Matching Overview

When random assignment is not possible, you can use matching to create a good comparison group to help you learn what works.

THE EVALUATION CHALLENGE

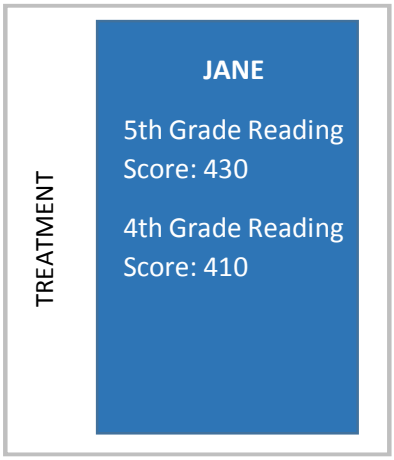
You want to test whether a technology is effective, but it is impossible to simultaneously observe what happens when an individual uses a technology and doesn't use that same technology. If you introduce a technology and watch what happens, you may notice improvements, for example in student test scores. 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 be causing the technology users to outperform non-users). Those differences can make an ineffective technology look effective, or vice versa.

MATCHED COMPARISON DESIGN

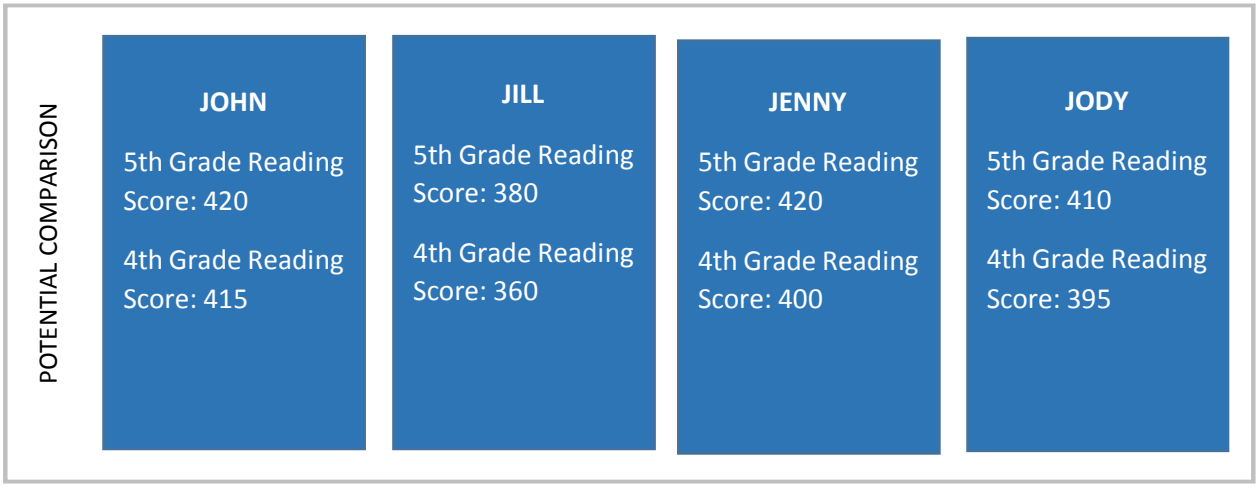
Solution: You can match educational technology users to similar non-users, using pretest measures and background characteristics. Once you have created two similar groups, you are comparing apples to apples—the only real difference between users and non-users is their exposure to the technology. Then, if you see differences in outcomes (such as student achievement scores) you can be confident that the new technology is moving the needle.

How it Works: Suppose you have a student, Jane, who uses the reading technology U-Read. Jane has a higher reading score on the 5th grade state reading test than other students. To know if U-Read is having a positive effect on the reading score, you would like to be able to observe a Jane in a parallel universe. "Parallel Jane" is exactly the same but has no access to U-Read. If this parallel Jane scored lower on the 5th grade state reading test, you could conclude that U-Read moved the needle for Jane.



Without Matching: You can observe that Jane has a higher score than the students who do not use the technology. So, you might conclude that there is a correlation, or relationship, between U-Read and reading scores. But you cannot conclude that U-Read causes higher reading scores because you don't know if there are other factors that led to Jane's higher score, such as having better reading skills before using U-Read.

With Matching: Your initial comparison compares Jane to the average student. Matching attempts to find something as close as possible to parallel Jane. When successful, you can conclude that any differences in achievement are due to the technology, not other factors.



You can compare Jane to one of the four students above. The better the match, the more confident you can be in your conclusion that U-Read is leading to higher test scores. *Who should you compare Jane to?*

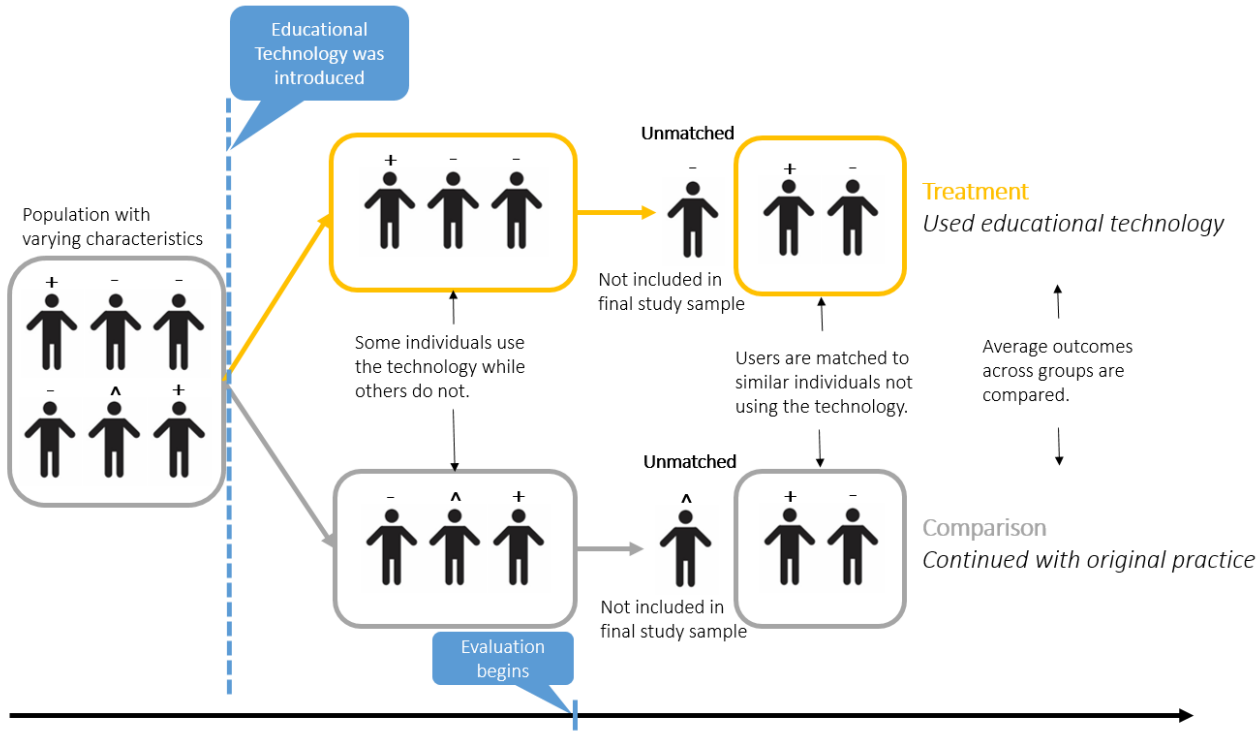
Assuming U-Read was introduced the first day of Jane's 5th grade year, we want to match her with someone who had a similar 4th grade reading score. Matching on a pre-test is fundamental for this technique to work. In this case, we would match Jane to John. Then we can use other statistical techniques to compare their 5th grade achievement and determine if U-Read is moving the needle, or not, for Jane and others using U-Read.

In practice, you will want to use more than one observable characteristic to find matches for groups of students using a technology. For example, imagine that you observe that English is the second language for some of the students using U-Read and that others have Individualized Education Programs. You may want to include these characteristics in your matching strategy because they are good predictors of who is using U-Read and of student achievement. To do so, the RCE Coach Matching tool uses a statistical technique called nearest neighbor matching. Once you create a valid comparison group using the RCE

Coach, you can use the RCE Coach's impact estimation tool to determine if U-Read is moving the needle or not.

NOTE: If the technology is targeted towards a very specific group (or if a very specific type of individual is likely to use the technology), the Coach will be less able to identify a good matched comparison.

For your RCE, the matching process will look something like this:



BASELINE EQUIVALENCE

An important question to ask is, "Was the matching successful?" A common way to assess this is to compare the average values of the groups' background characteristics after matching. 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." 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. The RCE Coach's matching dashboard automatically calculates the difference (measured as an effect size) between the users group average and the matched non-users group average, for any variables you specify. For student achievement outcomes, the US Department of Education's What Works Clearinghouse (WWC) checks the

equivalency of pre-intervention test scores according to the standards defined in Table 1. Meeting the WWC standard for baseline equivalence helps bolster confidence any effects you find are the result of the technology you studied.

Table 1. *WWC standards for baseline equivalence*

Absolute value of difference between groups	WWC conclusion on baseline equivalence
effect size Ò0.05	Satisfies baseline equivalence requirement
0.05 < effect size Ò0.25	Requires statistical adjustment
0.25 < effect size	Does not satisfy baseline equivalence requirement

If the Coach's matching dashboard finds that your groups have a difference in an important baseline characteristic greater than 0.25 effect size units, it will prevent you from moving forward until you can develop a better a match.

NOTE OF CAUTION: The amount of confidence you have in the results of a matched comparison analysis is based on the similarity between the groups of users and matched non-users. Two important notes follow from this:

1. Using observable characteristics to create a matched comparison will not necessarily yield two similar groups. It's important to be thoughtful about the variables you include. Too few variables can lead to groups that aren't actually similar, and too many variables, particularly variables that aren't important for the outcome, will make it too difficult to match similar students. Focus on the variables that you think are important to the outcome or the likelihood of using the technology.

2. Even if you have a large set of observable characteristics, a matched comparison analysis cannot remove the possibility that individuals using the technology are different from those not using it in some *unobservable* way. For example, students using the technology may have parents who are more involved and advocate for additional attention or educational resources than students who are not using the technology. Or, teachers who put in the extra effort to learn and use a new technology may work hard enhance other aspects of their instruction as well.

The only way to remove those potential differences between the two groups is to use random assignment to select the treatment and comparison groups. When properly conducted, random assignment assures that the two groups are similar in both observed and unobservable characteristics.