Overview: Creating Groups with Randomization

If you select pilot participants randomly (by chance), you will create two groups that are similar on observed *and* unobserved characteristics. When randomization is well implemented, you can be confident that any differences in outcomes are due to the educational technology you are testing. Because of this, random assignment is considered the best choice for evaluations of effectiveness and should be used whenever possible.

THE EVALUATION CHALLENGE

You want to test whether a technology is effective, but it is impossible to simultaneously observe what happens when an individual uses a technology and doesn't use that same technology. If you introduce a technology and watch what happens, you may notice improvements, for example in student test scores. However, you *cannot* assume that the technology *caused* improved student outcomes. Many other factors (including regular classroom teaching, other programs, student maturation, etc.) could have contributed to the increases.

To overcome these challenges, it is important to compare a group of technology users to a group of non-users, on the assumption that the only real difference between them is whether or not they are using the technology. However, comparing technology users and non-users brings an additional set of challenges. When we make comparisons without trying to ensure similarities between groups, it is possible that those who use the technology are different in any number of ways from those who do not use the technology. For example, hard workers might be more likely to try new technology, but they also perform better on tests. This might cause you to confuse the effect of the technology with the effect of working hard (because either one could cause the technology users to outperform non-users). Those differences can make an ineffective technology look effective, or vice versa.

RANDOMIZED PILOT

Solution: The best test of whether or not your technology works is to randomly select some of the eligible users (a subset of the group you are interested in) to pilot the technology and others to continue classroom practice as usual. Once the groups are assigned by chance, you can be confident that you are comparing apples to apples—that the two groups are the same in every way except the technology use. Then, if you see differences in outcomes (such as student achievement scores) you can be convinced that the new technology is moving the needle.

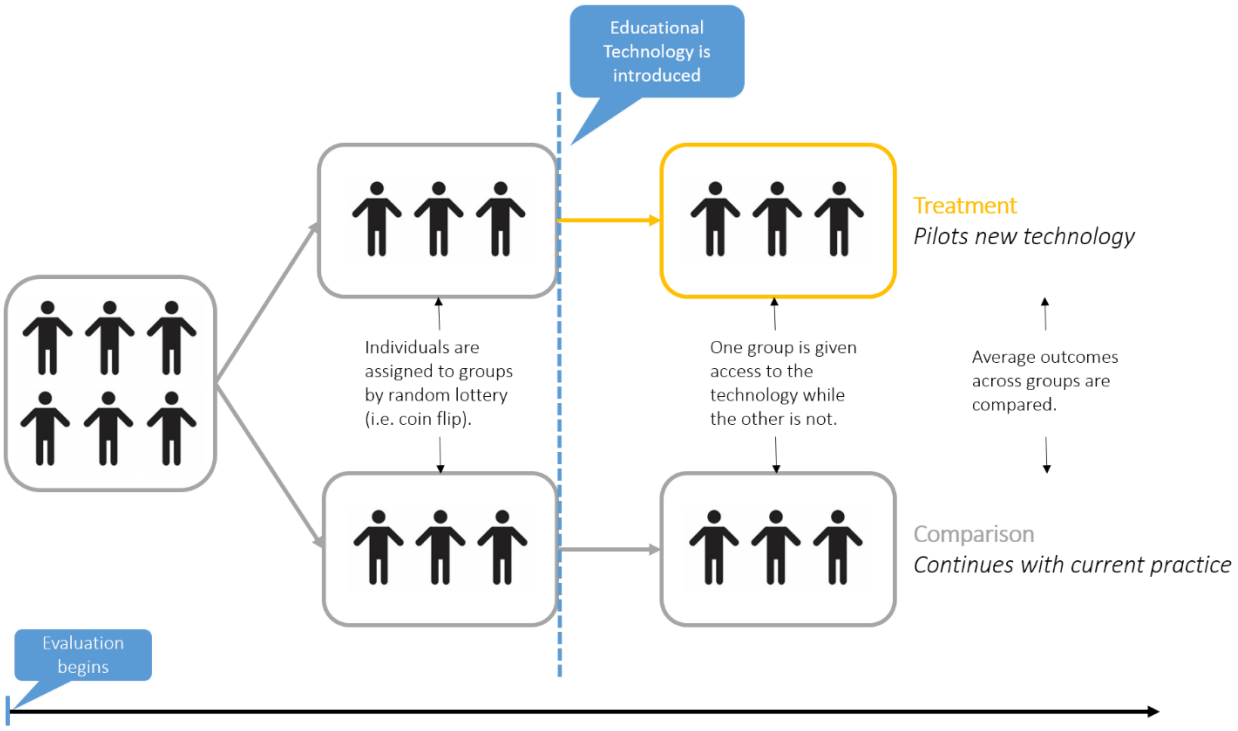
How it Works: Suppose you want to determine the impact of the reading technology U-Read on 5th grade reading test scores in your school. To know if U-Read is having an effect on end-of-year reading scores, you would like to be able to observe your users in a parallel universe. In this universe the parallel non-users are *exactly* the same as the users, but have no access to U-Read. If this parallel group scored lower on the end-of-year reading test, you could conclude that U-Read moved the needle for users.

> *Without Random Assignment:* You can observe that, after implementing U-Read for all 5th graders, reading test scores improved. So, you might conclude that there is a correlation, or relationship, between U-Read and reading scores. However, you cannot conclude that U-Read

caused higher reading scores because their reading scores may have increased anyway. Teachers probably help students learn, regardless of what technology they use.

With Random Assignment: You can randomly assign students at the beginning of the year, using a method equivalent to a coin flip, to one of two groups – users and non-users. You have two groups that are the same on observable and unobservable characteristics *on average*. While they aren't really parallel users, the only difference between the groups in the aggregate is whether or not they use the technology.

Therefore, you can conclude that any differences in achievement are due to the technology, not other factors.



BASELINE CHARACTERISTICS

An important question to ask is, "Was the random assignment successful?" A common way to assess this is to compare the average values of the groups' background characteristics. For characteristics such as test score results from previous years or demographic characteristics, we can quantify the difference between the two groups using a measure called an "effect size." The RCE Coach's randomization dashboard automatically calculates the difference (measured as an effect size) between the group of users and the group of non-users, for any variables you specify. The Coach's randomization dashboard will run and re-run its randomization until you have baseline equivalence on the characteristics you selected.

DEFINITIONS:

*An **effect size** is used to measure different characteristics using the same yardstick. It is calculated by dividing the difference in means between the two groups by the standard deviation of the entire sample.*

Background characteristics are also used in the final analysis, after you implement your technology and collect outcome data. It is necessary to include background characteristics in your analysis of the results if a lot of participants drop out of one group. The dropout rate in an evaluation is called attrition. (For instance, if all students remain in the technology user group but 25 percent drop out of the comparison group, attrition is high, and you might be suspect that the two groups are now different). Attrition may occur for many reasons; for example, students may be absent on the day the test is administered or move to a different district or state, or teachers may decide not to participate in the study after random assignment. Accounting for background characteristics can rebalance the two groups, if the initial differences were not too large. However, it is a good idea to include some background characteristics in your analysis even if your evaluation went as planned. (If you want to learn more about attrition, see the US Department of Education's What Works Clearinghouse (WWC) overview of attrition standards.)