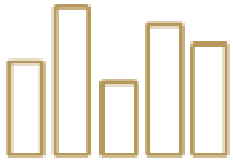


STRATEGIC DATA PROJECT

IDENTIFY: DATA SPECIFICATION GUIDE

SDP TOOLKIT

FOR EFFECTIVE DATA USE IN EDUCATION AGENCIES

www.gse.harvard.edu/sdp/toolkit

Toolkit Documents

VERSION: 1.5
Last Modified: September 3, 2013

| Authored by Todd Kawakita and the SDP Research Team

1. Identify: Data Specification Guide

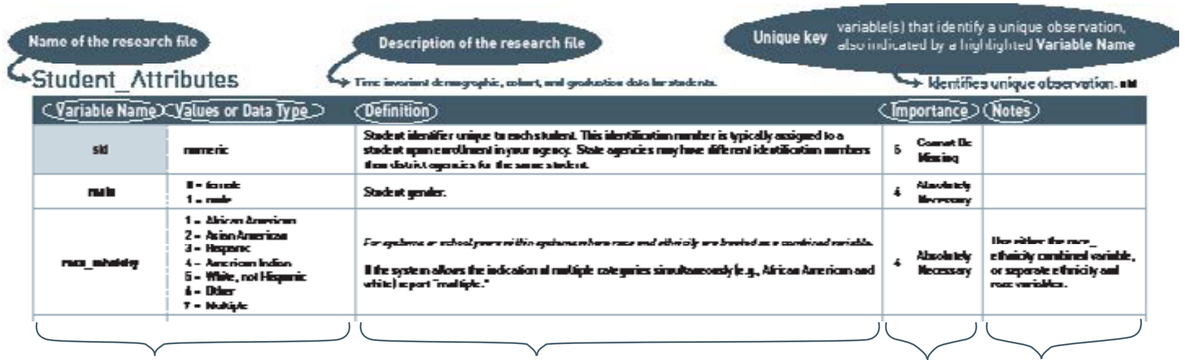
Identify essential data elements for analysis from your organization;

Identify: Data Specification Guide is a resource to identify data elements required to analyze student achievement, postsecondary attainment, and teacher effectiveness data. To address these different areas, we organize data elements into **research files** that contain important information at the student-, school-, and teacher- levels. These research files comprise the elements needed to **Clean**, **Connect**, and then **Analyze** your data.

These columns indicate files necessary to answer questions about college-going success or human capital.

		College-going	Human Capital	page
STUDENT DATA FILES				
Student_Attributes	Demographic, cohort, and graduation data for students.	■	■	7
Student_School_Year	School year and attendance data for students.	■	■	8
Student_School_Enrollment	School enrollment/withdrawal data for students.	■	■	9
Student_Class_Enrollment	Class enrollment, grades, and credits earned data for students.	■	■	10
Student_Test_Scores	Standardized test data for students (state standardized tests, advanced placement, SAT, ACT, etc). Every attempt at a test by a student should be recorded.	■	■	11
Student_NSC_Enrollment	The National Student Clearinghouse (NSC) Student Tracker student-level data report that provides information on postsecondary outcomes.	■		12
SCHOOL DATA FILES				
School	Location and classification of schools.	■	■	13
Class	Class level scheduling data.		■	14
STAFF DATA FILES				
Staff_Attributes	Demographic and recruitment data of staff.		■	15
Staff_School_Year	Pay, experience, school placement, and job codes of staff.		■	16
Staff_Degrees	Educational achievement of staff. Each degree a staff member received should be recorded once.		■	17
Staff_Certifications	Teaching certifications received by staff.		■	18

HOW TO READ A RESEARCH FILE



The **Variable Name** column indicates the name of the variable. The **Values** or **Data Type** indicate values the variables should take.

The **Definition** column provides a detailed description of each variable.

The **Notes** column contains comments on variables, including alternative sources for the data.

The **Importance** column gives the necessity of each variable. For instance, if you know a variable is of poor quality in your agency and it is marked 1-Not Essential in the importance column, you may decide to drop it from the analysis.

Variables designated as 5-Cannot Be Missing and 4-Absolutely Necessary are required for your analysis. The difference between these two is that Absolutely Necessary variables can be blank or missing values, whereas Cannot Be Missing variables cannot be. This is because 5-Cannot Be Missing variables are part of the **Unique Key**, the variables that uniquely identify an observation in the research file.

Based on the Student_Attributes research file, data you collect should look something like this:

sid	male	race_ethnicity	...
1000056	1	2	...
1000189	0	3	...
...	...	...	...

The **Variable Names** span the first row of the research file. Each variable is populated with a **Value** or **Data Type**. Also, because the **Unique Key** is **sid**, each row is uniquely identified by the student id.

THE DISTINCTION BETWEEN RAW AND CLEAN DATA

Consider Student Attributes again. In many cases, your data might be **raw**, where each row is not uniquely identified by the student id. Compare this to the **clean** data on the right.

Raw Data

sid	male	race_ethnicity	...
1000056	1	2	...
1000056	0	2	...
1000056	1	3	...
1000189	0	3	...
1000189	0	3	...
...	...	...	...

Here, a sid is listed more than once so each row is NOT uniquely identified by the student id. Student 1000056 is listed with different values for **male** and **race_ethnicity**, even though these are time-invariant variables. Student 1000189 has the same values for **male** and **race_ethnicity** but has the same row listed twice.

This data is considered **raw**.

The research file specification for Student Attributes specifies the data to look like below (this is the same data shown on the previous page):

Clean Data

sid	male	race_ethnicity	...
1000056	1	2	...
1000189	0	3	...
...	...	...	...

Here, a sid is listed only once so each row is uniquely identified by the student id. Student 1000056 has only one value for male and one value for race_ethnicity. Student 1000189 similarly has only one value for male and one value for race_ethnicity.

This data is considered **clean** and matches the specifications here in **Identify**.

The next step of the toolkit

Clean: Data Building Guide will help you get your data from a **raw** to a **clean** state.

ADVICE ON COLLECTING YOUR DATA

If you gather your agency's data to work through the toolkit (rather than using the practice files) use these "**5 W's of Data Collection**" to guide your efforts:

WHY	**WHAT**	**WHEN**	**WHERE**	**WHO**
am I collecting data? What research questions am I trying to answer?	data are needed to answer these questions?	(over what date range) do I need data for? How reliable are historical data?	are the data stored?	owns the systems where the data live and is responsible for delivery of the data?

Answering these questions will prepare you for potential issues with data collection, data reliability, and any other roadblocks to your analyses.

After you answer the questions above and are ready to assemble data, you should use this Data Specification Guide to structure your data files. The files in the specification are defined generically to accommodate necessary data elements for a wide variety of research questions regarding human capital and college going success.

THE ROLE OF DATABASE ARCHITECTS

Database Architects play an important role equipping educational analysts and data strategists with data. SDP recognizes that education agencies have pre-established systems to collect and manage student-, teacher- and human resource- databases. This toolkit enables database architects to pull data available in pre-existing warehouses and transform it for analytic work. Database architects should feel free to modify the data specification to meet their individual agency's needs. However, to support analysts and data strategists interested in pursuing the rest of the **SDP Toolkit**, following the standards set in this specification will promote efficiencies in the way variables are coded and defined.

Student_Attributes

Demographic, cohort, and graduation data for students.

Identifies unique observation: **sid**

Variable Name	Values or Data Type	Definition	Importance	Notes
sid	numeric	Student identifier unique to each student. This identification number is typically assigned to students 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	Student race and ethnicity. For systems where race and ethnicity are treated as a combined variable.	4 Absolutely Necessary	Use either the race_ ethnicity combined variable, or separate ethnicity and race variables If the system allows the indication of multiple categories simultaneously (e.g., African American and white) report "multiple"
race	1 = African American 2 = Asian American 4 = American Indian 5 = White 6 = Other 7 = Multiple	Student race. For systems or school years within systems where race and ethnicity are treated as separate variables.	4 Absolutely Necessary	Use either the race_ ethnicity combined variable, or separate ethnicity and race variables If the system allows for the indication of multiple categories simultaneously (e.g., African American and white) report "multiple"
ethnicity	0 = not Hispanic 1 = Hispanic	Student ethnicity. For 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 the student was a 9th grader for the first time. For this variable, report what the system recorded for 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 received a high school diploma from the system.	4 Absolutely Necessary	
hs_diploma_type	use local values	Any locally defined description of diploma the student received. Include instances when more than one type of diploma is awarded, (i.e. 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 the 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	

Student_School_Year

Yearly classification and attendance data for students.

Identifies unique observation: **sid** + **school_year**

Variable Name	Values or Data Type	Definition	Importance	Notes
sid	numeric	Student identifier unique to each student. This identification number is typically assigned to students upon enrollment in your agency. State agencies may have different identification numbers than district agencies for the same student.	5 Cannot Be Missing	
school_year	spring calendar year	Academic school year from fall to spring, denoted here as the spring calendar year.	5 Cannot Be Missing	
grade_level	-9 = ungraded -1 = any pre-kindergarten 0 = kindergarten 1-12 = grades 1–12 13+ = additional grade levels	Student grade level	4 Absolutely Necessary	Additional grade levels may include e.g. vocational training, special education past year 12
frpl	0 = not participating 1 = reduced lunch 2 = free lunch null = no status	Status in the free or reduced price lunch program.	4 Absolutely Necessary	
iep	0 = no IEP 1 = has IEP	Indicator for students who have an individualized education plan (IEP).	4 Absolutely Necessary	
iep_classification	use local values	Local IEP or special education classification. Generally these classifications follow the standard special education classifications.	2 Good to Have	
ell	0 = not ell 1 = ell	Indicator for students who are classified as English Language Learners (ELL). Some systems refer to this category as Limited English Proficient (LEP) or English as a Second Language (ESL).	4 Absolutely Necessary	
ell_classification	use local values	Local classification of level of English language learner status.	2 Good To Have	
gifted	0 = not enrolled in a gifted education program 1 = enrolled in a gifted education program	Indicator variable for students enrolled in gifted and talented education programs.	2 Good to Have	
gifted_classification	use local values	Local classification, if any, for gifted eligible students.	1 Not Essential	
total_days_enrolled	number of days	Total number of days over the school year a student was enrolled.	2 Good to Have	Can be calculated by school days between enrollment_date and withdrawal_date in Student_School_Enrollment for all schools, or total_days_present + total_days_absent
total_days_present	number of days	Total number of days over the school year a student was present. Cannot exceed the number of days enrolled.	2 Good to Have	Can sometimes be unreliable
total_days_absent	number of days	Total number of days over the school year a student was marked absent. Cannot exceed the number of days enrolled.	2 Good to Have	Can sometimes be unreliable
days_suspended_out_of_school	number of days	Total number of days over the school year a student experienced out of school suspension.	2 Good to Have	
days_suspended_in_school	number of days	Total number of school days during the year the student experienced in school suspension.	2 Good to Have	

Student_School_Enrollment

School enrollment/withdrawal data for students.

Identifies unique observation: **sid** + **school_year** + **school_code** + **enrollment_date**

Variable Name	Values or Data Type	Definition	Importance	Notes
sid	numeric	Student identifier unique to each student. This identification number is 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	
school_year	spring calendar year	Academic school year from fall to spring, denoted here as the spring calendar year.	5 Cannot Be Missing	
school_code	use local values	The local numeric or alpha-numeric code for the school.	4 Absolutely Necessary	
enrollment_date	date format (yyyy-mm-dd)	When the student enrolled at the school. In some systems an enrollment date is recorded when a student matriculates from a different school (e.g., moving from an elementary to middle school within the system, or moving from one middle school to a different middle school). In other systems enrollment dates are recorded at the beginning of each school year (e.g., 8th graders are assigned an enrollment date at the begining of the year even if they were enrolled at the same school the year before). The latter case is preferred. It is also what we observe in most agencies.	4 Absolutely Necessary	
withdrawal_date	date format (yyyy-mm-dd)	The date the student withdrew from the school. In some systems a withdrawal date is recorded when the student moves to a different school (e.g., moving from elementary to middle school within the system, or moving from one middle school to a different middle school). In other systems withdrawal dates are recorded at the end of each school year (e.g., 8th graders are assigned a withdrawal date at the end of the school year even if they plan to attend the same school next year). The latter case is preferred.	4 Absolutely Necessary	
enrollment_code	use local values	The local numeric or alpha-numeric code describing enrollment reason into the school, if available.	1 Not Essential	
enrollment_code_desc	text	Description of the enrollment_code.	1 Not Essential	
withdrawal_code	use local values	The local numeric or alpha-numeric code that describes withdrawal reason from the school.	4 Absolutely Necessary	
withdrawal_code_desc	text	Description of the withdrawal_code.	4 Absolutely Necessary	
days_enrolled	number of days	Number of school days during the school year the student was enrolled at a given school. The system's data sources may report this directly, or you may calculate it based on enrollment data.	2 Good to Have	Can be calculated by days_present + days_absent
days_present	number of days	Number of school days during the school year the student was present at a given school. The system's data sources may report this directly, or you may calculate it based on enrollment data.	2 Good to Have	
days_absent	number of days	Number of school days during the school year the student was marked absent at a given school. The system's data sources may report this directly, or you may calculate it based on enrollment data.	2 Good to Have	

Student_Class_Enrollment

Class enrollment, grades, and credits earned data for students.

Identifies unique observation: **sid + cid + enrollment_date**

Variable Name	Values or Data Type	Definition	Importance	Notes
sid	numeric	Student identifier unique to each student. This identification number is 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	
cid	value from Class table, page 21	Variable that links students to teachers by grouping students in the same room at the same time. One unique value should be assigned to each combination of variables: school_year + school_code + course_code + section_code + period_bell + room_number + tid.	5 Cannot Be Missing	
class_enrollment_ date	date format (yyyy-mm-dd)	The date the student enrolled in the class. In some cases an enrollment date may not be explicitly recorded in the system's data. Even if it is not recoreded, the enrollment date can often be derived.	2 Good to Have	Can be substituted with days enrolled in a course
class_withdrawal_ date	date format (yyyy-mm-dd)	The date the student withdrew from class. In some cases a withdrawal date may not be recorded in the system's data. Even if it is not explicitly recorded, the withdrawal date can often be derived.	2 Good to Have	Can be substituted with days enrolled in a course
credits_earned	use local values	The number of credits the student earned for the course.	3 Necessary for Multiple Analyses	Can be calculated by using creditspossible and creditattainment rules
final_grade_mark	use local values	The final grade or mark the student received in the class. ("final" means last, cumulative grade assigned). Grades can range from "Alpha Plus" (A+ through F), or a numeric scale (0.0 - 4.0 or 0-100).	4 Necessary for Multiple Analyses	

Student_Test_Scores

Standardized test data for students (state standardized tests, advanced placement, SAT, ACT, etc). Every attempt at a test by a student should be recorded.

Identifies unique observation: **sid + test_code + test_date**

Variable Name	Values or Data Type	Definition	Importance	Notes
sid	numeric	Student identifier unique to each student. This identification number is assigned to a student upon enrollment. State agencies may have different identification numbers than district agencies for the same student.	5 Cannot Be Missing	
test_code	use local values	These values identify individual tests, (usually expressed as a sequence of letters and numbers). For example, a state test such as "MCAS 6th Grade Math in Massachusetts" or college entrance exam such as "SAT Math".	4 Absolutely Necessary	Can be a concatenation of component variables below (e.g. test_type and test_subject)
test_date	date format (yyyy-mm-dd)	The exact date (or at a minimum school year) the test was completed. Note that students who re-take tests or are retained may have multiple observations for a single test_code; these should be differentiated by test_date.	4 Absolutely Necessary	
test_code_desc	text	Description of test_code.	4 Absolutely Necessary	
test_type	use local values	The category of test, e.g. MCAS (state test), SAT, ACT, or AP.	4 Absolutely Necessary	
grade_level	-9 = ungraded -1 = any pre-kindergarten 0 = kindergarten 1-12 = grades 1–12 13+ = additional grade levels (i.e. vocational training, special education past year 12)	Numeric grade level of the test. May be unavailable for SAT, ACT, or AP tests.	2 Good to Have	Can be pulled from Student_ School_Year if not available here.
test_subject	1 = math 2 = English language 3 = science 4 = social studies 5 = other 6 = writing	Subject of test.	4 Absolutely Necessary	Subjects other than those listed can be numbered from 7 onwards.
test_version		Test version for different standardized exams in the same agency (e.g. SAT9 vs. STAR).	2 Good to Have	
language_version	E = English S = Spanish	Language of test.	2 Good to Have	
raw_score	numeric	Student's raw score if available. May be unavailable for SAT, ACT, or AP tests.	4 Absolutely Necessary	Any test score should be acquired, including available percentile ranks
scaled_score	numeric	Student's scaled score.	4 Absolutely Necessary	Any test score should be acquired, including available percentile ranks
performance_level	code	Student's performance level (e.g., not proficient, proficient, advanced). May be unavailable for SAT, ACT, or AP tests.	2 Good to Have	
standardized_score	numeric	If the system or state provides a standardized score (i.e., mean zero, s.d. one) for state tests include it, and note any information on what distribution was used for standardization. If the system or state does not provide a standardized score for state tests, leave this blank.	1 Not Essential	Can be calculated using scaled_score with test_code, test_date, and grade_level

Student_NSC_Enrollment

National Student Clearinghouse Student Tracker student-level data that provides information on postsecondary outcomes.

Identifies unique observation: **sid** + **college_code_branch** + **enrollment_begin** + **enrollment_end**

Variable Name	Values or Data Type	Definition	Importance	Notes
The list of variables included in this file changes as the National Student Clearinghouse (NSC) updates its methodology. Do NOT use the below to request data, rather this represents the layout of the data you may receive from the NSC (as of December 2011).				
sid	numeric	Student identifier unique to each student. This identification number is assigned to a student upon enrollment. State agencies may have different identification numbers than district agencies for the same student.	5 Cannot Be Missing	
college_code_branch	use local values	OPE/FICE code of the college that the student attended. This is usually a six-digit college code followed by a hyphen and a two-digit branch code.	2 Good to Have	
enrollment_begin	date format (yyyy-mm-dd)	Start of enrollment.	4 Absolutely Necessary	
enrollment_end	date format (yyyy-mm-dd)	End of enrollment.	4 Absolutely Necessary	
record_found_yn	Y = yes N = no	Whether or not the NSC has college enrollment data for the student.	4 Absolutely Necessary	
high_school_code	use local values	Local code for student's high school.	1 Not Essential	
high_school_grad_dt	date format (yyyy-mm-dd)	Student's high school graduation date; if graduated.	1 Not Essential	
college_name	text	The name of the college.	2 Good to Have	
college_state	two letter state abbreviation	The state where the college is located.	2 Good to Have	
year4year	4 = 4-year 2 = 2-year L = less than 2-year	Length of degree program.	4 Absolutely Necessary	"L" and "2" are typically grouped together
public_private	public / private	Type of college:public or private.	2 Good to Have	
enrollment_status	A = leave of absence D = deceased F = full-time H = half-time L = less than half-time W = withdrawn	Student's college enrollment status.	3 Necessary for Multiple Analyses	
graduated	Y = yes N = no	Whether or not student has graduated college.	2 Good to Have	
graduation_date	date format (yyyy-mm-dd)	Student's college graduation date.	2 Good to Have	
college_sequence	1, 2, 3,...	Sequence that student progresses through college.	1 Not Essential	
degree_title	use local values	Degree title.	1 Not Essential	
major	use local values	Student's major.	1 Not Essential	

School

Yearly location and classification information for schools.

Identifies unique observation: **school_code** + **school_year**

Variable Name	Values or Data Type	Definition	Importance	Notes
school_code	use local codes	The local numeric or alpha-numeric code for the school.	4 Absolutely Necessary	
school_year	spring calendar year	Academic school year from fall to spring, denoted here as the spring calendar year.	4 Absolutely Necessary	
school_name	text	Name of school.	4 Absolutely Necessary	
pid	use local values	Principal identifier unique to each school. This identification number is typically assigned to a Principal upon entrance to agency. State agencies may have different identification numbers than district agencies for the same Principal.	1 Not Essential	
campus_code	use local values	Identifies the campus where the school is located, if co-located with other schools. For example, when smaller schools-within-schools are established at large traditional high school.	1 Not Essential	
local[district or cluster or etc]	use local values	One or more variables. The administrative sub-unit(s) to which the school is assigned.	1 Not Essential	
grade_span	K–5 K–8 6–8 9–12 [or as appropriate]	The span of grade levels served by the school. Note this variable should reflect the intended grade span, and not the actual grades observed.	1 Not Essential	A school with mostly K–5 students and a handful of 6th graders should not be recorded as K–6
elementary	0 = not elementary school 1 = elementary school	Indicator variable identifying schools that serve grades K–5.	2 Good to Have	
middle	0 = not middle school 1 = middle school	Indicator variable identifying schools that serve grades 6–8.	2 Good to Have	
high	0 = not high school 1 = high school	Indicator variable identifying schools that serve grades 9–12.	2 Good to Have	
charter	0 = not charter school 1 = charter school	Indicator variable identifying schools which are charter schools.	4 Absolutely Necessary	
alternative	0 = not alternative school 1 = alternative school	Indicator variable identifying schools which are alternative schools.	4 Absolutely Necessary	
sped	0 = not sped 1 = sped	Indicator variable identifying schools which are special education schools (i.e. school for the blind).	4 Absolutely Necessary	
frpl	use local codes	Percent of students in a school with a status of free or reduced price lunch, calculated by the agency.	2 Good to Have	May be used to group schools by poverty status

Class

Class level scheduling data.

Identifies unique observation: **cid**

Variable Name	Values or Data Type	Definition	Importance	Notes
cid	numeric or string	Variable that links students to teachers by grouping students in the same room at the same time. One unique value should be assigned to each combination of variables: school_year + school_code + course_code + section_code + period_bell + room_number + tid.	5 Cannot Be Missing	Constructed by the anaylst to link students to teachers
school_year	spring calendar year	Academic school year from fall to spring, denoted here as the spring calendar year.	5 Cannot Be Missing	
school_code	use local values	The local numeric or alpha-numeric code for the school.	4 Absolutely Necessary	
course_code	use local values	The local numeric or alpha-numeric code for the course. A definition of how to interpret the variable or variables is important to capture in the codebook when the description of the code's sub-elements is not sufficient.	4 Absolutely Necessary	
course_code_desc	text	Description of the course_code. For example, "Pre-Algebra" or "English 7".	4 Absolutely Necessary	
section_code	use local values	The local numeric or alpha-numeric code that identifies individual sections.	4 Absolutely Necessary	These variables determine the "class", the exact set of students in the same room at the same time. Some or all of these variables may be necessary for this purpose
period_bell	use local values	The local numeric or alpha-numeric code for the period when the class meets.	2 Good to Have	
room_number	use local values	The local numeric or alpha-numeric code for the classroom where class meets.	2 Good to Have	
tid	numeric	Unique teacher or student identifier. State agencies may have different identification numbers than district agencies for the same staff/teacher.	4 Absolutely Necessary	Staff includes teachers and other agency employees
semester_term_year	100 = year long course 010 = first semester course 020 = second semester course 001 = first term course 002 = second term course 004 = fourth term course	The duration of the class and when the class occurred during the school year. (i.e., year, semester, or term).	2 Good to Have	
graduation_requirement	0 = not required for high school graduation 1 = required for high school graduation	Indicator variable for all high school level courses required for graduation.	3 Necessary for Multiple Analyses	
credits_possible	use local values	Credits possible from the course.	1 Not Essential	
instructional_level	1 = remedial 2 = standard 3 = honors/advanced	Instructional level of the course.	2 Good to Have	Sometimes embedded in course code or course title
subject	1 = math 2 = English language arts 3 = science 4 = social studies 5 = other	General subject matter of the course.	3 Necessary for Multiple Analyses	Sometimes embedded in course code or course title

Staff_Attributes

Time invariant demographic and recruitment data related to staff.

Identifies unique observation: **tid**

Variable Name	Values or Data Type	Definition	Importance	Notes
tid	numeric	Unique staff or teacher identifier. State agencies may have different identification numbers than district agencies for the same staff/teacher.	5 Cannot Be Missing	Staff includes teachers but also other employees in the agency
male	0 = female 1 = male	Staff gender.	2 Good to Have	
race_ethnicity	1 = African American 2 = Asian American 3 = Hispanic 4 = American Indian 5 = White, not Hispanic 6 = Other 7 = Multiple	For systems where race and ethnicity are treated as a combined variable. If the system allows multiple categories (e.g., African American and white) report "multiple."	2 Good to Have	Use either the race_ethnicity combined variable, or separate ethnicity and race variables
race	1 = African American, not Hispanic 2 = Asian American 4 = American Indian 5 = White 6 = Other 7 = Multiple	For systems where race and ethnicity are treated as separate variables. If the system allows for multiple categories (e.g., African American and white) report "multiple."	2 Good to Have	Use either the race_ethnicity combined variable, or separate ethnicity and race variables
ethnicity	0 = not Hispanic 1 = Hispanic	For systems where race and ethnicity are separate and Hispanic or Latino origin is asked separately.	2 Good to Have	Use either the race_ethnicity combined variable, or separate ethnicity and race variables
birth_date	date format (yyyy-mm-dd)	Staff birth date.	1 Not Essential	
zip_code	xxxxx or xxxxx-yyyy	Zip code of the staff member's home address.	1 Not Essential	
offer_date_first	date format (yyyy-mm-dd)	The date the staff member was offered a job.	2 Good to Have	
offer_date_most_recent	date format (yyyy-mm-dd)	If the staff member left and was re-hired, the most recent date they were offered a job to work in the system.	2 Good to Have	
hire_date_first	date format (yyyy-mm-dd)	The first date the staff member was hired to work.	2 Good to Have	Can be substituted with start date or first paycheck. Used to determine late hires
hire_date_most_recent	date format (yyyy-mm-dd)	If the staff member left and was re-hired, the most recent date they were hired to work in the system.	2 Good to Have	
termination_date_first	date format (yyyy-mm-dd)	The first date staff members terminated employment.	1 Not Essential	
termination_date_most_recent	date format (yyyy-mm-dd)	If the staff member left, was re-hired and then left again, the most recent date they left the system.	1 Not Essential	
certification_path	use local values	The teacher's certification pathway. For example, "university/college" , "alternative", or "uncertified."	2 Good to Have	Where data exists, TFA or NYC Teaching Fellows can be included

Staff_School_Year

Yearly pay, experience, school placement, and job codes for staff.

Identifies unique observation: **tid** + **school_year**

Variable Name	Values or Data Type	Definition	Importance	Notes
tid	numeric	Unique staff or teacher identifier. State agencies may have different identification numbers than district agencies for the same staff/teacher.	5 Cannot Be Missing	Staff includes teachers but also other employees in the agency
school_year	spring calendar year	Academic school year from fall to spring, denoted here as the spring calendar year.	5 Cannot Be Missing	
salary_group_code	use local values	Separate variables that describe different components of salary. For example: salary_group_degree, salary_group_experience, salary_group_nbct, etc. Identifies the teacher's base salary, and generally includes a component for experience, degrees earned, and other considerations (e.g., NBPTS certification). A definition of how to interpret the variable or variables is particularly important to capture in the codebook.	Only necessary if we do not know teaching experience and advanced degree	
additional_pay	amount	The amount of salary above the base salary (captured in salary_group_code) which the staff member received.	1 Not Essential	
teaching_ experience	0 or positive integer	Identifies the total teaching experience of the staff member in years, (0 denotes a new teacher).	4 Absolutely Necessary	Can be backed out from salary. Note: we only need to differentiate between the first 1-5 years. It is okay (but not ideal) to group more experienced teachers together
district_experience	0 or positive integer	Identifies teaching experience of the staff member as an employee of the district.	2 Good to Have	
school_code1	use local values	The local code for the primary school in which a staff member is employed.	4 Absolutely Necessary	Can be determined from course file
job_code1	use local values	The local code for the job or position held in the primary school.	2 Good to Have	
school_code2	use local values	The local code for the secondary school, if more than one.	1 Not Essential	If a teacher splits time between schools and has a different job_code for each
job_code2	use local values	The local code for the job or position held in secondary school.	1 Not Essential	
school_code#	use local values	The local code for the #th school, if more than two.	1 Not Essential	
job_code#	use local values	The local code for the job or position held in the #th school, if more than two.	1 Not Essential	

Staff_Degrees

Educational achievement for staff. Each degree a staff member has received should be recorded once.

Identifies unique observation: **tid** + **degree_type** + **degree_institution_code** + **degree_date**

Variable Name	Values or Data Type	Definition	Importance	Notes
tid	numeric	Staff/teacher identifier unique to each staff/teacher. This identification number is typically assigned to a staff/teacher upon entrance to your agency. State agencies may have different identification numbers than district agencies for the same staff/teacher.	5 Cannot Be Missing	
degree_type	1 = less than four year degree 2 = four year degree 3 = master's degree 4 = doctoral degree	The type of degree received by the staff member.	3 Necessary for Multiple Analyses	Can be backed out from salary grade.
degree_institution_ code	use local codes	Local code for the degree granting institution.	2 Good to Have	
degree_date	date format (yyyy-mm-dd)	The date (or at minimum year) on which the degree was granted.	1 Not Essential	
degree_opeid	OPEIDs	OPEID for the degree granting institution. Identification number used by the U.S. Department of Education's Office of Postsecondary Education (OPE) to identify schools.	2 Good to Have	Can be obtained from a simple match if degree_ institution name is known.
degree_major	use local codes	Description of the degree major or subject.	2 Good to Have	

Staff_Certifications

Teaching certifications received by staff.

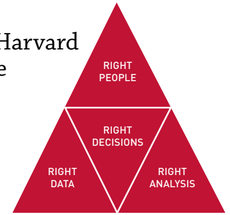
Identifies unique observation: **tid** + **certification_code** + **certification_effective_date**

Variable Name	Values or Data Type	Definition	Importance	Notes
tid	numeric	Unique staff or teacher identifier. State agencies may have different identification numbers than district agencies for the same staff/teacher.	5 Cannot Be Missing	Staff includes teachers but also other employees in the agency
certification_code	use local codes	One or more variables (e.g., grade level may be separate from subject area). Local code for the type or class of certification or license.	2 Good to Have	
certification_ effective_date	date format (yyyy-mm-dd)	The date the certification took effect.	2 Good to Have	School year can also be used
certification_code_ desc	text	Descriptive text for certification_code.	2 Good to Have	
certification_ expire_date	date format (yyyy-mm-dd)	The date when certification expired or expires.	2 Good to Have	School year can also be used
special_education	0 = not special education certified 1 = special education certified	An indicator variable for staff with Special Education certification.	1 Not Essential	
english_language_ learners	0 = not English language learners certified 1 = English language learners certified	An indicator variable for staff with English language learners certification.	1 Not Essential	
nbct	0 = not National Board certified 1 = National Board certified	An indicator variable that are National Board certified.	2 Good to Have	

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

23 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 USE

SDP'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