USING DIFFERENCE-IN-DIFFERENCES TO EVALUATE EDUCATIONAL INITIATIVES

QUESTION

How can we measure the true impact of an educational initiative when a randomised experiment isn't possible?

AUTHORS

Professor Markus Meierer, Alona Shmygel
Geneva Research Lab for Digital Impact
University of Geneva

Difference-in-Differences (DiD) is a **widely used method to estimate causal effects** of programs or policies when randomised experiments are not feasible. The method compares changes in outcomes over time between a group that received an intervention (the treatment group) and a group that did not (the control group). By looking at the difference in trends rather than levels, DiD accounts for pre-existing differences between groups and for common time trends that affect everyone equally. This makes it particularly suitable for evaluating educational initiatives, where randomly assigning schools or classrooms to receive or not receive a program is often impractical or politically unfeasible (Angrist & Pischke, 2009).

In its standard form, DiD **requires data for both treated and untreated units** before and after the initiative. In educational settings, the outcome may be pupil performance, attendance, dropout rates, or other relevant educational indicators. Consider a program that introduces a new teaching method in selected schools. Test scores in these schools are compared to test scores in similar schools that did not adopt the method. The DiD estimate is the change in average test scores in treated schools (after minus before), minus the change in average test scores in control schools (after minus before). This double differencing removes both time-invariant school characteristics and common shocks that affect all schools simultaneously.

The validity of DiD rests on the **parallel trends assumption**, which states that treated and control groups would have followed the same trajectory in the absence of the intervention. This assumption cannot be tested directly, but it can be assessed by examining whether outcomes evolved similarly in both groups before the program started. If pre-treatment trends diverge, the DiD estimate may be biased. Researchers should therefore always present visual evidence of pre-treatment trends and, where possible, conduct formal tests of trend differences.

DiD **also offers solutions when the evaluation setting is more complicated** than a simple before-and-after comparison of two groups. A common complication arises when an initiative is rolled out at different times across schools or regions rather than all at once. In such staggered designs, the standard two-way fixed effects estimator may be biased because schools that have already adopted the initiative may (incorrectly) be used as the comparison group for schools that adopt later. Callaway and Sant'Anna (2021) propose an estimator that addresses this by computing group-time average treatment effects, comparing each treatment cohort only to not-yet-treated or never-treated units. Their approach is now considered best practice for staggered designs and is implemented in widely available statistical software.

DiD has been **successfully applied to study educational interventions** across a range of contexts. In the United States, Dettling et al. (2018) used a generalized DiD design to estimate the effect of high-speed internet availability on student outcomes, providing evidence from a high-income setting. In developing countries, Duflo (2001) applied DiD to assess the impact of a large school construction program in Indonesia on educational attainment and wages, demonstrating the method's value where large-scale policy changes create natural experiments. Both studies illustrate how DID can yield credible causal evidence when combined with careful attention to the identification assumptions.

For practitioners planning an evaluation, **collecting baseline data before the initiative starts is essential**. At minimum, researchers need two time periods of outcome data (before and after) for both groups. More pre-treatment periods strengthen the assessment of parallel trends. The choice of control group should prioritize units that are similar to treated units in observable characteristics and institutional context. Transparent reporting of the research design, including trend graphs and robustness checks, is critical for the credibility of the findings.

References

Angrist, J. D., & Pischke, J.-S. (2009). Mostly Harmless Econometrics: An Empiricist's Companion. Princeton University Press.
Callaway, B., & Sant'Anna, P. H. C. (2021). Difference-in-Differences with Multiple Time Periods. Journal of Econometrics, 225(2), 200–230.
Dettling, L. J., Goodman, S., & Smith, J. (2018). Every Little Bit Counts: The Impact of High-Speed Internet on the Transition to College. The Review of Economics and Statistics, 100(2), 260–273.
Duflo, E. (2001). Schooling and Labor Market Consequences of School Construction in Indonesia: Evidence from an Unusual Policy Experiment. American Economic Review, 91(4), 795–813.