Prepare Your Data for Analysis: Randomized Pilot

INTRODUCTION

To answer your 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:

Missing data is replaced with "NA" so that the Coach skips that cell.

Categorical variables must be converted to binary variables with two possible values: "1" instead of "yes" and "0" instead of "no."

AnonStudentID	SchoolID	Grade	Treatment	Fall_ Score	Spring_ Score	Female	EL	Low_SES	Black	White	Asian	Other
159508	100	3	1	320	35	1	1	1	1	0	0	0
694677	100	3	0	450	52	0	0	0	0	1	0	0
807588	100	4	0	NA	37	1	0	0	1	0	0	0
482489	100	4	1	410	89	0	1	0	0	0	1	0
555123	100	4	0	534	34	1	1	1	0	0	1	0
124226	200	3	1	604	67	0	0	0	1	0	0	0
232721	200	3	1	378	59	1	0	1	0	0	0	1
834305	200	4	1	NA	58	1	1	NA	NA	NA	NA	NA
490514	300	3	0	380	NA	NA	NA	NA	NA	NA	NA	NA
573401	300	3	0	468	45	NA	NA	NA	NA	NA	NA	NA
275321	300	4	1	523	78	NA	NA	NA	NA	NA	NA	NA
288475	300	3	0	375	37	NA	NA	NA	NA	NA	NA	NA

Each row represents a single observation (student, teacher, school, etc.).

Each variable has its own column

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.

- ☐ **Treatment Status (required):** This variable is in the dataset that you downloaded from the Coach after your technology users and your comparison group were randomly selected. It 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 (comparison group).

  CAUTION: You must use the original file with the original treatment status variable, even if someone in your comparison group found a way to use the technology. The Coach will calculate the effect of being randomly assigned to use the technology on your outcome of interest. If you move participants from one group to another after randomization, your groups may be unbalanced, and you will not be able to make reliable conclusions about the technology's effect.

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

- ☐ **Pre-test data (recommended):** This is a measure of the outcome prior to the introduction of the technology (such as an assessment from the beginning of the school year). Pre-test data are 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 (recommended):** Background characteristics provide data on observable traits for each participant. These could include grade level, gender, ethnicity, Individualized Education Program (IEP) status, English Learner (EL) status, socioeconomic status and more. It is necessary to include background characteristics in your analysis if a lot of participants drop out of one group. (For instance, if all students remain in the technology user group but 25 percent drop out of the comparison group, you might be suspect that the two groups are now different). Accounting for background characteristics can rebalance the two groups. However, the Coach recommends that you include some background characteristics in your analysis even if your evaluation went as planned.

  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.*

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.

CONSIDER: It is important that all of your data is recorded using **individual identifiers**. These are unique codes for each participant. The identifiers are what will allow you to combine (merge) datasets. These could be a Student ID Number, School ID Number, Teacher ID number, etc. If you have students in multiple classrooms or teachers in multiple schools, you should include an ID for both the individual (i.e. student or teacher) and group (i.e. class or school).

[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

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.

NOTE: The dataset that you downloaded from the Coach after random assignment may look different than the dataset you now need for your analysis. The analysis dataset should contain one row for each individual or group for which you measured outcomes.

For example, if you randomly assigned students to use the technology, and you want to look at student test score outcomes, you probably need to add only the posttest scores to the original dataset of students. However, if you randomly assigned teachers/classes to use the technology, and you want to look at student test score outcomes, your original dataset had one row per teacher, but this dataset should have one row per student. You must include both student and teacher/class IDs. Also, student characteristics should be individual to each student, not the average of each class.

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.

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 binary or 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 "Race" (where the options are (1) Asian, (2) Black, (3) White, and (4) Other) you will need to create a binary or dummy variable for each option. For example, "Asian" would be one column (with 0 representing non-Asians and 1 representing Asians), "Black" would be a second column, "White" would be a third column, and "Other" would be a fourth column. (See Example Dataset 3b below.)

YES: Continue on to D.

D. Is all missing data coded as "NA" 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.

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/classrooms/schools 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 DATASETS

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

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

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	EL Status	SES	Race
159508	100	F	EL	Low	Black
694677	100	M	Not EL	Medium	White
807588	100	F	Not EL	High	Black
482489	100	M	EL	High	Asian
555123	100	F	EL	Low	Asian
124226	200	M	Not EL	Medium	Black
232721	200	F	Not EL	Low	Other
834305	200	F	EL	Missing	Missing

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

AnonStudentID	SchoolID	Female	EL	Low_SES	Black	White	Asian	Other
159508	100	1	1	1	1	0	0	0
694677	100	0	0	0	0	1	0	0
807588	100	1	0	0	1	0	0	0
482489	100	0	1	0	0	0	1	0
555123	100	1	1	1	0	0	1	0
124226	200	0	0	0	1	0	0	0
232721	200	1	0	1	0	0	0	1
834305	200	1	1	NA	NA	NA	NA	NA

The Coach requires that all categorical variables are converted to binary or dummy variables with values of 0 and 1. In the case of SES, you may group medium and high SES together (1=low and 0=medium/high).

NOTE: Research shows that students of higher SES often have an academic advantage over students of lower SES because of differences in early education access, home enrichment, levels of stress, food access, and many other factors. Therefore, if possible, it is good to control for SES when analyzing your data. However, some schools or districts may not have access to SES measures or may not be able to use them in this evaluation because of privacy concerns.