tif TEACHER INCENTIVE FUND

tsl TEACHER & SCHOOL LEADER INCENTIVE PROGRAM

Matched-Comparison Group Design: An Evaluation Brief for Educational Stakeholders

White **Paper**

Makoto Hanita
Dana Ansel
Karen Shakman
Education Development Center

January 2017

Table of Contents

Introduction

Schools and districts frequently implement and test new educational interventions with the goal of improving student learning and outcomes. Sometimes these interventions are classroom-based, and other times they involve changes to teacher training, support, and compensation systems. High-quality program evaluations are essential to understanding which interventions work and their impact.

Randomized controlled trials, or RCTs, are considered the gold standard for rigorous educational evaluation. However, RCTs of educational interventions are not always practical or possible. In such situations, a quasi-experimental research design that schools and districts might find useful is a matched-comparison group design. A matched-comparison group design allows the evaluator to make causal claims about the impact of aspects of an intervention without having to randomly assign participants.

This brief provides schools and districts with an overview of a matched-comparison group design and how they can use this research methodology to answer questions about the impact and causality of aspects of an educational program. It also includes a case study of how one Teacher Incentive Fund (TIF) grantee used this methodology as part of its evaluation of the impact of the district's TIF program.

Overview of the Matched-Comparison Group Design

A matched-comparison group design is considered a "rigorous design" that allows evaluators to estimate the size of impact of a new program, initiative, or intervention. With this design, evaluators can answer questions such as:

- What is the impact of a new teacher compensation model on the reading achievement of ninth graders on the state assessment?

- What is the impact of an instructional coaching program on the pedagogical skills of teachers in schools that serve poor and/or minority students?

A matched-comparison group design consists of (1) a treatment group and (2) a comparison group whose baseline characteristics are similar to those of the treatment group at the beginning of the intervention. The more similar the two groups are at baseline, the more likely that the observed difference between the two groups after the intervention can be attributed to the intervention itself, and not to other preexisting differences (either observable or unobservable) between the two groups. Unlike RCTs, in matched-comparison group designs, the treatment and the comparison groups are typically identified after the treatment has already been implemented.

Key Considerations in Developing Matched-Comparison Groups

The most important aspect of this research design is that an evaluator must identify two similar groups, one consisting of individuals who participate in the intervention (treatment group), and the other consisting of those who do not (comparison group). Because in most educational interventions the treatment group is already established, the challenge is to find or create a comparison group. In order to maximize the validity of the comparison, these two groups must be as similar as possible in terms of characteristics prior to the implementation of the intervention. To do this, the evaluator needs data on baseline characteristics of schools, teachers, or students.

There are other relevant considerations to make the match as similar as possible.

Which baseline characteristics to match on? At the end of the intervention, the two groups will be compared in terms of the outcome of interest (e.g., teacher evaluation ratings, student test scores). Therefore, the evaluator needs data on baseline characteristics that could potentially affect the outcome. Such baseline characteristics are called confounders because they could bias (or confound) the estimate of the intervention's effect if they are not controlled through the matching process.

Box 1. Definition of key terms

Treatment group. The group of students, teachers, or schools that participates in the intervention.

Comparison group. The group of students, teachers, or schools that does not participate in the intervention.

Variable. Anything that has a quantity or quality that varies and can be measured.

Outcome variable. Variable of interest that the intervention is designed to improve, such as teacher evaluation ratings or student test scores.

Baseline characteristics. Characteristics of students, teachers, or schools that are measured before the implementation of the intervention.

Selection. Individual tendency to choose to participate or not to participate in the intervention.

Confounders. Characteristics of students, teachers, or schools that affect the outcome of interest, such as a teacher's years of experience or certification.

Proxy variable. Variables that serve as good substitutes for potential confounders due to their similarity to the confounders or high correlation with them.

Propensity score. A summary measure (or score), based on the aggregation of several confounders, that represents the likelihood of an individual's participation in the intervention.

Matching on confounders. Some confounders, however, affect not only the outcome of interest but also selection. Selection refers to an individual's tendency to participate in the program. Those confounders that are associated with selection are especially important to control through matching. For example, an evaluator may be asked to estimate the impact of a performance-based pay program on the quality of classroom instruction. However, teachers with more years of experience may be more likely to participate in the performance-based pay program. In this case, the evaluator may want to match the treatment group teachers and the comparison group teachers in terms of their years of experience. Why? Because otherwise the treatment group would end up consisting of teachers who have more years of experience than the comparison group, which would make the groups fundamentally different and would likely bias results. If more years of experience is associated with higher quality classroom instruction and, in turn, higher evaluation ratings, the observed difference in evaluation ratings between the two groups of teachers would not reflect the impact of the performance-based pay program accurately because it would also include the impact of the teachers' years of experience.

In matched-comparison group designs, an evaluator can only ensure equal distribution of potential confounders that she can measure and for which she has data. This means that the two groups are "equivalent" on only some of the potential confounders. For example, if the evaluator has teacher data on years of experience and advanced degrees, she will be able to match the treatment and comparison groups on

Box 2. Propensity Score Matching

In some matched-comparison group designs, a "propensity score" is used. A propensity score is the likelihood of a particular case being in the treatment group. In our example of teacher participation in a performance-based pay program, a propensity score refers to the estimated likelihood of an individual teacher's participation in the program.

A propensity score is calculated by using a set of potential confounders for prediction. So, considering the example of teachers' participation in a performance-based pay program, variables such as student growth percentiles, teaching experience, and certifications are used as inputs for predicting teacher likelihood of participating in a performance-based pay program. Once calculated, the propensity score could be treated as a summary measure for all the potential confounders that were used for its calculation. As such, matching on the propensity score is analogous to matching on all those confounders – but since the evaluator needs to consider one variable to match, finding a good match becomes far easier when she has the single score.

Calculating propensity score
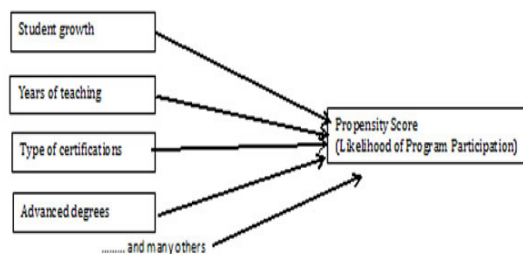
these two variables. However, if she lacks data on teacher skills and motivation, she will not be able to ensure that the two groups are equivalent on those other teacher characteristics that may affect the outcome. So, in planning a matched-comparison group design, an evaluator must make a list of potential confounders, checking for which ones she may have data.

Matching on proxy variables. Often, no data exist for some of the potential confounders. In this situation, an evaluator can consider the use of a proxy, which is a variable that is similar to the potential confounder for which she does not have data or a variable that is highly correlated with the confounder variable. For example, in place of teacher pedagogical skills, an evaluator may use student growth percentiles as the proxy – based on the reasoning that the teacher's student growth percentiles should reflect the teacher's pedagogical skills at baseline. The evaluator may then proceed with matching teachers on this variable in place of pedagogical skills.

Matching on outcome variable measured at baseline. Whenever possible, evaluators should match the treatment and comparison groups on the outcome variable measured at baseline. For example, if the outcome variable is the teacher evaluation rating, evaluators can match the two groups of teachers on this measure taken before their exposure to the intervention (the prior year's teacher evaluation rating).

The outcome measure taken at baseline typically is highly correlated with the outcome measure after the intervention and also likely correlates with other confounders. So, just by matching two groups on the outcome measure at baseline, an evaluator has already made the treatment and the comparison groups somewhat similar in terms of all other confounders. For example, by matching two groups of teachers on the previous year's teacher evaluation rating (the outcome measure at baseline), those two groups become more similar in terms of other confounders such as years of experience, advanced degrees, and skills and motivation, among others.

Matching on multiple baseline characteristics. Once the evaluator determines a set of baseline characteristics that she will use to match groups, the evaluator identifies untreated cases (i.e., schools, teachers, or students who did not participate in the intervention) that match treated cases (i.e., schools, teachers, or students who did participate in the intervention) on these characteristics. Depending on the number of untreated cases, and also the number of baseline characteristics to match, the evaluator may or may not be successful in finding suitable matches. In general, finding matches becomes more difficult when:

1) The pool of untreated cases is small;

2) The number of baseline characteristics to match is large; and

3) The criteria defined for the match are strict.

For example, compiling a matched-comparison group of teachers would be relatively easy if there are only 100 teachers who participate in the intervention, and there are 1,000 nonparticipants in the district—so long as the evaluator is using just a few matching variables (e.g., student growth percentile and years of experience). However, matching would quickly become difficult once the evaluator starts including additional variables (e.g., certifications, courses taught, race/ethnicity). For variables such as growth percentile and years of experience, the evaluator needs to make a decision as to how similar the teachers' scores should be in order to be considered a match. To call two

teachers a match, should the growth percentile be exactly the same, within 3 percentage points of difference, or within 5 percentage points of difference? This is a decision the evaluator must make. Naturally, adopting a stricter criterion for the variable creates more difficulty in finding a match.

To verify that the treatment group and the comparison group are similar, the evaluator will compare the two groups. At minimum, the evaluator must verify that the treatment and the comparison groups have a similar mean for the outcome measured at baseline. For example, she must verify that the two groups of teachers are similar in terms of their mean teacher evaluation rating at baseline, if the outcome of interest is their teacher evaluation rating. (If no data exist for the outcome at baseline, a proxy could be used instead.)

These are important tradeoffs to consider because including more matching variables and applying stricter criteria will make the two groups more similar to each other, which will make for a better comparison. However, if taken too far, there will not be enough matches to allow for the comparison.

Analyzing the Results

Analysis of outcome data for a matched-comparison group design is similar to that for a randomized experiment. In both designs, theoretically the difference in the mean between the treatment and the comparison groups reflects the impact of the intervention. In reality, however, some adjustments are needed. Randomization typically produces two very similar groups, but even so, they are seldom identical. Likewise, matching when done skillfully could produce two very similar groups, but they are never identical. For this reason, the difference in the mean between the two groups is often a bit "off," and requires some statistical adjustment to arrive at a valid estimate of impact.

Box 3. Covariate adjustment

Evaluators typically rely on a technique called covariate adjustment to correct the difference in the mean between the treatment and the comparison groups. Covariate adjustment is a statistical method for adjusting the mean difference so that it would be free of bias resulting from the residual difference between the two groups in terms of confounders that the matching could not eliminate. For example, let's say that the evaluator was asked to estimate the impact of a performance-based pay program on teachers' pedagogical skill, which was measured through classroom observations. The evaluator calculates a mean score on the classroom observation rubric (the outcome measure) of 7.6 for the treatment group and 4.9 for the matched comparison group. The evaluator matched the two groups to be very similar at baseline, based on their baseline classroom observation scores, student growth percentiles, years of experience, but they were not identical. For this reason, differences between groups at the start of the intervention could bias the mean difference of 2.7 between the treatment- and the matched-comparison group in the classroom observation outcome measure. After applying covariate adjustments, the evaluator arrives at the adjusted mean difference of 2.5—which she reports to the districts as the size of impact that the performance-based pay program had on the pedagogical skill of participating teachers.

Box 4. Grantee Spotlight: Miami-Dade School District

Miami-Dade school district used Teacher Incentive Fund (TIF) funds to implement the iHEAT initiative, which supplements an existing districtwide performance-based compensation program. Nine schools have participated in iHEAT, and teacher participation is voluntary. iHEAT provides (1) additional financial incentives for teachers and (2) an opportunity for highly effective teachers to become peer review teachers who identify and address professional development needs of other teachers.

Miami-Dade wanted to know the impact of iHEAT on the increase in the number of effective and highly effective teachers at participating schools. To do this, Miami-Dade and its evaluator used a matched-comparison group research design. They calculated propensity scores and then used those scores to identify nine comparison schools that matched well to the nine iHEAT schools. Next, they contrasted three groups of teachers in the district: (1) iHEAT participants in iHEAT schools, (2) iHEAT nonparticipants in iHEAT schools, and (3) iHEAT nonparticipants in non- iHEAT schools. The outcomes of interest included various teacher evaluation scores (i.e., classroom observation score, value-added score, combined score, summative evaluation rating). The evaluator had access to the baseline data for those teacher evaluation scores, taken during the school year prior to iHEAT (SY 2012–13).

The evaluation results showed that after 1 year of iHEAT implementation (2013–14) all three groups of teachers experienced a decrease in mean value-added scores and an increase in their classroom observation scores. Overall, iHEAT teachers had scores that were the same or slightly above other non-iHEAT teachers in treatment and control schools at baseline, indicating that there was somewhat more room for growth among the non-iHEAT teachers. Comparison school teachers (at the non-iHEAT schools) had the smallest decrease in value-added scores and largest increase in classroom observation scores. Between participants and nonparticipants at iHEAT schools, however, the iHEAT teachers had better results on both their VAM and observation scores. In the following year (2014–15), which is the most recent year of data, Miami-Dade reports that 31% of iHEAT-participating teachers earned incentives as highly effective (the top level) compared with 29% of nonparticipating teachers at iHEAT schools, suggesting that over time the iHEAT initiative may be positively affecting the VAM and classroom observation scores that determine teachers' ratings. However, information about statistical significance is not available at this time.

The district will continue to examine teacher and administrator performance data across all groups to identify potential opportunities for improvement. An ongoing challenge for the district, which might influence comparisons, is the persistent gap in the performance of teachers in high-need schools, of which the TIF schools are a subset, and the relatively small number of non-high-need schools in the district.

For other grantees interested in implementing a similar evaluation design, Miami-Dade district leaders recommend tempering expectations regarding conducting student outcomes' analyses until states have settled on assessment and accountability systems. The constant changes at the state level have affected the district's ability to study outcomes longitudinally. If possible, and with the district's cooperation, TIF grantees might consider administering their own assessment to treatment and comparison schools. However, Miami-Dade leaders advise others to think carefully before adding any additional assessment burdens on teachers and students.

Conclusion

Matched-comparison group design is an excellent option for schools and districts interested in evaluating the impact of new interventions. While not as rigorous as RCTs, it does allow for conclusions regarding the impact of an intervention and can be an effective research design option. However, grantees should consider the following key points when using this design:

- Focus on matching two groups on potential confounders—the background characteristics that affect the outcome of interest.

- A match is only possible on the background characteristics for which the evaluator can get data.

- If at all possible, the evaluator should match the two groups on the outcome of interest taken at the baseline.

- There is a trade-off between the number of background characteristics to match on and the ease of finding a good match. One way to deal with this issue is the use of propensity score for matching.

- The accuracy of the results depends on the treatment and the comparison groups' similarity in terms of their baseline characteristics, and especially in terms of confounders. Therefore, the evaluator must verify the equivalence of these two groups on confounders at baseline.

- The difference in means between the treatment and comparison groups, after a statistical adjustment in terms of residual difference, reflects the size of causal impact of the intervention.

Appendix A.
Additional Resources

Coalition for Evidence-Based Policy. (2014). *Which comparison-group ("quasi-experimental") study designs are most likely to produce valid estimates of a program's Impact*? Retrieved October 27, 2016, from http://coalition4evidence.org/wp-content/uploads/2014/01/Validity-of-comparison-group-designs-updated-January-2014.pdf

National Center for Education Evaluation and Regional Assistance. (2003). Identifying and implementing educational practices supported by rigorous evidence: A user friendly guide. Washington, DC: U.S. Department of Education, Institute of Education Sciences. Retrieved from https://www2.ed.gov/rschstat/research/pubs/rigorousevid/rigorousevid.pdf

Appendix B.
Checklist for Determining the Feasibility of a Matched-Comparison Group Design

1. Has the program already started?

   ☐ Yes, go to #2
   ☐ No, is it possible to randomly assign teachers (or schools) to the treatment and business-as-usual?

      ☐ Yes, conduct a randomized experiment instead of a matched-comparison group design.
      ☐ No, go to #2

2. Are there some teachers (or schools) in the district who are not participating in the program?

   ☐ Yes, identify alternative way to find a comparison group, such as looking at another district.
   ☐ No, got to #3

3. What is your outcome of interest? Is it about students (e.g., academic performance)? Is it about teachers (e.g., evaluation rating)? Or is it about schools (e.g., school rating)?

   ☐ If the outcome is about students, consider matching students and/or group of students (classrooms, schools).
   ☐ If the outcome is about teachers, consider matching teachers and/or group of teachers (schools).
   ☐ If the outcome is about schools, consider matching schools.

4. What are potential confounders (baseline characteristics that are associated with the outcome of interest)?

   Make a list. Pay special attention to those that are associated with both program participation and outcome of interest.

5. For which potential confounders do you have data?

   ☐ For those potential confounders for which you do not have data, are there any variables you could use as their proxies?
   ☐ Are those potential confounders and proxies measured accurately?

6. Do you have data on the outcome measured at baseline? If you do, make sure to match on this as well.

7. Attempt to match the groups. You can use a combination of potential confounders plus the outcome at baseline (if available), or calculate a propensity score (combining confounders into a single score) plus the outcome at baseline.

8. Check the quality of the match by comparing means and frequencies.

   ☐ If the quality of match is good, go to #9.
   ☐ If the quality of match is not good, try matching on a different set of variables, try using a different set of criteria to determine a match, or try using a different propensity score model. You may need to trim parts of the treatment and/or comparison groups that do not have good overlap. Go back to #6.

9. Analyze data. First, calculate an unadjusted mean difference between the treatment and the comparison groups. Then, adjust the mean difference using potential confounders and the outcome at baseline as statistical controls. Report the adjusted mean difference as the estimate of impact of the intervention.