

This section describes the information available on the DOH HYS and AskHYS.net websites to verify your analysis results. When running data analysis in STATA it's always a good idea to verify your results by looking at previously produced results.

This section also includes information about an online tool for testing statistical significance when you are comparing two estimates that have 95% confidence intervals.

AskHYS.net Website

address is: <http://www.askhys.net>

AskHYS.net has three primary features:

1. Fact sheets on important HYS topics at the state and local level results
2. Reports of all HYS frequency results at the state and local levels
3. An interactive data query system to analyze state and local data.

1. AskHYS Fact Sheets

Currently, there are 21 topical fact sheets available with results from 2002, 2004, 2006, 2008 and 2010. State, ESD and County fact sheets available to all, and District and Building fact sheets are available to those with permission from district superintendents (must go through an approval process with OSPI).

Fact sheets can also be run by gender, but the general rules for crosstabs apply (at least 5 respondents in every cell for state fact sheets and 10 per cell for local).

The fact sheets include the following topics:

- Unintentional Injury (seat belts, helmets, life vests, and driving or riding with a drinking driver)

- Violent Behaviors (fighting, weapon carrying and not feeling safe at school, missing school because of safety issues, and gang membership)
- Harassment, Intimidation, and Bullying (bullying, harassment for sexual orientation, harassment by cell phone)
- Weight and Obesity
- Dietary Behaviors (eating fruits/vegetables, not eating with family, and drinking sweetened drinks and school and buying them at school)
- Physical Activity (not meeting physical activity recommendations, too much screen time and not participating in physical education)
- Asthma (current asthma, lifetime asthma, asthma attacks, and ER visits, missing school and using medication for asthma)
- Depression & Suicide (depression, considering, planning and attempting suicide, and likeliness to seek help if depressed)
- Sexual Behavior (ever had sex, age of first sex, number of sexual partners, condom use during last sexual intercourse)
- Current Substance Use (current cigarettes, alcohol, marijuana, methamphetamine, inhalants, Ritalin and prescription pain killer use)
- Alcohol Use (current and lifetime alcohol use, binge drinking, levels of alcohol use, anti-alcohol messages, and sources of, access to and perceptions of alcohol)
- Alcohol Use - a report for all grade levels combined on one sheet
- Tobacco Use (current cigarette, cigar, chew, bidis, cloves, pipes and hookah use)
- Marijuana Use (current use, perception of harm, ease of access, perception of acceptance from adults and peers)
- Community Risk Factors (availability of drugs, pro-drug use laws and norms, availability of handguns, low neighborhood attachment, and access to substances)
- Community Protective Factors (rewards including neighbors noticing good work, encouraging best and being proud, and opportunities including sports, service and activity clubs, and adults to talk to)
- School Risk Factors (academic failure including usually getting low grades and grades worse than others, and low commitment to school including school not meaningful or important for future, and cut school)
- School Protective Factors (opportunities including making decisions, talking to teacher, class involvement, other activities, and rewards including teachers saying good work, praising hard work and notifying parents of good work)
- Peer-Individual Risk Factors (perceived risk of drug use, attitudes favor drug use, friends drug use, and intentions to use drugs)
- Peer-Individual Protective Factors (per interaction, belief in a moral order, social skills)
- Family Protective Factors (opportunities including discussing problems, having fun and involved in decisions, and rewards including parents noticing good work, being proud and enjoying being with mom or dad)

Most fact sheets also include a chart with the association between one of the topics and academic achievement, e.g., cigarette smoking and academic achievement.

2. Q x Q Analysis

The Q x Q is an interactive data query system to analyze state and local frequencies and crosstabs. HYS data from 2002, 2004, 2006, 2008 and 2010 are available to analyze. State, ESD and County data can be accessed by all, and District and Building data are available to those with permission from district superintendents (must go through an approval process with OSPI).

When running a crosstab, you need to think about how you want your results to turn out before you select your variables. The variable that you drop into the first box will be the group you are interested in finding out more information about. The second variable you select is your outcome variable. For example:

- Do you want to know the prevalence of smoking among different race groups? Then select Demographics - Race/Ethnicity -[G06] Race/Ethnicity as your first variable and Tobacco - Current Use - [D14] Current Cigarette Smoking as your second variable.
- Do you want to know the prevalence of drinking alcohol among youth who smoke? Then select Tobacco - Current Use - [D14] Current Cigarette Smoking as your first variable and Alcohol - Current Use - [D20] Current Alcohol Drinking as your second.
- Or do you want to know the flip side, what is the prevalence of smoking among youth who drink alcohol? Then select Alcohol first and Cigarettes second.

Crosstabs on the Q x Q also have to follow the requirements for a minimum number of respondents per cell in order to produce results:

- For state level analysis you must have 5 or more respondents in each cell.
- For sub-state level analysis you must have 10 or more respondents in each cell.

Online Tool for Determining Statistical Significance

AskHYS.net Website

www.AskHYS.net/Training

There is an “Excel Tool for Determining Statistical Significance” on the HYS administration website reporting page. Scroll down to 3rd bullet under Information and Tools. The tool itself has cells to enter local and state data, but you can use this tool to test the difference between any two estimates with 95% confidence intervals.

For example, to test for differences in drinking any soda between 10th and 12th graders in 2010, statewide using the following results:

- 10th grade: 35.6% (± 3.0)
- 12th grade: 38.8% (± 2.8)

Input section		
	Percent	Plus or minus
Local Result	35.6	3.0
State Result	38.8	2.8

Output section	
p-value:	0.1264150

Notice that the tool provides cells to enter the percent and confidence interval for a “local” and a “state” result, but you can enter any two results you would like to compare. In this example, the results are both “state”, but one is the 10th grade result and the other is the 12th grade result.

Interpretation: P-value is .1264, which is greater than 0.05, so there is no difference in drinking any soda between 10th and 12th graders, in 2010 statewide.

Test for differences in drinking soda between 10th grade males and females in 2010, statewide using the following results:

- 10th grade males: 41.1% (± 3.1)
- 10th grade females: 30.7% (± 3.7)

Input section		
	Percent	Plus or minus
Local Result	41.1	3.1
State Result	30.7	3.7

Output section	
p-value:	.0000241

Interpretation: P-value is .00002, which is less than 0.05, so 10th grade males are more likely to drink any soda compared to 10th grade females, in 2010 statewide.

Additional Resources

Web Resources

Here are a few helpful resources on the Healthy Youth Survey, STATA, and statistical analysis. The links provided here are not in any way imply that the sources are endorsed by DOH. They are just some sites that we have found to be helpful.

Healthy Youth Survey

- AskHYS Website: <http://www.askhys.net>
- Department of Health HYS Website: <http://www.doh.wa.gov/DataandStatisticalReports/HealthBehaviors/HealthyYouthSurvey.aspx>
- HYS 2010 Administration Website : <http://www.hys.wa.gov/Reporting>
- History of Risk and Protective Factors: <http://www.hys.wa.gov/Reporting/RPHistory.pdf>

Stat Transfer

- Allows you to convert data files into STATA datasets. Available free trial at: <http://www.stattransfer.com>

STATA Resources

- STATA: <http://www.stata.com>
- UCLA: <http://www.ats.ucla.edu/stat/STATA>
- UNC: http://http://www.cpc.unc.edu/research/tools/data_analysis/statatutorial
- Princeton: <http://dss.princeton.edu/training>

Statistical Analysis

- AssessNow: <http://www.assessnow.info>
- Emory (Biostatistics): <http://www.sph.emory.edu/bios>
- Florida: <http://www.stat.ufl.edu/vlib/statistics.html>
- StatSoft: <http://www.statsoft.com/textbook/stathome.htm>
- JoinPoint regression program: <http://srab.cancer.gov/joinpoint>

Previously Used Computed/Calculated Variables

Asthma – Asthma Severity scale (available for HYS 2002, 2004, 2006 only)

Asthma may be classified according to the severity of symptoms experienced by an individual. This “symptom severity” classification is an indication of the extent to which an individual is affected by their asthma, not a measure of the seriousness of their asthma. For example, a person with clinically less severe asthma, but who does not manage their exposure to triggers or use medication appropriately, may have severe symptoms. Conversely, a person with clinically very severe asthma may control their condition well and experience relatively fewer symptoms.

For more discussion of asthma symptom severity, please refer to “The Burden of Asthma in Washington State 2005”, pages 63-65, available at:

http://www.doh.wa.gov/portals/1/Documents/Pubs/345-201_TheBurdenofAsthmaInWashingtonState.pdf

```
* Asthma severity scale
gen ast_sev=.
replace ast_sev=1 if(h74==1|h74==2|h73==1|h73==2|h73==3)
replace ast_sev=2 if(h74==3|h73==4)
replace ast_sev=3 if(h74==4|h73==5)
replace ast_sev=4 if(h74==5|h73==6)
replace ast_sev=0 if(currentasthma==0)
lab var ast_sev "Asthma Severity Scale"
lab def severe 1 "Mild intermittent" 2 "Mild Persistent" 3 "Moderate Persistent"
4 "Severe Persistent"
lab val ast_sev severe
```

Nutrition – Recode for Fruits and Vegetables Servings “Five a Day”

The past CDC recommendations for fruit and vegetable consumption for youth is five or more servings per day. The Healthy Youth Survey asks students how often they have eaten several common fruits and vegetables in the past week, and the responses are combined into an estimated daily consumption pattern. Note that the individual responses (for example, fv5, the frequency of carrot consumption alone) are not considered useful, and not included in any HYS reports.

It is important to recognize that HYS questions are framed as *times* per day, which is different than *servings*. Also providing serving size information doubles the estimated percent eating “five a day” when number of servings is asked ; see Bensley, L., Van Eenwyk, J, and Bruemmer, BA. (2003). Journal of the American Dietetic Association, 103:1530-1532. Thus we can estimate the percent of students who meet past nutrition guidelines (“five a day”) using this measure, but it is likely to be an over-estimate if students eat multiple servings at the time they eat fruits or vegetables.

```

* Average number of fruits/vegetables per day
gen numday1=fv1
recode numday1 1=0 2=.286 3=.714 4=1 5=2 6=3 7=4
gen numday2=fv2
recode numday2 1=0 2=.286 3=.714 4=1 5=2 6=3 7=4
gen numday3=fv3
recode numday3 1=0 2=.286 3=.714 4=1 5=2 6=3 7=4
gen numday4=fv4
recode numday4 1=0 2=.286 3=.714 4=1 5=2 6=3 7=4
gen numday5=fv5
recode numday5 1=0 2=.286 3=.714 4=1 5=2 6=3 7=4
gen numday6=fv6
recode numday6 1=0 2=.286 3=.714 4=1 5=2 6=3 7=4
gen numday=(numday1 + numday2 + numday3 + numday4 + numday5 + numday6)
replace numday=. if (numday1==. | numday2==. | numday3==. | numday4==. |
numday5==. | numday6==.)

* Number times fruits and vegetables were eaten per day
gen h07=.
replace h07=4 if numday8>=5
replace h07=3 if numday8<5
replace h07=2 if numday8<3
replace h07=1 if numday8<1
replace h07=. if numday8==.
lab def h07 1 "less than 1" 2 "1 to less than 3" 3 "3 to less than 5" 4 "5 or
more"
lab val h07 h07

* Poor nutrition - getting fewer than "five a day"
gen poornut=h07
recode poornut 1=0 2=0 3=0 4=1
lab def poornut 0"Fewer than 5 a day" 1"5+ fruit-veggies a day"
lab val poornut poornut

```

Disability Screener

The youth disability screener (YDS) was developed by the Seattle Quality of Life Group at the University of Washington. The questions were not asked on the HYS asked in 2010. This creates a new variable "disable" that can be used to compare variables for youth with disabilities and youth without disabilities. Bear in mind, the YDS questions are only available for 8th, 10th, and 12th graders and only on Form B.

```

* Youth disability screener
gen disable=.
replace disable = 1 if (h18==1 | h19==1 | h20==1 | h21==1)
replace disable = 2 if (h18==2 | h18==3) &(h19==2 | h19==3) & (h20==2 | h20==3)
& (h21==2 | h21==3)
lab def disable 1 "with disabilities" 2 "without disabilities"
lab val disable disable

```

More information is available at: <http://depts.washington.edu/seaql/YDS>

State Level Enrollments by Year and Coding for Synthetic High School Weights

For more information on calculating synthetic high school estimates is available in Chapter 9: Combining Grade Levels. The state enrollments for 2010 are on page 86.

2008-2009 State Enrollment

Grade	Enrolled	% High School
9th	83,125	0.2540
10th	80,446	0.2458
11th	80,091	0.2447
12th	83,616	0.2555
Total	327,278	1.0000

2008 weight coding:

```
gen hswt=.
replace hswt=(83125/327278*100*.5) if grade==8
replace hswt=((83125/327278*100*.5)+(80446/327278*100*1) ///
+(80091/327278*100*.5)) if grade==10
replace hswt=((80091/327278*100*.5)+(83616/327278*100*1)) if grade==12
```

2006-2007 State Enrollment

Grade	Enrolled	% High School
9th	90,444	0.2721
10th	84,476	0.2542
11th	80,193	0.2413
12th	77,242	0.2324
Total	332,355	1.0000

2006 weight coding:

```
gen hswt=.
replace hswt=(90444/332355*100*.5) if grade==8
replace hswt=((90444/332355*100*.5)+(84476/332355*100*1) ///
+(80193/332355*100*.5)) if grade==10
replace hswt=((80193/332355*100*.5)+(77242/332355*100*1)) if grade==12
```

2004-2005 State Enrollment

Grade	Enrolled	% High School
9th	89,970	0.2769
10th	83,315	0.2564
11th	77,443	0.2383
12th	74,248	0.2285
Total	324,976	1.0000

2004 weight coding:

```
gen hswt=.
replace hswt=(89970/324976*100*.5) if grade==8
replace hswt=((89970/324976*100*.5)+(80877/324976*100*1) ///
+(77443/324976*100*.5)) if grade==10
replace hswt=((77443/324976*100*.5)+(74248/324976*100*1)) if grade==12
```

2002-2003 State Enrollment

Grade	Enrolled	% High School
9th	87,842	0.2763
10th	80,877	0.2544
11th	76,759	0.2415
12th	72,404	0.2278
Total	317,882	1.0000

2002 weight coding:

```
gen hswt=.
replace hswt=(87842/317882*100*.5) if grade==8
replace hswt=((87842/317882*100*.5)+(83315/317882*100*1) ///
+(76759/317882*100*.5)) if grade==10
replace hswt=((76759/317882*100*.5)+(72404/317882*100*1)) if grade==12
```