Ed Tech Rapid Cycle Evaluation Coach

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 an educational technology is effective, but it is impossible to simultaneously observe what happens when an individual uses or doesn't use that technology. If you introduce a technology and watch what happens, you could 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, and so on) could have contributed to the increases.

To overcome these challenges, it is important to compare a group of technology users to a group of nonusers, on the assumption that the only real difference between them is whether they are using the technology. However, comparing technology users and nonusers 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 differ in any number of ways from those who do not use the technology. For example, hard workers might be more likely to try a 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 could cause the technology users to outperform nonusers). Those differences can make an ineffective technology look effective, or vice versa.

MATCHED COMPARISON DESIGN

Solution. You can match educational technology users to similar nonusers using pre-test measures and background characteristics. After you have created two similar groups, you are comparing apples to apples—the only observed difference between users and nonusers is their exposure to the technology (though there might be unobserved differences). 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.

Exhibit 1. Jane's reading scores

TREATMENT
JANE
5th-grade reading score: 430
4th-grade reading score: 410

With matching. Your initial comparison, without matching, compares Jane with 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.

Exhibit 2. Potential comparisons

JOHN	JILL	JENNY	JODY
5th-grade reading score: 420	5th-grade reading score: 380	5th-grade reading score: 420	5th-grade reading score: 410
4th-grade reading score: 415	4th-grade reading score: 360	4th-grade reading score: 400	4th-grade reading score: 395

You can compare Jane with one of the four students shown in Exhibit 2. The better the match, the more confident you can be in your conclusion that U-Read is leading to higher test scores. *With whom should you compare Jane?*

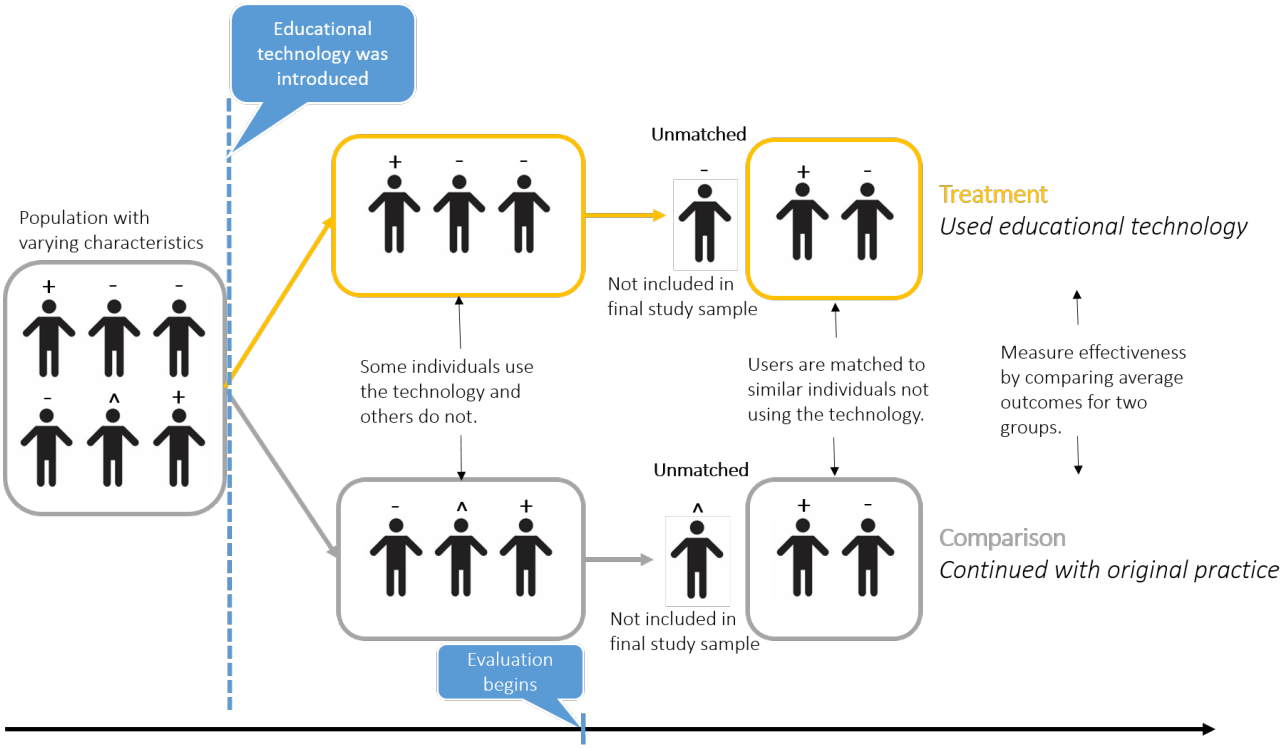
Assuming U-Read was introduced the first day of Jane's 5th-grade year, we want to match her with someone who had a similar 4th-grade reading score. Matching on a pre-test is fundamental for this technique to work. In this case, we would match Jane to John. Then we can use other statistical techniques 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 observed 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 might 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. When you create a valid comparison group using the RCE Coach, you can use the RCE Coach's impact estimation tool to determine whether U-Read is moving the needle.

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 the outcome in Exhibit 3.

Exhibit 3. The matching process



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. Researchers use an effect size 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 nonusers group average, for any variables you specify. For student outcomes, the U.S. Department of Education's What Works Clearinghouse (WWC) checks the equivalency of pre-intervention test scores according to the standards defined in Exhibit 4. Meeting the WWC standard for baseline equivalence helps bolster confidence that any effects you find are the result of the technology you studied.

Exhibit 4. WWC standards for baseline equivalence

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

If the Coach's matching dashboard finds that your groups differ 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.

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 nonusers. Two important notes follow from this:

1. Using observed characteristics to create a matched comparison will not necessarily yield two similar groups. It's important to think 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 observed characteristics, a matched comparison analysis cannot remove the possibility that individuals using the technology differ from those not using it in some unobserved way. For example, students using the technology might 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 might work hard to enhance other aspects of their instruction as well. If so, the measured effect of the intervention could be inaccurate, even if the groups appear to be well matched on observed characteristics.

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 ensures that the two groups are similar in both observed and unobserved characeristics.