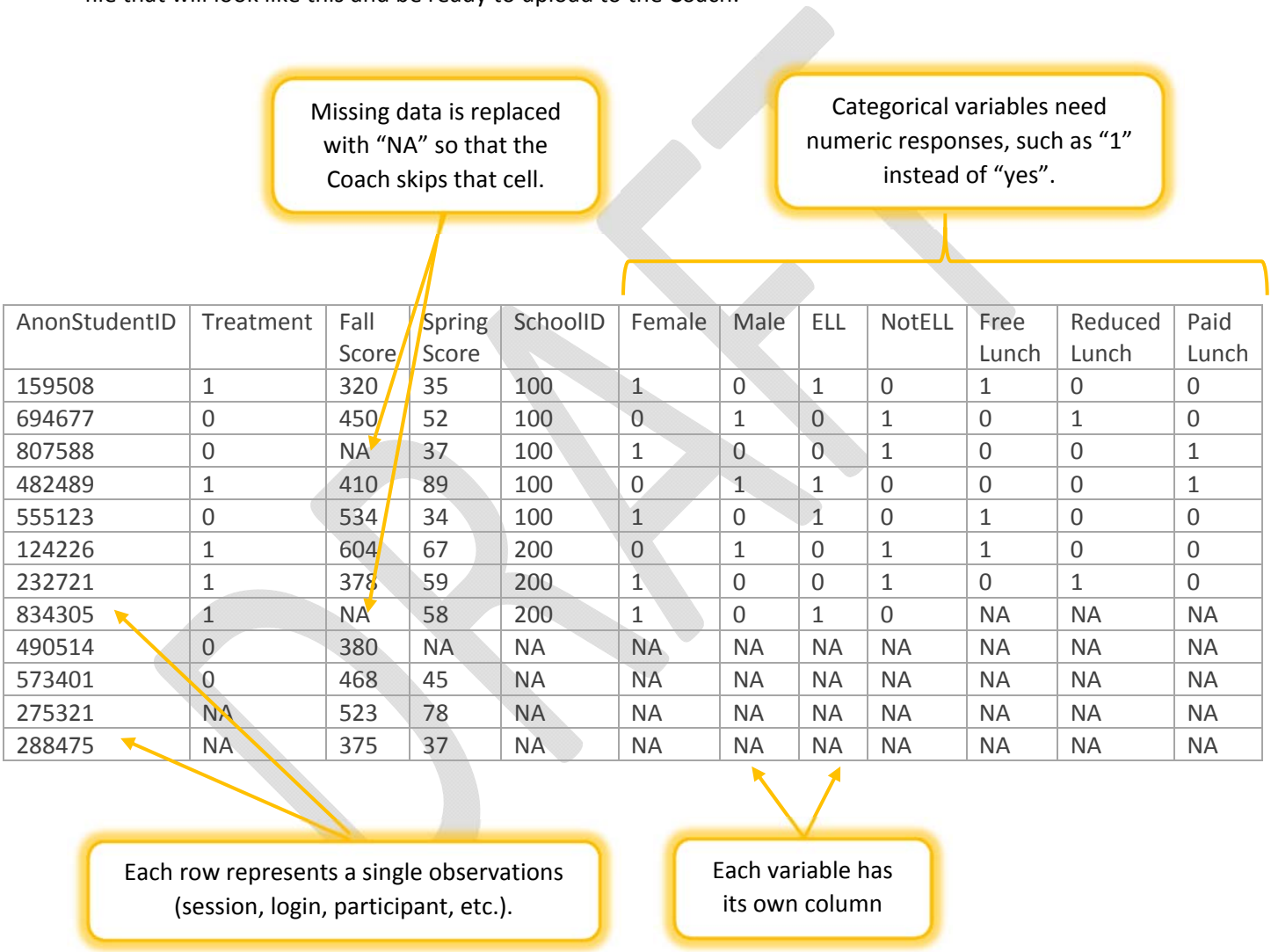
Prepare Your Data for Analysis

INTRODUCTION

To answer a research question you will need data from multiple sources. If you need guidance on processing your data, this document will take you through best practices. At the end you will have a data file that will look like this and be ready to upload to the Coach:



AnonStudentID	Treatment	Fall Score	Spring Score	SchoolID	Female	Male	ELL	NotELL	Free Lunch	Reduced Lunch	Paid Lunch
159508	1	320	35	100	1	0	1	0	1	0	0
694677	0	450	52	100	0	1	0	1	0	1	0
807588	0	NA	37	100	1	0	0	1	0	0	1
482489	1	410	89	100	0	1	1	0	0	0	1
555123	0	534	34	100	1	0	1	0	1	0	0
124226	1	604	67	200	0	1	0	1	1	0	0
232721	1	378	59	200	1	0	0	1	0	1	0
834305	1	NA	58	200	1	0	1	0	NA	NA	NA
490514	0	380	NA	NA	NA	NA	NA	NA	NA	NA	NA
573401	0	468	45	NA	NA	NA	NA	NA	NA	NA	NA
275321	NA	523	78	NA	NA	NA	NA	NA	NA	NA	NA
288475	NA	375	37	NA	NA	NA	NA	NA	NA	NA	NA

NOTE: You can work with data in a number of different programs. If you do not have access to statistical software, you can use Microsoft Excel to prepare your data. We have included some tips throughout this guide that will help you manage your data in Excel with ease.

STEP 1. IDENTIFYING DATA SOURCES

There are multiple types of data you will use in your analysis. Below is a checklist and description of the data you should have. There are examples at the end of this guide showing what each dataset should look like.

- ☐ **Outcome data (required):** This is data on the outcome you are using to determine the effect of your educational technology, such as test scores.

- ☐ **Treatment Status (required):** This is the data set that indicates which participants were given access to the educational technology and which were not. Typically, a 1 in this column indicates access to the technology (treatment or intervention group) and a 0 indicates no access to the technology (control or comparison group).

- ☐ **Pre-Intervention Assessment data (required for certain designs):** Assessment data from before the intervention (such as an assessment from the beginning of the school year) is incredibly important. It is necessary to create similar groups if you are not using randomization and can also be useful to control for pre-intervention abilities in a randomized pilot.

- ☐ **Background Characteristics:** Background characteristics provide data on observable traits for each participant. These could include gender, ethnicity, IEP status, EL status, free and reduced price lunch status and more. Background traits are particularly important for matched comparison evaluations because you can use them to make sure your intervention and comparison groups are well matched (balanced).

  CAUTION: Some of these characteristics, like EL status, can change over time. It is important to know if these characteristics were recorded before or after the introduction of the educational technology because some technologies may affect them (e.g., the introduction of a reading software). *It is preferable to measure all background characteristics before the introduction of the educational technology to avoid confounding effects.*

- ☐ **Usage Data:** Sometimes you want to determine how much of an intervention someone actually received. This is the "dosage." You can collect this data by recording how much time someone actually used an educational technology for. This can tell you things about the integrity of implementation as well as about the relationship between time using the technology and performance on the outcome you're interested in. This often comes directly from the technology vendor.

  CONSIDER: It is important that all of your data should be recorded using **individual identifiers**. These are unique codes for each participant. The identifiers are what will allow you to combine (merge) data sets. These could be a Student ID Number, School ID Number, Teacher ID number, etc.

STEP 2. PROCESSING YOUR DATA

Once you have identified the data elements and data sources, the second step is to combine all data elements into one *tidy* dataset and prepare the variables that will be used for analysis. (We explain below how to do this.) We recommend generating tidy datasets not only because it is a requirement to use the RCE Coach and most statistical software packages for analyses, but because a tidy dataset is easy to manipulate, model, and visualize.[1] The dataset at the beginning of this guide is an example of a tidy dataset. This section will take you through a series of questions that will help you create your own.

Having **tidy data** means that you've used a standardized way to structure your dataset. Specifically:

1. Each variable forms a column.
2. Each observation forms a row.
3. Each type of observational unit forms a table.

A. Is each observation a row and each variable a column?

Example:

AnonStudentID	Treatment	Fall Score	Spring Score	Gender
159508	1	320	35	F
694677	0	450	52	M
807588	0	999	37	F
482489	1	410	89	M

NO: Re-organize your dataset so that each row represents one observation. Each variable you are interested in should be its own column.

YES: Continue on to B.

B. Do you have one dataset that contains all of the variables you will need for analysis?

NO: You will need to merge your existing datasets into one complete dataset. This will be easy to do using the unique identifiers. If you are using Excel to manage your data you can do this using a VLOOKUP function.

CAUTION: Some observations may be present in some datasets but not in others. Therefore, when merging these datasets you may introduce some missing data. For example, a student in a dataset consisting of test scores may not exist in another dataset and therefore could have missing data for other variables (such as treatment status and background characteristics) after combining the datasets.

YES: Continue on to C.

[1] If you want to learn more about Tidy Data you can refer to: Wickham, H. (2014). Tidy Data. Journal of Statistical Software, 59(10), 1 - 23. doi:http://dx.doi.org/10.18637/jss.v059.i10

C. Are all of the categorical variables that will be used for your analysis numeric?

> **NOTE:** If you are using a statistical software package you will want to make sure that all of these variables are recognized as "numeric" values and not "string" or "character."

NO: Convert all of your categorical variables, or variables that include names or labels, into numbers. This might mean you have to change a variable into a dummy variable. A dummy variable uses 1 to indicate "yes" or that a condition was met and 0 to indicate "no" or that a condition was not met.

For example, if your background characteristics include gender as a variable, you may have "male" or "female," or "M" or "F" in each cell of that column. Instead, you should change the variable from 'Gender' to 'Female', and change each cell that indicates the participant is a female to 1, and each cell that indicates the participant is male to 0. You can do this for every variable that is non-numeric.

Example:

Student ID	Gender
159508	F
694677	M
807588	F

→

Student ID	Female
159508	1
694677	0
807588	1

> **NOTE:** If your categorical variable contains more than two options, such as "Free or Reduced Price Lunch" (where the options are (1) free, (2) reduced, and (3) full price) you will need to create a dummy variable for each option. For example, Free would be one column (with 0 representing that they don't receive a free lunch and 1 representing that they do), Reduced price would be a second column, and Paid (meaning they pay full price) would be a third column. (See Example Dataset 3b below)

YES: Continue on to D.

D. Is all missing data coded consistently in your dataset?

NO: If you have merged datasets and there is missing data, make sure you are consistently coding that as "NA" to ensure that the Coach can analyze your data. You want to be extra careful to make sure missing data has not been given a number designation, like 0 or 999. These values will get incorporated into the analysis. If you are using your own statistical software, such as SAS or Stata, it will be helpful to code the missing data as a period ".".

YES: Congratulations! You have a tidy dataset!

STEP 3. CHECKING THE QUALITY OF YOUR DATA

After constructing your data file and converting your variables, the final step is to check the quality of your data. The following checks can be run to identify potential data issues that warrant additional investigation:

- **Check the minimum and maximum values of variables.** This check may help to identify extremely low or high values that are outliers in your distribution or may signal a special missing code that needs to be converted to a missing value. You may want to check with someone who is familiar with the data to confirm the value range makes sense.

  If you are working in Excel you can use MIN and MAX functions to easily find these values.

- **Consider the impact of missing data.** The dashboards, and some statistical software packages, will automatically drop observations that contain missing data. You should try to understand why data is missing and how excluding students with incomplete data will affect your results.

  If you are working in Excel you can sort and filter your data to view missing values. To determine exactly how many values are missing for a single variable you can use the COUNTIF function, and to determine how many observations have at least one missing value you can use a nested COUNTIF with OR function.

EXAMPLE DATA SETS

Pre- and post-test scores do not need to come from the same test, because they will not be compared directly to one another.

Dataset 1: *Test Scores*

AnonStudentID	Fall Score	Spring Score
159508	320	35
694677	450	52
807588	NA	37
482489	410	89
555123	534	34
124226	604	67
232721	378	59
834305	NA	58
490514	380	NA
573401	468	45
275321	523	78
288475	375	37

These are missing values.

Dataset 2: *Treatment Status*

AnonStudentID	Treatment
159508	1
694677	0
807588	0
482489	1
555123	0
124226	1
232721	1
834305	1
490514	0
573401	0

Dataset 3a: *Background Characteristics (with non-numeric categorical variables)*

AnonStudentID	SchoolID	Gender	ELL_Status	Free_Reduced_Lunch
159508	100	F	ELL	Free
694677	100	M	Not ELL	Reduced
807588	100	F	Not ELL	Paid
482489	100	M	ELL	Paid
555123	100	F	ELL	Free
124226	200	M	Not ELL	Free
232721	200	F	Not ELL	Reduced
834305	200	F	ELL	Missing

Dataset 3b: *Background Characteristics (with numeric categorical variables)*

AnonStudentID	SchoolID	Female	ELL	Free	Reduced	Paid
159508	100	1	1	1	0	0
694677	100	0	0	0	1	0
807588	100	1	0	0	0	1
482489	100	0	1	0	0	1
555123	100	1	1	1	0	0
124226	200	0	0	1	0	0
232721	200	1	0	0	1	0
834305	200	1	1	NA	NA	NA