# Prepare Your Data for Analysis: Matched Comparison Design

## 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:

### **Exhibit 1.** Summary table

Missing data are replaced with NA so that the Coach doesn't include a value for that observation

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 |
| 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 |

EL = English learner; SES = socioeconomic status.

Each row represents a single observation (student, teacher, school, and so on).

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.

# Ed Tech Rapid Cycle Evaluation Coach

## STEP 1. IDENTIFYING DATA SOURCES

You will use multiple types of data in your analysis. A checklist and description of the data you should have follow. At the end of this guide are examples showing what each data set should look like.

**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 enable you to combine (merge) data sets. These could be a student ID number, school ID number, teacher ID number, and so on.

- **Outcome data (required).** These are 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 received access to the educational technology and which did 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).
- **Pre-test data (required).** This is a measure of the outcome before 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 measuring an outcome other than student achievement, you should include a variable that is highly correlated with your outcome of interest.
- **Background characteristics (recommended).** Background characteristics provide data on observable traits for each participant. These could include gender, ethnicity, individualized education program status, English learner (EL) status, and socioeconomic status (SES). 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, such as 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 could affect them (for example, introducing a reading software). *It is preferable to measure all background characteristics before introducing the educational technology.*

## STEP 2. PROCESSING YOUR DATA

After you have identified the data elements and data sources, the second step is to combine all data elements into one **tidy**[1] dataset and prepare the variables that will be used for analysis. (We explain below how to do this.) We recommend generating tidy data sets not only because it is a requirement

[1] Having **tidy data** means that you've used a standardized way to structure your data set. Specifically: 1.Each variable forms a column; 2. each observation forms a row; 3. each type of observational unit forms a table. To learn more about tidy data, see Wickham, H. "Tidy Data." Journal of Statistical Software, vol. 59, no. 10, 2014, pp.1–23. doi:http://dx.doi.org/10.18637/jss.v059.i10

to use the RCE Coach and most statistical software packages for analyses, but because a tidy data set is easy to manipulate, model, and visualize. The data set at the beginning of this guide is an example of a tidy data set. This section will take you through a series of questions that will help you create your own.

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

## **Exhibit 2.** Example: Columns and rows

| 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:** Reorganize your data set 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 data set that contains all of the variables you will need for analysis?

**NO:** You will have to merge your existing data sets into one complete data set. 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 might be present in some data sets but not in others. Therefore, when merging these data sets you could introduce some missing data. For example, a student in a data set consisting of test scores might not exist in another data set and, therefore, could have missing data for other variables (such as treatment status and background characteristics) after combining the data sets.

**YES:** Continue on to C.

**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 values.

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

**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 might 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.

### **Exhibit 3.** Example: Numeric variables

| 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 (in which the options are Asian, Black, White, and Other) you will have 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 (Exhibit 7).

**YES:** Continue on to D.

**D.** Is all missing data coded as NA in your dataset?

**NO:** If you have merged data sets and there are 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 have not been given a number designation, such as 0 or 999. These values will get incorporated into the analysis.

**YES:** Congratulations! You have a tidy data set!

## 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. You can run the following checks to identify potential data issues that warrant additional investigation:

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

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

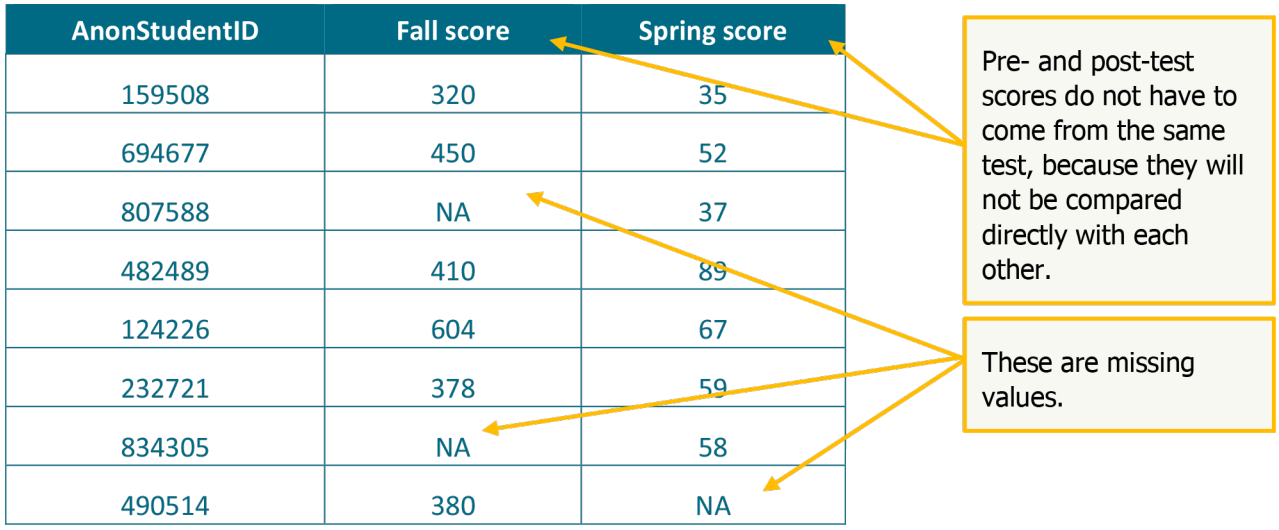
- **Consider the impact of missing data.** The dashboards, and some statistical software packages, will automatically drop observations that contain missing data. You should try to understand why data are missing and how excluding students with incomplete data will affect your results.

**NOTE:** 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; to determine how many observations have at least one missing value you can use a nested COUNTIF with OR function.

## EXAMPLE DATA SETS

### Exhibit 4. Data set 1: Test scores

| AnonStudentID | Fall score | Spring score |
|---|---|---|
| 159508 | 320 | 35 |
| 694677 | 450 | 52 |
| 807588 | NA | 37 |
| 482489 | 410 | 89 |
| 124226 | 604 | 67 |
| 232721 | 378 | 59 |
| 834305 | NA | 58 |
| 490514 | 380 | NA |




### Exhibit 5. Dataset 2: Treatment status

| AnonStudentID | Treatment |
|---|---|
| 159508 | 1 |
| 694677 | 0 |
| 807588 | 0 |
| 482489 | 1 |

## **Exhibit 6.** 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 |

**Data set 3a:** Background characteristics (with non-numeric categorical variables)

**Data set 3b:** Background characteristics (with numeric categorical variables)

EL = English learner; SES = socioeconomic status.

| 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 |

EL = English learner; SES = socioeconomic status.

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 can choose to group medium and high SES together (1 = low and 0 = medium or high) if you are only interested in the effect of low SES on your outcome of interest.

**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 match students with similar levels of SES and to control for SES when analyzing your data. However, some schools or districts might not have access to SES measures or might not be able to use them in this evaluation because of privacy concerns.