STRATEGIC DATA PROJECT

AN INTRODUCTION TO THE SDP TOOLKIT FOR EFFECTIVE DATA USE

A GUIDE FOR CONDUCTING DATA ANALYSIS IN EDUCATION AGENCIES

www.gse.harvard.edu/sdp/toolkit

Toolkit Documents

An Introduction to the SDP Toolkit for Effective Data Use

Identify: Data Specification Guide

Clean: Data Building Guide for College-Going
Clean: Data Building Guide for Human Capital BETA

Connect: Data Linking Guide for College-Going
Connect: Data Linking Guide for Human Capital BETA

Analyze: College-Going Success Analysis Guide
Analyze: Human Capital Analysis Guide BETA

Adopt: Coding Style Guide

SDP Stata Glossary

VERSION: 1.2
Last Modified: September 2, 2013

| Authored by Todd Kawakita and the SDP Research Team

Dear Data Strategist in Education,

The **SDP Toolkit for Effective Data Use** is a resource that the Strategic Data Project (SDP) has developed to help data strategists in education learn to lead and effect change through data. Having worked with many partner educational agencies, SDP is well-versed in the difficulties in conducting large-scale analyses. We understand that collecting "silo-ed" data of varying quality, cleaning that data, and transforming that data into a decision-making tool can be a daunting process. Our toolkit will empower you to more effectively collect and analyze data that reveal key trends in your agency.

The toolkit is divided into 5 discrete steps that we believe capture the essence of an analytic process: **Identify**, **Clean**, **Connect**, **Analyze**, and **Adopt**. **Identify** provides a clear and detailed set of guidelines to help you know exactly what data to collect and how it should be coded.

Clean and **Connect** provide fully-developed synthetic data sets and instructions to help you clean data and link your data together into a single, powerful analysis file. **Clean** also includes a reference glossary of decision rules that provide standards for resolving common data problems.

Analyze uses the analysis file from **Connect** to develop key indicators to track performance along a student's education pathway. **Analyze** will help you produce data visualizations that answer questions such as:

- How are students in my agency transitioning from 9th grade to 10th grade? Graduating?
- Are students who graduate from my agency's high schools persisting and ultimately graduating from post-secondary institutions? How does this vary by high school?

We believe that any education agency committed to preparing every child for postsecondary success must embrace rigorous research that answers these fundamental questions. Agency leaders must believe that answers to these questions can be produced by data strategists in education agencies who are fully equipped and trained to conduct rigorous research from within. We believe that **Analyze** and its preceding steps can help equip and train education analysts and data strategists to do this work.

The last step in the toolkit, **Adopt**, is a guide toward best data management and coding practices designed to facilitate shared analysis and transparency. Many in the education analyst and data strategist community have cited the usefulness of this document for setting a standard in statistical coding practice.

At SDP, our mission is to **transform the use of data in education to improve student achievement**, and we hope that this toolkit empowers you to transform your use of educational data. It is our sincere hope that these standards of data practice and process become useful to you and adopted by the larger educational sector. Whatever your experience with data may be, fellow data geeks, data strategists, and educational leaders, we welcome you on this journey and hope you find the experience useful. Onto a new frontier of educational data use!

Sincerely,

Patty Diaz, Director of Education and Outreach
Todd Kawakita, Manager of Product Development

The SDP Toolkit helps **YOU, a data strategist in an education agency, do two crucial things:** (1) Build Clean Datasets, and (2) Produce Meaningful Analyses.

The toolkit consists of the following five steps:

1. Identify: Data Specification Guide
Identify essential data elements for analysis across your organization;

2. Clean: Data Building Guide
Clean and process data files you identify;

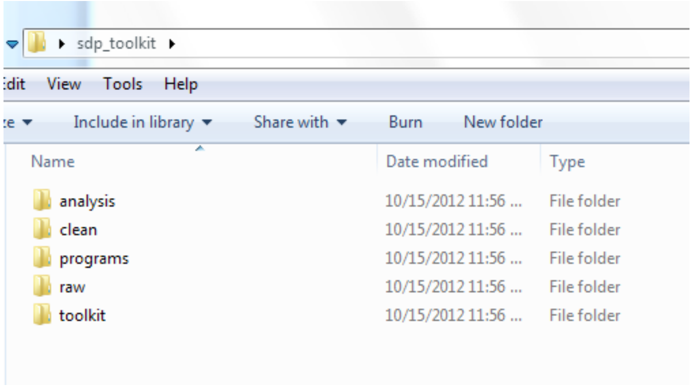
3. Connect: Data Linking Guide
Link cleaned data files into one analysis file;

4. Analyze: College-Going Analysis Guide
Conduct analyses that help answer key questions for your agency;

5. Adopt: Coding Style Guide
Adopt best practices to facilitate sharing and replication.

Steps 1-4 of the toolkit build upon knowledge sequentially. Thus, the logical place to start is **Identify**. However, based on your familiarity with data collection, cleaning, and analysis, you may decide to start elsewhere, though we do recommend skimming some of the preceding tools for reference.

Everything you need for the toolkit is available online at **<u>www.gse.harvard.edu/sdp/toolkit</u>** as a **zip file**. When unzipped, this file reveals an infrastructure including all the steps of the toolkit, the data files you need, and template files that serve as a shell for Stata code.



Name	Date modified	Type
analysis	10/15/2012 11:56 ...	File folder
clean	10/15/2012 11:56 ...	File folder
programs	10/15/2012 11:56 ...	File folder
raw	10/15/2012 11:56 ...	File folder
toolkit	10/15/2012 11:56 ...	File folder

CONTENT COVERAGE

This toolkit is based on the intensive work the Strategic Data Project (SDP) completes with partner agencies (school districts, states, and charter management organizations). The two focus areas of this work and the toolkit are: College-Going Success and Human Capital (Teacher Effectiveness).

SDP refers to analyses in these two research areas as "diagnostics." The table below indicates how each step of the toolkit supports exploration of questions relating to the College-Going and Human Capital (Teacher Effectiveness) diagnostics.

	COLLEGE-GOING DIAGNOSTIC	**HUMAN CAPITAL DIAGNOSTIC**
1. Identify: Data Specification Guide Identify essential data elements for analysis across your organization;	**Fully supports** identification of data elements required for both diagnostics.	
2. Clean: Data Building Guide Clean and process data files you identify;	**Fully supports** data cleaning for the College-Going Diagnostic.	**Fully supports** data cleaning for the Human Capital Diagnostic. This is currently a BETA release.
3. Connect: College-Going Data Linking Guide Link cleaned data files into one analysis file;	**Only supports** creation of analysis file for College-Going Diagnostic.	**Only supports** creation of analysis files for the Human Capital Diagnostic. This is currently a BETA release.
4. Analyze: College-Going Analysis Guide Conduct analyses that answer key questions for your agency;	**Only supports** analyses for College-Going Diagnostic.	**Only supports** analyses for Human Capital Diagnostic. This is currently a BETA release.
5. Adopt: Coding Style Guide Adopt best practices to facilitate sharing and replication.	**Independent** of any diagnostic. Provides generalizable best practices for data management and coding style.	

EXPECTATIONS AND PRIOR KNOWLEDGE

To progress through the toolkit, you will need:

- **Data** to get hands-on experience cleaning, connecting, and analyzing data. There are two ways to obtain data:
  1. **Use your agency's data**. Go through the **Identify** step to collect data you need and continue through the toolkit. This way you will produce data files and analytical results for your agency.
  2. **Use the provided files**. You may not have immediate access to your agency's data or may want to ensure you understand the toolkit before jumping in. If so, use the data we provide online to begin learning the toolkit and then check your answers with the provided solutions. Data files are available for download as part of the **zip file** containing the toolkit infrastructure or as separate data files at **www.gse.harvard.edu/sdp/toolkit**.

- **Knowledge of a statistical package (ideally Stata) and access to the software**. Statistical packages (Stata, SAS, SPSS, or R) are the preferred tools to manipulate large, longitudinal datasets. This toolkit uses Stata and its programming language. SDP provides a Stata glossary if you choose to use it.

- **A basic understanding of statistics**. The current version of the toolkit only requires a basic knowledge of statistics (i.e. you should understand what means, modes, medians and standard deviations are). This knowledge is not required for **Identify**.

For those less familiar with statistical programming: statistical programming is an important skill to manipulate large datasets in a robust, replicable way. Programming, or writing code, allows you to maintain a record of changes made to data that can be easily adapted and replicated later. If you're not familiar with statistical programming, this toolkit will help you learn it!

SUMMARIES OF EACH STEP

1. **Identify**: Data Specification Guide

Identify essential data elements for analysis across your organization

Identify: Data Specification Guide is a resource that helps you identify data elements required to conduct analyses of student achievement, postsecondary attainment, and teacher effectiveness. This guide helps you understand what information you should collect and how information should be organized to facilitate data cleaning and linking.

One of the goals of **Identify** is to help you think about how **disparate**, or **siloed**, data can be collected and brought together to support later stages of the toolkit (particularly **Connect** and **Analyze**).

Identify is also a resource for database architects who equip analysts with data. SDP recognizes that education agencies have pre-established systems to collect and manage student-, teacher- and human resource- databases. This toolkit allows database architects to pull data available in pre-existing warehouses and transform that data for analysis.

2. **Clean**: Data Building Guide

Clean and process data files you identify;

Clean: Data Building Guide is a series of tasks with step-by-step instructions to build clean data files from raw data you identify. Starting with a raw input file you obtain from **Identify** (e.g. Student Attributes), each task guides you through the process of preparing clean output files that can be linked together in Connect.

3. Connect: Data Linking Guide

Link cleaned data files into an analysis file;

Connect: Data Linking Guide is a tool to help you bring together the data you identified and cleaned in the previous steps into a single, powerful analysis file that captures student information through high school and college. This file will serve as the basis for all analyses conducted in **Analyze**.

4. Analyze: Analysis Guide

Conduct analyses that answer key questions for your agency;

Analyze: Analysis Guide provides a set of step-by-step instructions to generate key data visualizations from SDP's diagnostic analyses. These analyses are meant to help you answer important questions about student pathways through high school and college. The guide provides explanations of each visualization's purpose, along with guiding questions to lead a discussion of the results.

5. Adopt: Coding Style Guide

Adopt best practices to facilitate sharing and replication.

Adopt: Coding Style Guide outlines best practices related to file organization and programming to help others (and yourself down the line) understand your analytic work. Great coding is similar to great writing: it is well-organized and clear. By following our data management and coding conventions, you will organize data and write code in a clear, understandable way. These conventions will build capacity within your organization to use data more effectively.

SDP Stata Glossary

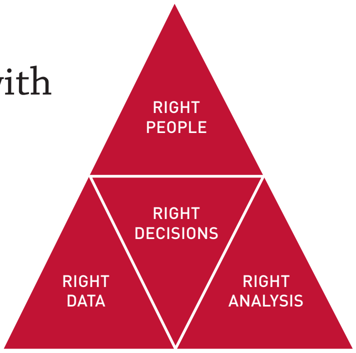
Brush up on Stata using a glossary of commonly used commands within the toolkit.

The **SDP Stata Glossary** is meant to help new and existing users of Stata learn the commonly used commands within the toolkit. The glossary outlines many useful commands and functions relevant to data cleaning and exploration. The glossary is provided in either alphabetical order or by topic.

The Strategic Data Project

OVERVIEW

The Strategic Data Project (SDP), housed at the Center for Education Policy Research at Harvard University, partners with school districts, school networks, and state agencies across the US. **Our mission is to transform the use of data in education to improve student achievement.** We believe that with the right people, the right data, and the right analyses, we can improve the quality of strategic policy and management decisions.



SDP AT A GLANCE

56 AGENCY PARTNERS
14 SCHOOL DISTRICTS
7 STATE EDUCATION DEPARTMENTS
2 CHARTER SCHOOL ORGANIZATIONS

79 FELLOWS
54 CURRENT
25 ALUMNI

CORE STRATEGIES

1. Placing and supporting top-notch analytic leaders as "Fellows" for two years with our partner agencies

2. Conducting rigorous diagnostic analyses of teacher effectiveness and college-going success using existing agency data

3. Disseminating our tools, methods, and lessons learned to many more education agencies

SDP DIAGNOSTICS

SDP's second core strategy, conducting rigorous diagnostic analyses using existing agency data, focuses on two core areas: (1) college-going success and attainment for students and (2) human capital (primarily examining teacher effectiveness).

The diagnostics are a set of analyses that frame actionable questions for education leaders. By asking questions such as, "How well do students transition to postsecondary education?" or "How successfully is an agency recruiting effective teachers?" we support education leaders to develop a deep understanding of student achievement in their agency.

ABOUT THE SDP TOOLKIT FOR EFFECTIVE DATA USESDP's third core strategy is to disseminate our tools, methods, and lessons learned to many more educational agencies. This toolkit is meant to help analysts in all educational agencies collect data and produce meaningful analyses in the areas of college-going success and teacher effectiveness. Notably, the analyses in this release of our toolkit primarily support questions related to college-going success. The data collection (Identify) and best practices (Adopt) stages of the toolkit, however, are applicable to any sort of diagnostic and convey general data use guidelines valuable to any analysts interested in increasing the quality and rigor of their analyses. Later releases will address analyses relating to teacher effectiveness.

Center for Education Policy Research
HARVARD UNIVERSITY

CENTER FOR EDUCATION POLICY RESEARCH
STRATEGIC DATA PROJECT
50 CHURCH ST., 4TH FLOOR, CAMBRIDGE, MA 02138
VOX 617.496.1563
FAX 617.495.2614
WWW.GSE.HARVARD.EDU/SDP