Prepare Your Data for Randomization

REQUIRED

At minimum, you must include a list of users with an individual identifier (ID) for each individual, class, or school that will be randomly assigned.

Option 1: If you are assigning individuals, your dataset should include an ID for each student or teacher.

AnonStudentID
159508
694677
807588
482489
555123
124226
232721
834305
490514
573401
275321
288475

Option 2a: If you are assigning groups (classes or schools), your dataset may include student *and* class/school IDs.

AnonStudentID	SchoolID
159508	100
694677	100
807588	100
482489	100
555123	100
124226	200
232721	200
834305	200
490514	300
573401	300
275321	300
288475	300

Option 2b: If you are assigning groups, but you do not yet have individual information or won't need it for your final analysis, your dataset may include only class/school IDs.

SchoolID
100
101
102
103
104
201
202
203
301
302
303
304

Each row represents a single observation (student, teacher, school, etc.). Each observation must have an **individual identifier**. This is a unique code for each participant that will be assigned to the technology user group or the comparison group. The identifiers will allow you to determine who will use the technology, so participants can be notified. They will also allow you to combine (merge) datasets.

RECOMMENDED

The Coach recommends that you also include in your dataset pre-test data and background characteristics. This information will be used to make sure that the randomly assigned treatment and comparison groups are equivalent or balanced before you introduce the technology. If you include these additional variables, your dataset will look like this:

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	Fall_Score	Female	EL	Low_SES	Black	White	Asian	Other
159508	100	3	320	1	1	1	1	0	0	0
694677	100	3	450	0	0	0	0	1	0	0
807588	100	4	NA	1	0	0	1	0	0	0
482489	100	4	410	0	1	0	0	0	1	0
555123	100	4	534	1	1	1	0	0	1	0
124226	200	3	604	0	0	0	1	0	0	0
232721	200	3	378	1	0	1	0	0	0	1
834305	200	4	NA	1	1	NA	NA	NA	NA	NA
490514	300	3	380	NA	NA	NA	NA	NA	NA	NA
573401	300	3	468	NA	NA	NA	NA	NA	NA	NA
275321	300	4	523	NA	NA	NA	NA	NA	NA	NA
288475	300	3	375	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. In the rest of this guide, we have included some tips that will help you manage your data in Excel.

STEP 1. IDENTIFYING DATA SOURCES

There are several types of data you should use to create this initial dataset. Below is a checklist and description of the data. There are examples at the end of this guide showing what each dataset should look like.

□ **List of users who will be randomly assigned (required):** Compile a list of all potential technology users. For example, if you plan to randomly select students to use the technology, you will need an identifier for every student who may be assigned to either the technology user group or the comparison group. If the students are in different schools or different classrooms, you should also include an identifier for the classroom or school. If you plan to randomly select teachers or schools, you will need an identifier for each teacher or each school. (Remember, larger sample sizes are better.)

□ **Pre-test data (recommended):** If available, you should include an outcome measure from before the intervention (such as an assessment from the beginning of the school year). The Coach will double check that the two groups are equivalent before giving you the final list of assignments. If the groups are not balanced, the Coach will re-randomize the list of users until balance is achieved.

□ **Background Characteristics (recommended):** Background characteristics provide data on observable traits for each participant. These could include gender, ethnicity, Individualized Education Program (IEP) status, English Learner (EL) status, socioeconomic status (SES), and more. If you include background traits in your dataset, the Coach can make sure your intervention and comparison groups are well balanced prior to introducing the technology.

CAUTION: Some of these characteristics, like EL status, can change over time. *It is preferable to measure and record 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.

[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	Gender
159508	1	320	F
694677	0	450	M
807588	0	999	F
482489	1	410	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?

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 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 DATASETS

Dataset 1a: *List of participants*

Student_Name	AnonStudentID	School	SchoolID
Guillermo Gonzalez	159508	Alan Elementary	100
Robert Rice	694677	Alan Elementary	100
Sophia Smith	807588	Alan Elementary	100
Patricia Pacheco	482489	Alan Elementary	100
Mariana McGregor	555123	Alan Elementary	100
Brian Brown	124226	Egan Elementary	200
Elena Espinoza	232721	Egan Elementary	200

You will need identifiers for each participant. If you are randomizing groups of participants (classes or schools), you must also include a group ID.

Dataset 1b: *List of participants without personally identifiable information (PII)*

AnonStudentID	SchoolID
159508	100
694677	100
807588	100
482489	100
555123	100
124226	200
232721	200

Eliminate personally identifiable information (PII), such as student and school names, from the data that you upload to the Coach.

Dataset 2: *Test Scores*

AnonStudentID	Fall_Score
159508	320
694677	450
807588	NA
482489	410
555123	534
124226	604
232721	378
834305	NA
490514	380
573401	468

These are missing values.

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 check that the two randomly assigned groups are balanced on SES. 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.