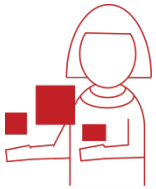
STRATEGIC DATA PROJECT

SDP TOOLKIT

FOR EFFECTIVE DATA USE

A GUIDE FOR CONDUCTING DATA ANALYSIS IN EDUCATION AGENCIES

1. Identify



5. Adopt

www.gse.harvard.edu/sdp/tools

Patty Diaz
Senior Program Manager, Fellows

Todd Kawakita
Manager of Product Development

Jared Silver
Data Architect and Manager

1 STRATEGIC DATA PROJECT

SDP TOOLKIT

FOR EFFECTIVE DATA USE

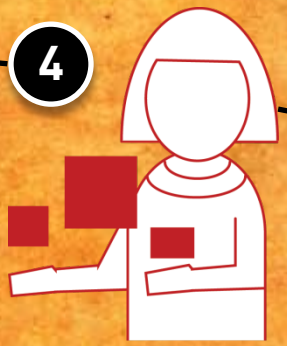
A GUIDE FOR CONDUCTING DATA ANALYSIS IN EDUCATION AGENCIES

2

3

1. Identify

4



5. Adopt

5 Q&A

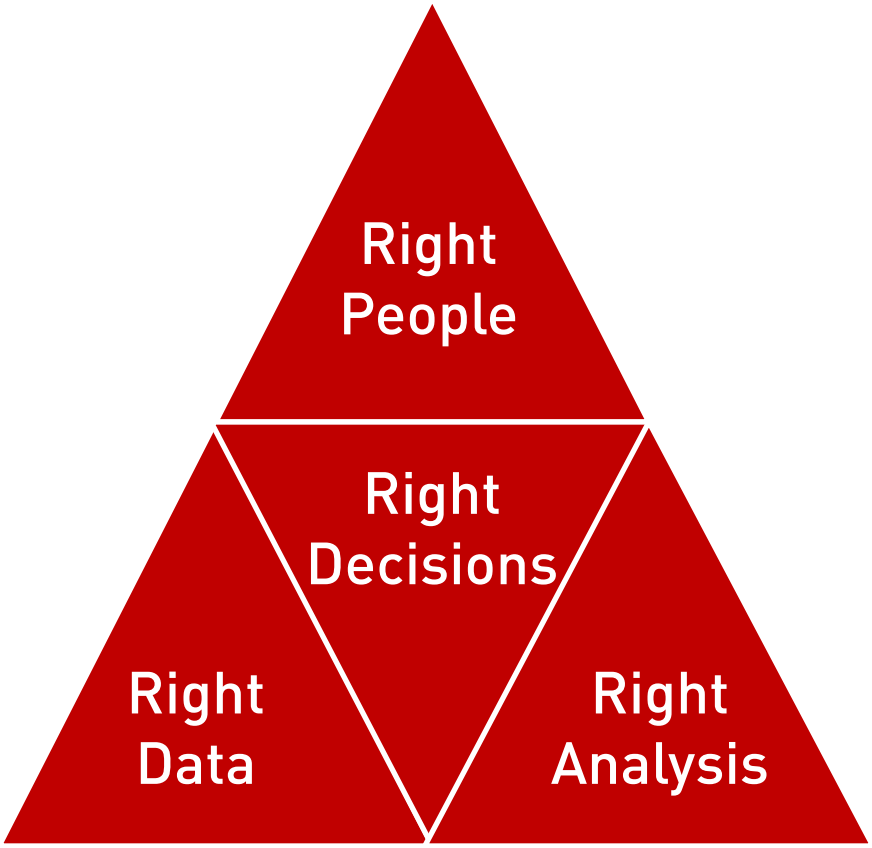
STRATEGIC DATA PROJECT

MISSION

Transform the use of data in education to improve student achievement.

STRATEGIC DATA PROJECT

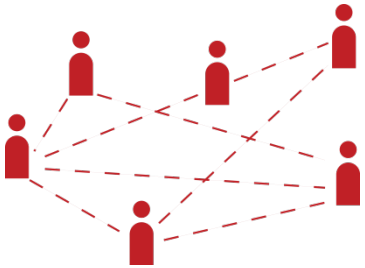
Theory of Action



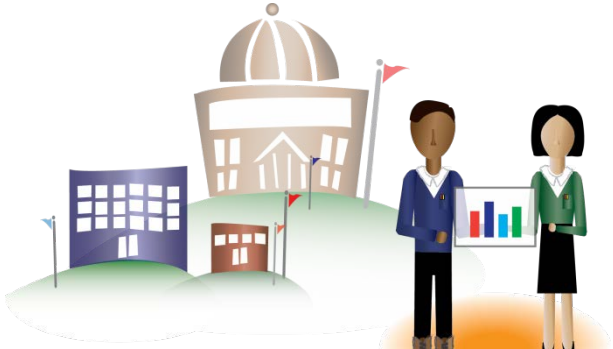
STRATEGIC DATA PROJECT

I. Fellows

Create and support a national network of high quality data analysts



who will influence policy at the local, state, and national levels.

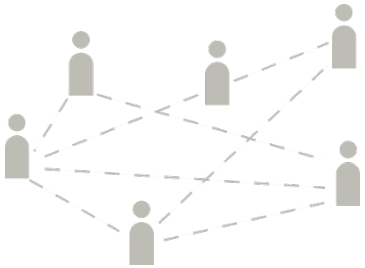


Place **Fellows** in partner agencies.

STRATEGIC DATA PROJECT

I. Fellows

Create and support a national network of high quality data analysts



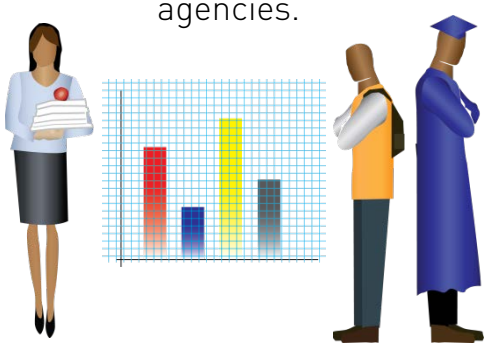
who will influence policy at the local, state, and national levels.

2. Diagnostics

Create policy- and management-relevant standardized analyses

for districts and states.

Perform **Diagnostics** in partner agencies.



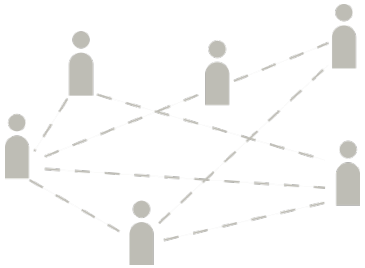
Human Capital
Teacher Effectiveness

College-Going Success

STRATEGIC DATA PROJECT

I. Fellows

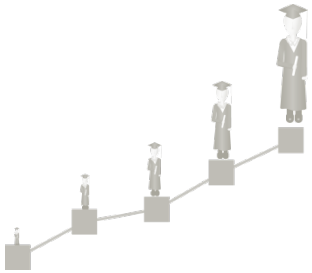
Create and support a national network of high quality data analysts



who will influence policy at the local, state, and national levels.

2. Diagnostics

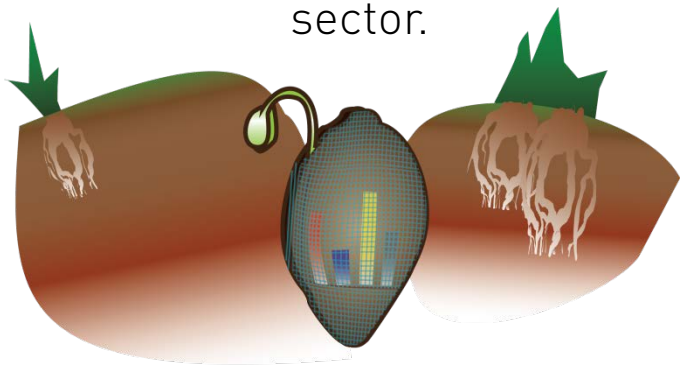
Create policy- and management-relevant standardized analyses



for districts and states.

3. Scale

Improve the way data is used in the education sector.

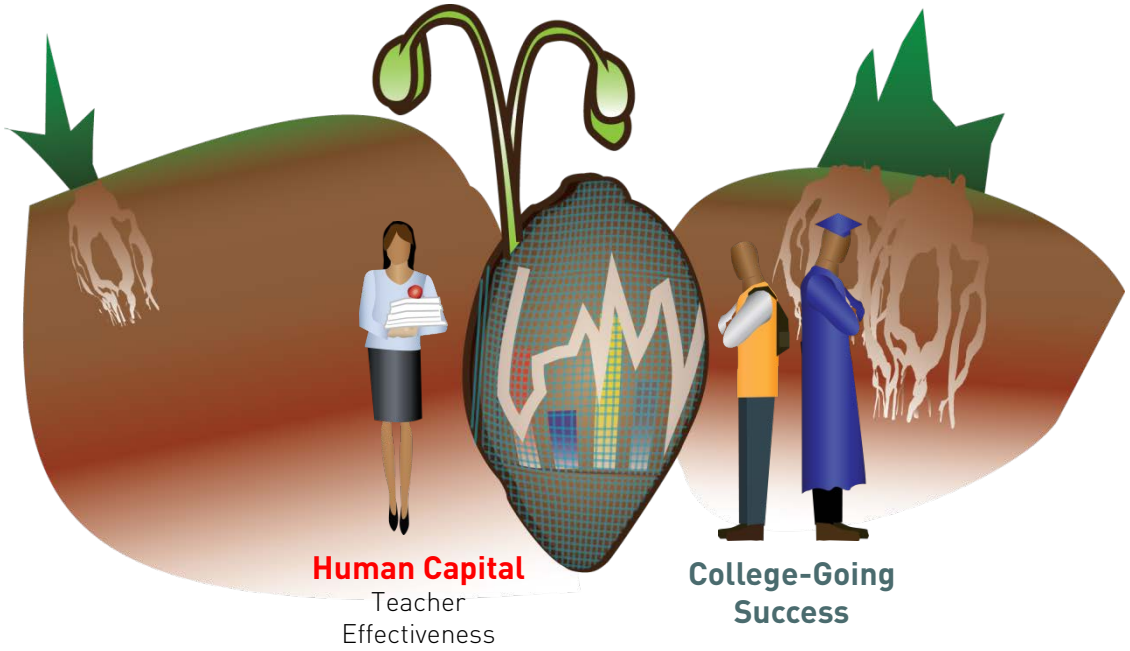


Achieve broad impact through wide dissemination of analytic **tools**, methods, and best practices.

SDP TOOLKIT

FOR EFFECTIVE DATA USE

A GUIDE FOR CONDUCTING DATA ANALYSIS IN EDUCATION AGENCIES



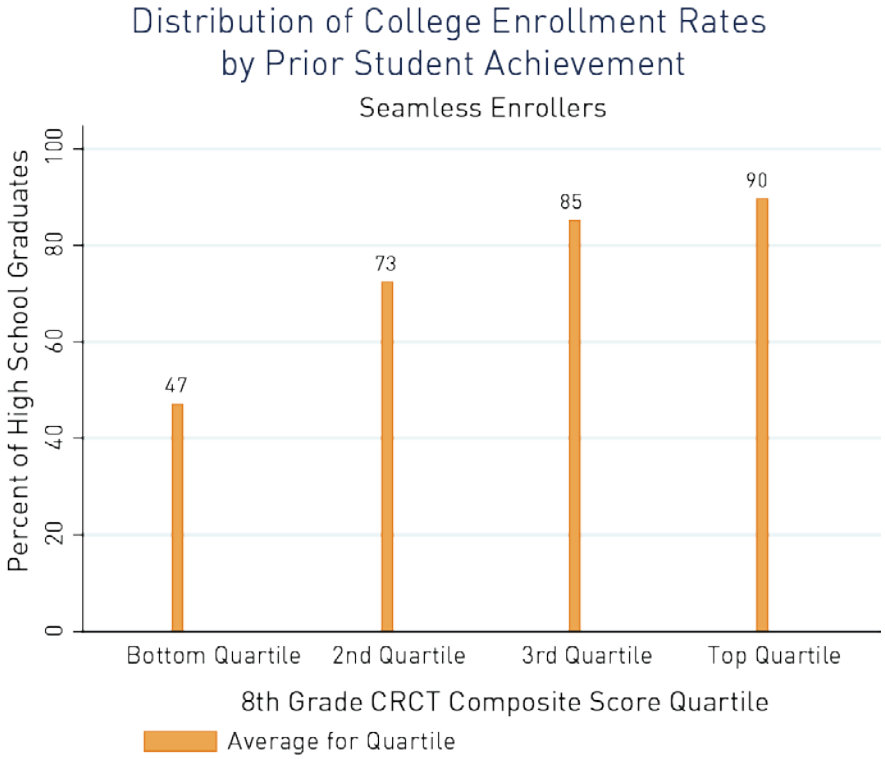
SDP TOOLKIT

FOR EFFECTIVE DATA USE

A GUIDE FOR CONDUCTING DATA ANALYSIS IN EDUCATION AGENCIES

College-Going Success

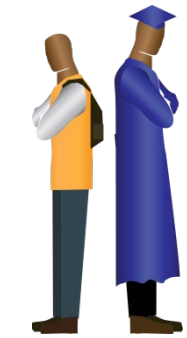
What is the relationship between 8th grade test scores and college enrollment rates?



SDP TOOLKIT

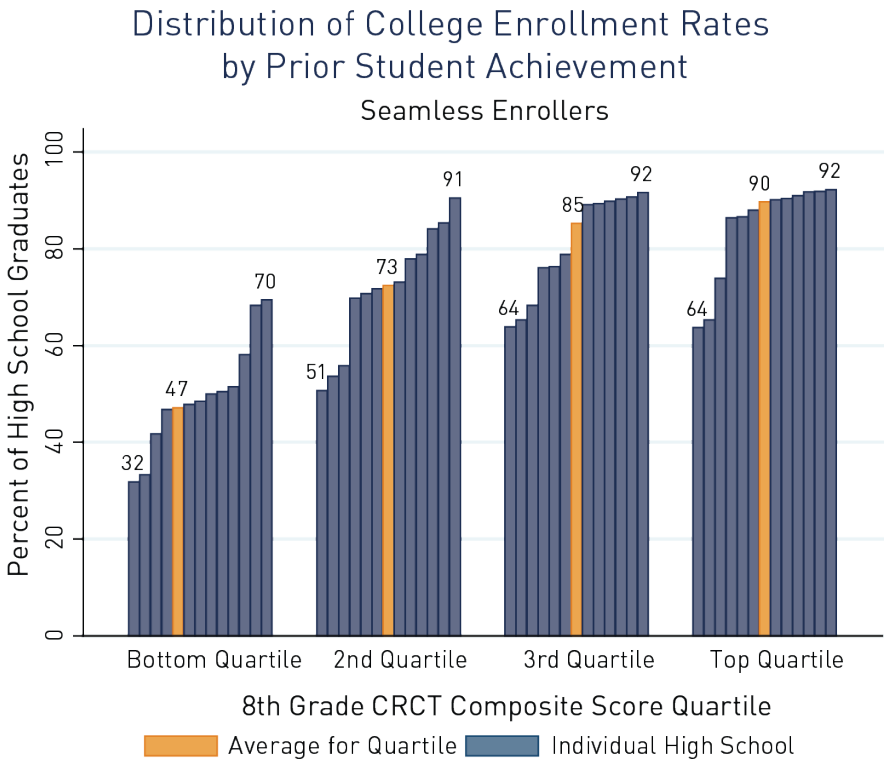
FOR EFFECTIVE DATA USE

A GUIDE FOR CONDUCTING DATA ANALYSIS IN EDUCATION AGENCIES



College-Going Success

What is the relationship between 8th grade test scores and college enrollment rates?



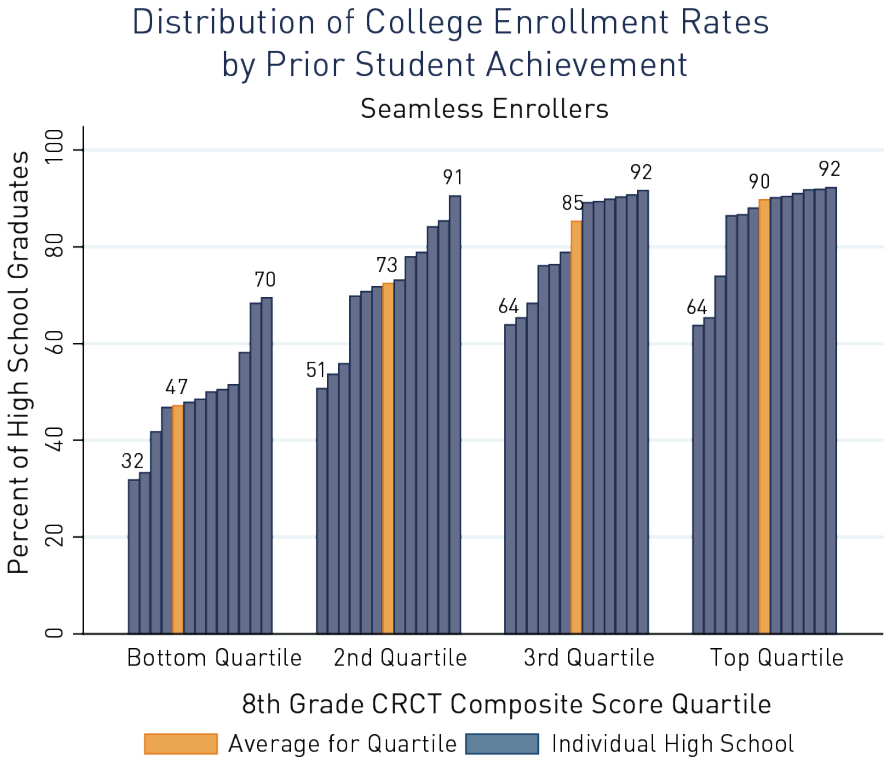
SDP TOOLKIT

FOR EFFECTIVE DATA USE

A GUIDE FOR CONDUCTING DATA ANALYSIS IN EDUCATION AGENCIES

1. Identify essential data elements

- Test Scores
- College Enrollment
- High School Graduation
- Student School Enrollment

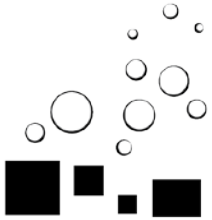


SDP TOOLKIT

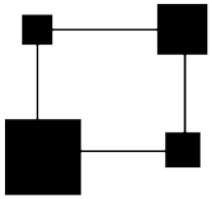
FOR EFFECTIVE DATA USE

A GUIDE FOR CONDUCTING DATA ANALYSIS IN EDUCATION AGENCIES

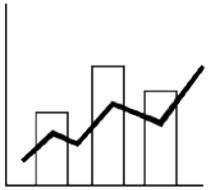
1. Identify essential data elements



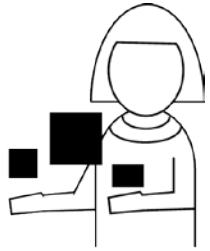
2. Clean check, and build variables for your datasets



3. Connect relevant datasets from different sources



4. Analyze your datasets

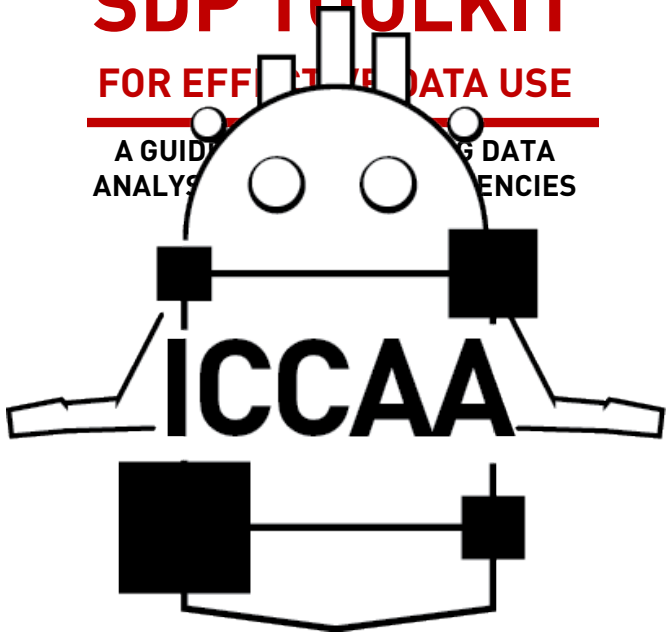


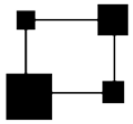
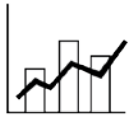
5. Adopt best practices to facilitate shared and replicable data analysis

SDP TOOLKIT

FOR EFF[illegible]ATA USE

A GUID[illegible]G DATA
ANALYS[illegible]ENCIES

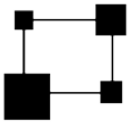
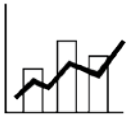


I	C	C	A	A
				
Identify: Data Specification Guide	**Clean**: Data Building Tasks	**Connect**: Data Linking Guide	**Analyze**: Diagnostic Analyses Guide	**Adopt**: Coding Style Guide

SDP TOOLKIT

FOR EFFECTIVE DATA USE

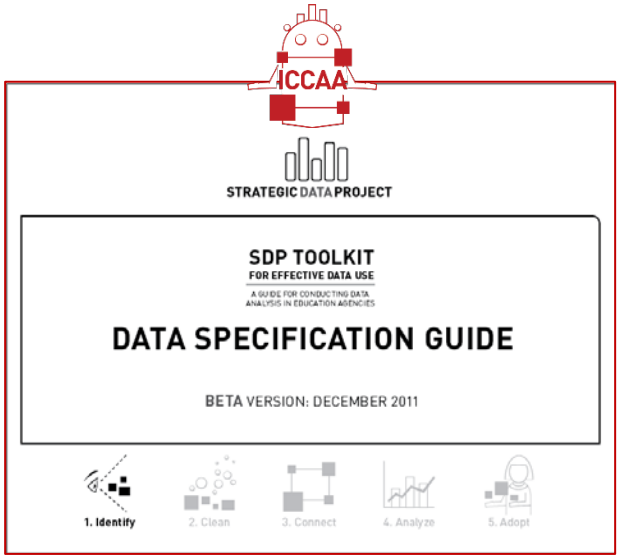
A GUIDE FOR CONDUCTING DATA ANALYSIS IN EDUCATION AGENCIES

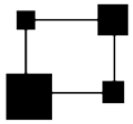
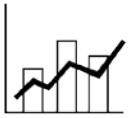
I	C	C	A	A
				
Identify: Data Specification Guide	**Clean**: Data Building Tasks	**Connect**: Data Linking Guide	**Analyze**: Diagnostic Analyses Guide	**Adopt**: Coding Style Guide

SDP TOOLKIT

FOR EFFECTIVE DATA USE

A GUIDE FOR CONDUCTING DATA ANALYSIS IN EDUCATION AGENCIES



I	C	C	A	A
				
Identify: Data Specification Guide	**Clean**: Data Building Tasks	**Connect**: Data Linking Guide	**Analyze**: Diagnostic Analyses Guide	**Adopt:** Coding Style Guide

1. Identify

Data Specification Guide

Successful data analysis begins with proper identification of data elements necessary to answer key questions of interest.

1. Identify
Data Specification Guide

5 **W**'s of Data Collection

WHY
are we collecting data?

Research questions?

WHAT

WHERE

WHO

WHEN

1. Identify
Data Specification Guide

5 **W**'s of Data Collection

WHY
are we collecting data?

Research questions?

WHAT
data is needed?

WHERE

WHO

WHEN

- **Student demographics** (ethnicity, gender, birth date)
- **Teacher demographics**
- **Student learning classifications** (ELL, SpEd, FRPL)
- **Teacher experience, pay, certifications** (HR data)
- **School enrollment, attendance, and graduation**
- **Class enrollment, class grades, student/teacher links**
- **Test scores**
- **College Enrollment** (NSC data)

1. Identify
Data Specification Guide

5 **W**'s of Data Collection

WHY	WHAT	WHERE	WHO	WHEN
are we collecting data? Research questions?	data is needed?	does the data live?		
	• **Student demographics** (ethnicity, gender, birth date) • **Teacher demographics** • **Student learning classifications** (ELL, SpEd, FRPL) • **Teacher experience, pay, certifications** (HR data) • **School enrollment, attendance, and graduation** • **Class enrollment, class grades, student/teacher links** • **Test scores** • **College Enrollment** (NSC data)	• **Student Information System** (SIS) • **Longitudinal Data Store** (LDS) or **Data Warehouse** (DW) • **HR Systems** • **Excel Spreadsheets** or **MS Access** • On **paper**!		

1. Identify
Data Specification Guide

5 **W**'s of Data Collection

WHY
are we collecting data?

Research questions?

WHAT
data is needed?

- **Student demographics** (ethnicity, gender, birth date)
- **Teacher demographics**
- **Student learning classifications** (ELL, SpEd, FRPL)
- **Teacher experience, pay, certifications** (HR data)
- **School enrollment, attendance, and graduation**
- **Class enrollment, class grades, student/teacher links**
- **Test scores**
- **College Enrollment** (NSC data)

WHERE
does the data live?

- **Student Information System** (SIS)
- **Longitudinal Data Store** (LDS) or **Data Warehouse** (DW)
- **HR Systems**
- **Excel Spreadsheets** or **MS Access**
- On **paper**!

WHO
Owns these systems and is responsible for delivering the data?

WHEN

1. Identify
Data Specification Guide

5 **W**'s of Data Collection

WHY
are we collecting data?

Research questions?

WHAT
data is needed?

- **Student demographics** (ethnicity, gender, birth date)
- **Teacher demographics**
- **Student learning classifications** (ELL, SpEd, FRPL)
- **Teacher experience, pay, certifications** (HR data)
- **School enrollment, attendance, and graduation**
- **Class enrollment, class grades, student/teacher links**
- **Test scores**
- **College Enrollment** (NSC data)

WHERE
does the data live?

- **Student Information System** (SIS)
- **Longitudinal Data Store** (LDS) or **Data Warehouse** (DW)
- **HR Systems**
- **Excel Spreadsheets** or **MS Access**
- On **paper**!

WHO
Owns these systems and is responsible for delivering the data?

WHEN
(over what date range/school years) do we need data for, in case of a longitudinal analysis?

Reliability of historical data elements?

1. Identify
Data Specification Guide

5 **W**'s of Data Collection

WHY
are we collecting data?

Research questions?

WHAT
data is needed?

- **Student demographics** (ethnicity, gender, birth date)
- **Teacher demographics**
- **Student learning classifications** (ELL, SpEd, FRPL)
- **Teacher experience, pay, certifications** (HR data)
- **School enrollment, attendance, and graduation**
- **Class enrollment, class grades, student/teacher links**
- **Test scores**
- **College Enrollment** (NSC data)

WHERE
does the data live?

- **Student Information System** (SIS)
- **Longitudinal Data Store** (LDS) or **Data Warehouse** (DW)
- **HR Systems**
- **Excel Spreadsheets** or **MS Access**
- On **paper**!

WHO
Owns these systems and is responsible for delivering the data?

WHEN
(over what date range/school years) do we need data for, in case of a longitudinal analysis?

Reliability of historical data elements?

HOW
should the data appear?

STUDENT ATTRIBUTES

Identifies unique observation: **sid**

Field Name	Values or Data Type	Definition	Importance	Notes
sid	numeric	Student identifier unique to each student. This identification number is typically assigned to a student upon enrollment in your agency. State agencies may have different identification numbers than district agencies for the same student.	5 Cannot Be Missing	
male	0 = female 1 = male	Student gender.	4 Absolutely Necessary	
race_ethnicity	1 = African American 2 = Asian American 3 = Hispanic 4 = American Indian 5 = White, not Hispanic 6 = Other 7 = Multiple	*For systems or school years within systems where race and ethnicity are treated as a combined variable.* If the system allows the indication of multiple categories simultaneously (e.g., African American and white) report "multiple."	4 Absolutely Necessary	Use either the race_ ethnicity combined variable, or separate ethnicity and race variables.
race	1 = African American 2 = Asian American 3 = American Indian 4 = White 5 = Other 6 = Multiple	*For systems or school years within systems where race and ethnicity are treated as separate variables.* If the system allows for the indication of multiple categories simultaneously (e.g., African American and white) report "multiple."	4 Absolutely Necessary	Use either the race_ ethnicity combined variable, or separate ethnicity and race variables.
ethnicity	0 = Hispanic 1 = not Hispanic	*For systems or school years within systems where race and ethnicity are treated as separate variables and Hispanic or Latino origin is asked as a separate question.*	4 Absolutely Necessary	Use either the race_ ethnicity combined variable, or separate ethnicity and race variables.
birth_date	date format (yyyy-mm-dd)	Student birth_date.	2 Good to Have	
first_9th_school_ year_reported	spring calendar year	The school year during which the student was a 9th grader for the first time. For this variable, report what the system explicitly recorded for first 9th grade school year. Not all systems will record this information.	1 Not Essential	
hs_diploma	0 = no high school diploma 1 = has high school diploma	Indicator variable equal to **1** if the student has received a high school diploma from the system.	4 Absolutely Necessary	Can sometimes be the same as a **graduated** flag.
hs_diploma_type	use local values	Any locally defined description of the type of diploma the student received. Include instances in which more than one type of diploma is observed, for example, Honors diploma, College Prep diploma, or General Education Diploma (GED) diploma.	4 Absolutely Necessary	Needed when multiple types of diplomas are issued.
hs_diploma_date	date format (yyyy-mm-dd)	The date on which the student received a high school diploma. If only a month and year, or only a school year is known report that partial information.	4 Absolutely Necessary	Can also be Graduation Date.
zip_code	xxxxx or xxxxx-yyyy	The zip code of the student's home address.	1 Not Essential	

- **Identifies uniqueness of observations** in each file to avoid data duplication (double counting)
- **Standardizes encoded values** for certain data points for consistency (i.e. ethnicity, subject matter)
- ***Importance* listed** for each data element

TABLE OF CONTENTS

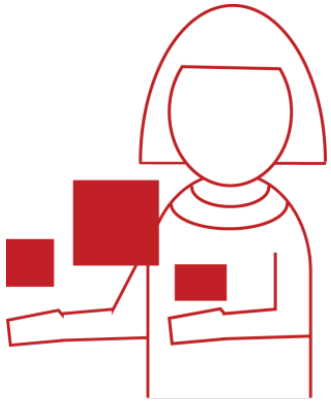
- **Broad range of data points** – covers many research questions, many data points common to different analyses
- **Structured in a way that facilitates analyses** and similar to many existing data systems

1. Identify
Data Specification Guide

- Specification meant to be a format to ***extract*** research data files to, **not as a stand alone database** structure

- May need to **merge** data across data stores to reach final layout

- **Software**
  - ***ETL:*** MS SSIS, Oracle WB, Informatica, DataStage
  - ***Reporting:*** Cognos, Crystal Reports, Oracle BI
  - ***Statistical:*** Stata, SAS, SPSS

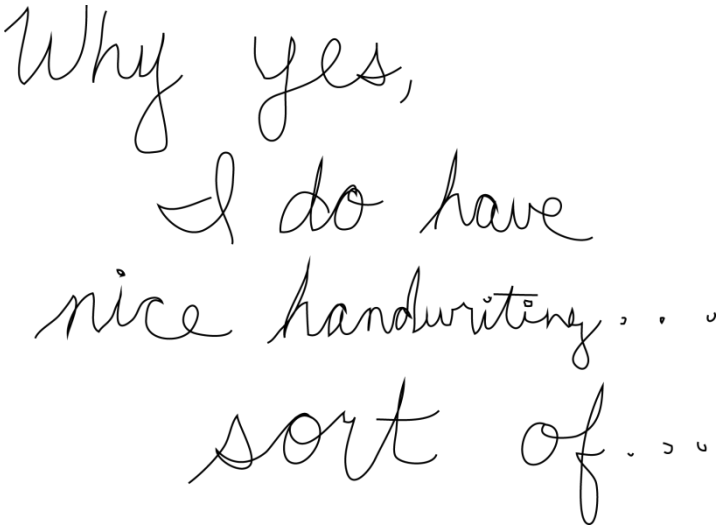


5. Adopt

Coding Style Guide

To ensure that statistical code is easily shared across a team and is replicable by future users, SDP and the Center for Education Policy Research (CEPR) recommends that you follow best coding, programming, and data management practices.

Coding Style



```
if $teacher == 1 {
local numyrs = 4
mat out = J(`numyrs',2,.)
local row = 1
local col = 1

foreach subj in math read {
use "$data/student_teacher_`subj'_vam.dta", clear
forval yr = 2(1)`numyrs' {gen late_exp_`yr' = ever_late_hire*t_exp`yr'}
}
}
```

```
if $teacher == 1 {

    local numyrs = 4

    // define empty matrix of Yr x Subj
    mat out = J(`numyrs',2,.)
    local row = 1
    local col = 1

    foreach subj in math read {
        use "$data/student_teacher_`subj'_vam.dta", clear

        forval yr = 2(1)`numyrs' {
            gen late_exp_`yr' = ever_late_hire*t_exp`yr'
        }
    } // end of loop on subject
 } // end of teacher processing
```

```
if $teacher == 1 {
local numyrs = 4
mat out = J(`numyrs',2,.)
local row = 1
local col = 1

foreach subj in math read {
use "$data/student_teacher_`subj'_vam.dta", clear
forval yr = 2(1)`numyrs' {gen late_exp_`yr' = ever_late_hire*t_exp`yr'}
}
}
```

```
if $teacher == 1 {

    local numyrs = 4

    // define empty matrix of Yr x Subj
    mat out = J(`numyrs',2,.)
    local row = 1
    local col = 1

    foreach subj in math read {
        use "$data/student_teacher_`subj'_vam.dta", clear

        forval yr = 2(1)`numyrs' {
            gen late_exp_`yr' = ever_late_hire*t_exp`yr'
        }
    } // end of loop on subject
} // end of teacher processing
```

```
/*************************************************************************
* File name:     crosswalk_masked_ids.do
* Author(s):   JSilver
* Date:        5/27/11
* Description: This program creates the crosswalk of student ids to random
*              research ids by:
*              1. Inputting the universe of student ids
*              2. Filtering the distinct set of student ids
*              3. Generating random ids and associating to student ids
*
* Inputs:      ../raw/students/studentyearsch.dta
*              ../raw/students/englang.dta
*
* Outputs:     ../data/bps_student_school_year.dta
*
* Update 1: TKawakita, 6/1/11 - Added check to ensure random ids are unique
*************************************************************************/

clear
set more off
capture log close
set mem 8000m

global raw  "//cepr-files/projects/DCPS/Raw"
global data "//cepr-files/projects/DCPS/Data"
global log  "//cepr-files/projects/DCPS/Log Files"

//******* Step 1: Input universe of student ids *******
...
//******* Step 2: Filter distinct set of student ids *******
...
//******* Step 3: Generate random ids and associate to student ids *******
...
//******* Update 1: Add check to ensure ids unique *******
...
```

```
gen t_late_hire = 0

 replace t_late_hire = 0 if t_hiredate <= td(1sep2006) & t_hiredate !=. & t_year==2007
 replace t_late_hire = 1 if t_hiredate > td(1sep2006) & t_hiredate <= td(1apr2007) ///
     & t_hiredate!=. & t_year==2007
 replace t_late_hire = 0 if t_hiredate > td(1apr2007) & t_hiredate!=. & t_year==2009

 replace t_late_hire = 0 if t_hiredate <= td(1sep2007) & t_hiredate !=. & t_year==2008
 replace t_late_hire = 1 if t_hiredate > td(1sep2007) & t_hiredate <= td(1apr2008) ///
     & t_hiredate!=. & t_year==2008
 replace t_late_hire = 0 if t_hiredate > td(1apr2008) & t_hiredate!=. & t_year==2008

 replace t_late_hire = 0 if t_hiredate <= td(1sep2008) & t_hiredate !=. & t_year==2009
 replace t_late_hire = 1 if t_hiredate > td(1sep2008) & t_hiredate <= td(1apr2009) ///
     & t_hiredate!=. & t_year==2009
 replace t_late_hire = 0 if t_hiredate > td(1apr2009) & t_hiredate!=. & t_year==2009

 replace t_late_hire = 0 if t_hiredate <= td(1sep2009) & t_hiredate !=. & t_year==2010
 replace t_late_hire = 1 if t_hiredate > td(1sep2009) & t_hiredate <= td(1apr2010) ///
     & t_hiredate!=. & t_year==2010
 replace t_late_hire = 0 if t_hiredate > td(1apr2010) & t_hiredate!=. & t_year==2010
```

27 changes

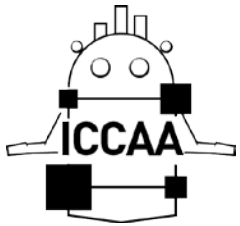
```
local num_yrs    "4"
local first_yr   "2007"
local cutoff1    "1sep"
local cutoff2    "1apr"

gen t_late_hire = 0

forval yr =`firstyr'(1)(`first_yr'+`numyrs'-1) {
    replace t_late_hire = 0 if t_hiredate <= td(`cutoff1'`yr') & t_hiredate !=. ///
       & t_year==`yr'
    replace t_late_hire = 1 if t_hiredate > td(`cutoff2'`yr') ///
       & t_hiredate<= td(`cutoff2'`yr') & t_hiredate!=. & t_year==`yr'
    replace t_late_hire = 0 if t_hiredate > td(`cutoff2'`yr') ///
       & t_hiredate!=. & t_year==`yr'
}
```

4 changes

TABLE OF CONTENTS

5. Adopt: CEPR Coding Style Guide CG HK

To ensure that statistical code is easily shared across a team and is replicable by future users, SDP and the Center for Education Policy Research (CEPR) recommends that you follow best coding, programming, and data management practices.



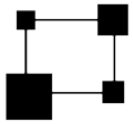
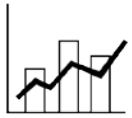
Q&A

SDP TOOLKIT

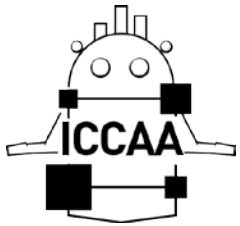
FOR EFFECTIVE DATA USE

A GUIDE FOR CONDUCTING DATA ANALYSIS IN EDUCATION AGENCIES

Will Be Released Prior to Webinar On:

I	C	C	A	A
	Thursday, February 2	Thursday, February 9	Thursday, February 16	
				
Identify: Data Specification Guide	**Clean**: Data Building Tasks	**Connect**: Data Linking Guide	**Analyze**: Diagnostic Analyses Guide	**Adopt**: Coding Style Guide

Thank You



*The toolkit is currently in **BETA**.*

Please send us your feedback at goo.gl/AAvdF.

Check www.gse.harvard.edu/sdp/tools for the most recent toolkit version.

Please contact us at sdp@gse.harvard.edu if you have any questions about the toolkit.