Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Volume 8, Number 24, November, 2003 ISSN=1531-7714

Estimating Policy and Program Effects with Observational Data: The "Differences-in-Differences" Estimator

Jack Buckley & Yi Shang
Boston College

When randomized field trials are impossible or impractical, researchers in education and the social sciences more broadly must use observational data, such as standardized test scores or responses to survey questions, to quantitatively evaluate the effects of policies or programs. A potential pitfall with these analyses, however, is that units of observation are not randomly assigned to participate; rather, they self-select to introduce the policy or program of interest. This becomes a problem for estimation and inference if the decision to institute the policy is correlated with the outcome measure—e.g. if states that are more likely to introduce high-stakes testing are also more likely to have a larger gain in test scores. In the econometrics literature, statistical techniques used to analyze these data are often referred to as "treatment effects" models (Goldberger, 1972; Maddala, 1983), where the policy of interest is the "treatment."

In this literature, which spans several disciplines (although perhaps is most developed in econometrics), several families of approaches to modeling treatment effects have been suggested, including:

- "Heckman-type" selection models (Goldberger 1972, based on Heckman's [1976] sample selection model) in which a selection equation and an outcome equation are jointly estimated, assuming a bivariate normal error term in the two equations;
- Instrumental variables estimators (widely used in econometrics, with Brundy and Jorgenson 1971 being a seminal reference) in which an variable can be defined and measured that is related to selection to treatment but not to the outcome measure, and this "instrument" is used to make unbiased inference;
- Nonparametric matching methods, most prominently propensity score matching (Rosenbaum and Rubin 1983), in which the probability of each unit selecting treatment is first estimate, and control observations are chosen by matching on this score to the treatment observations (see Schneider and Buckley 2003 for a recent application in education).

In this article, we consider one of the simplest and most powerful techniques for estimating treatment effects with observational data: the "difference-in-differences" (DiD) estimator. We will briefly describe the DiD model and its underlying assumptions, and then turn to a simple but topical example: applying the DiD to estimate the effects of high-stakes testing on student outcomes.

The Difference-in-Differences Estimator

The basic logic behind the DiD estimator (Ashenfelter, 1978; Ashenfelter & Card, 1985), or the "natural experiment approach," is to model the treatment effect by estimating the difference between outcome measures at two time points for both the treated observations and the controls (those not implementing or participating in the policy or program) and then comparing the difference between the groups—hence the difference-in-differences moniker. This strategy ensures that any variables that remain constant over time (but are unobserved) that are correlated with the selection decision and the outcome variable will not bias the estimated effect.

Clearly this technique thus requires repeated observations of the units. Note, however, that these may be either a true panel, where data is gathered on the same units at both times, or repeated cross-sections, such as two national random survey samples. This is both a strength and a limitation of the DiD model in comparison to the methods described above —the DiD model tends to be more powerful and thus better able to detect small treatment effects, but the other types of treatment effects models can be estimated with only a single cross-section of data.

The key assumption of the DiD model is that the average change in the outcome is presumed to be the same for both the non-participants and, counterfactually, for participants *if they had not participated*. In other words, the analyst must be comfortable in assuming that unmeasured factors, perhaps changes in economic conditions or other policy initiatives, affect both the participants and the non-participants in similar ways. Dee and Fu (2003) provide an excellent discussion of this assumption in an education research context and how to minimize the possibility of its violation through the careful selection of independent variables; Abadie (in press) provides a more technically complex solution that employs the propensity score matching technique to adjust the DiD sample.

We should note that, in addition to this unique identifying assumption, the DiD as we describe it here employs the ordinary least squares estimator and, as such, is sensitive to the usual violations of the Gauss-Markov assumptions (such as homoscedasticity, normality, and no autocorrelation).

If the researcher has pooled cross-sectional data, then the DiD can be estimated with the linear equation:

$$Y_{i,t} = \alpha + \beta D_{i,t} + \delta t + \gamma D_{i,1} + \varepsilon_{i,t}$$

where $Y_{i,t}$ denotes the outcome measure for every unit *i* at both times *t*, *t* itself is a variable coded 1 if the observation is in the second time period and 0 if it is in the initial period, $D_{i,t}$ is an indicator (or "dummy") variable coded 1 if unit is in the treatment group, 0 if in the control group, and $D_{i,1}$ is an indicator variable coded 1 if the observation is in the treatment group *and* in second time period, 0 otherwise. The estimable quantities of interest are thus: α, a common constant for all observations, β, a constant for treatment units only, δ, the effect of time on all units, and γ, the effect of treatment on the treated units (and the main target of inference). The final term, $\varepsilon_{i,t}$, is just an error or disturbance for each unit at each time period. If we assume that these disturbances are uncorrelated normal variates with mean 0 and unknown variance, then we can estimate the DiD model quite simply with the familiar ordinary least squares multiple linear regression estimator.

The expected values of quantities of interest can be shown in a simple table (Table 1, below) that also helps to clarify the model described above:

Table 1: Quantities of Interest in the DiD Model

	Pre-Treatment Outcome	*Post-Treatment Outcome*	*Difference*
Treated Units	$\alpha+\beta$	$\alpha+\beta+\delta+\gamma$	$\delta+\gamma$
Control Units	α	$\alpha+\delta$	δ
Difference-in-Differences			γ

If the researcher has access to true panel data, then the model is even more straightforward and statistically powerful. Differencing the time 1 and time 0 equations yields:

$$Y_{i,1} - Y_{i,0} = \delta + \gamma D_{i,1} + \varepsilon_i^*$$

where $Y_{i,1} - Y_{i,0}$ is the difference between the repeated outcome measures for each observation, $D_{i,1}$ is the treatment indicator, γ is the treatment effect, δ is still the effect of time on all units, and ε_i^* is the difference between errors at time 1 and time 0, which is itself a normal random variate with mean 0.

Practically speaking, to estimate the DiD model for panel data, all the researcher has to do is to compute the difference between observed outcome measures and then regress this on a constant and a dummy variable for whether-or-not the unit of observation elected to participate in the treatment or not. The repeated cross section version is a little more involved, as the researcher must construct the additional indicators for time and for time and treatment, but basically the same steps are required. Before we turn to an example, however, we must first consider a slightly more complicated model.

Extending the DiD Model

Often the simple DiD model may not be sufficient to capture the dynamics that our theory suggests are occurring in the real world. The easiest way to include additional factors to account for heterogeneous dynamics in a DiD model is to simply add them linearly to the regression equation. Say, for example, we have an additional demographic variable, X_i, that we wish to include. For the repeated cross section data, the model thus becomes:

$$Y_{i,t} = \alpha + \pi_t X_i + \beta D_{i,t} + \delta t + \gamma D_{i,1} + \varepsilon_{i,t}$$

where π_t are the effects of the new covariate on the outcome for each of the two time points which are practically computed by estimating a separate coefficient for X_i at time 0 and at time 1. Once again, the panel data model is slightly simpler due to differencing:

$$Y_{i,1} - Y_{i,0} = \delta + \pi X_i + \gamma D_{i,1} + \varepsilon_i^*$$

where π is just $\pi_1 - \pi_0$ from the previous model. Of course, either model can be extended to simultaneously consider several independent covariates instead of just a single X_i.

As Meyer (1995) points out, however, if the researcher believes that the treatment may actually have different effects on different units depending on these additional variables, then this simple linear model will not be sufficient to capture the heterogeneity of the dynamics. One possible solution that is easy to implement is the inclusion of interactions between the treatment indicator and the additional covariates, yielding:

$$Y_{i,t} = \alpha + \pi_t X_i + \beta D_{i,t} + \delta t + \gamma D_{i,1} + \lambda_t X_i D_{i,1} + \varepsilon_{i,t}$$

for the multiple cross-sections and:

$$Y_{i,1} - Y_{i,0} = \delta + \pi X_i + \gamma D_{i,1} + \lambda X_i D_{i,1} + \varepsilon_i^*$$

for panel data. Here λ_t or λ denotes the coefficient on the new interaction term. These models (once again extendable to multiple additional coefficients) allow the modeling of nonlinearity in the treatment effect due to differences in level of the additional covariates. To better clarify this point, as well as the usage of the previous models, we now turn to an example.

Example: DiD Models for the Effects of High-Stakes Testing

High-stakes tests and their uses by education policymakers remain a vital and contentious area of research. In a recent article, Amrein & Berliner (2002) make a strong contribution to the debate over whether such tests boost student achievement by examining the performance of states who have adopted high-stakes testing on a variety of independent measures, or audit tests, such as the SAT and the NAEP math and reading tests. Their findings are challenged, however, by Rosenshine (2003) who points out that they failed to include a proper control group for comparison in their analyses of NAEP 4th and 8th grade mathematics and 4th grade reading scores. In their response to Rosenshine, Amrein-Beardsley & Berliner (2003) concede this point, but report different results due to removing a greater number of states from the population of interest because of their "unclear" status (they had a changing rate of student exemptions from the NAEP tests).

To illustrate the various DiD models discussed above, let us consider the data on the 4th grade NAEP math scores in 1996 and 2000 (here we ignore the reclassification of unclear status and retain the full 35 observations, 18 of which are high-stakes states, examined in the original paper and by Rosenshine). Since the authors are considering state average scores at two time periods and the unit of analysis is individual states, the panel form of the DiD estimator is appropriate here. A simple model might thus be:

$$\text{Score}_{i,1} - \text{Score}_{i,0} = \delta + \gamma \text{High Stakes}_{i,1} + \varepsilon_i^*$$

where the dependent variable is the difference in scores between time 1 and time 0 and the variable "High Stakes" is an indicator coded 1 if state *i* requires high-stakes tests and 0 otherwise.

We may be interested in a one or more demographic measures of the states, however, such as racial composition of the students, or per capita educational spending. Accordingly, we could include a measure of 1998-1999 per public school pupil spending (in 1000's of dollars, data from NCES) and estimate a second model:

$$\text{Score}_{i,1} - \text{Score}_{i,0} = \delta + \gamma \text{High Stakes}_{i,1} + \pi \text{Spending}_i + \varepsilon_i^*.$$

Finally, what if high-stakes testing actually has a different effect on outcomes depending on the level of per pupil spending? We now need to estimate a third model that includes this interaction:

$$\text{Score}_{i,1} - \text{Score}_{i,0} = \delta + \gamma \text{High Stakes}_{i,1} + \pi \text{Spending}_i + \lambda \text{High Stakes}_{i,1} \times \text{Spending}_i + \varepsilon_i^*.$$

Table 2, below, compares the results of the three models. Note that example SPSS syntax for estimating all models

presented below (and their repeated cross-sectional counterparts as well) in is included below in the Appendix.

Table 2: Comparing DiD Models for 4th Grade Math NAEP Data

	Model 1	*Model 2*	*Model 3*
δ	2.117	-0.70	-1.995
(Constant time effect for all states)	(.609)	(2.31)	(3.119)
γ	2.382	2.410	6.493
(High-stakes testing treatment effect)	(.850)	(.850)	(4.520)
π		.345	.648
(Per pupil spending effect, 000's of $)		(.351)	(.482)
λ			-.648
(High-stakes testing and per pupil spending interaction effect)			(.705)
Root mean squared error of the model	2.512	2.514	2.520

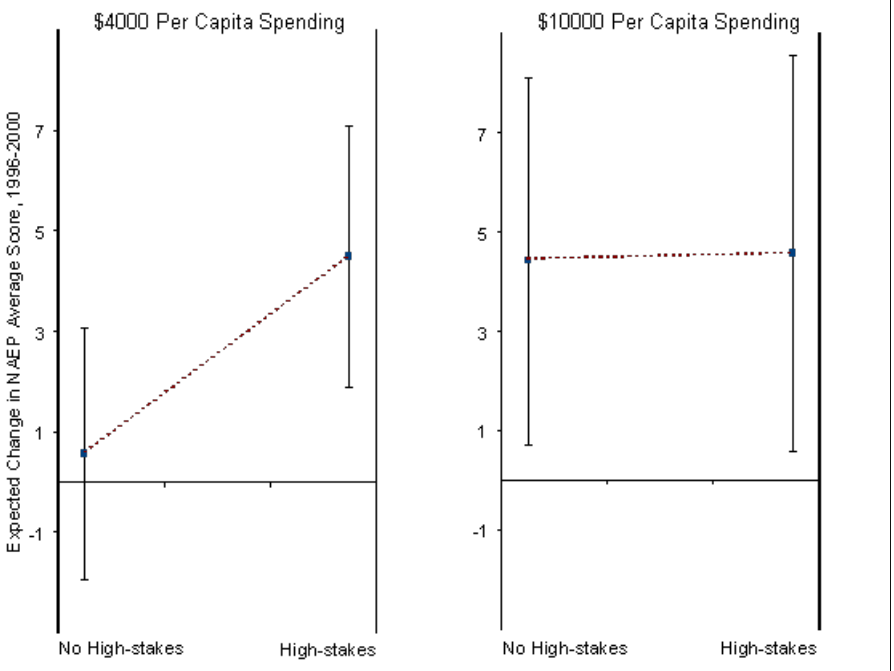
Note: Number of observations = 35. Results presented are estimated coefficients from ordinary least squares multiple linear regressions, standard errors in parentheses.

Before discussing the results, we should note that we neither test nor correct for violations of Gauss-Markov assumptions here, but suspect that heteroscedasticity is a potential problem. We suggest, at a minimum, that researchers who detect heteroscedasticity replace the conventional standard errors with White's (1980) consistent standard errors, or consider alternative approaches to estimating heteroscedastic regression models.

The first column of Table 2 presents the results of the simplest panel DiD. These results suggest that all states over this period had an average increase in NAEP 4th grade mathematics scores of 2.117, but that high-stakes test states had an average additional increase of 2.382 points (a substantively meaningful and statistically significant gain). The second column shows that considering per pupil spending does not meaningfully alter the high-stakes effect, but predicts an average gain of .345 points per $1000 spent (note however that there is a great deal of uncertainty about this estimate and it does not approach significance at conventional levels).

The third column, which reports the results of considering an interaction between spending and high-stakes testing, bears closer examination. Because it is notoriously difficult to directly interpret regression coefficients in an interaction model, we use the estimated variance-covariance matrix from the model to simulate predictions while varying quantities of interest in the model (King, Tomz, & Wittenberg, 2000; Tomz, Wittenberg, & King, 2000). This allows for easy interpretation but still proper accounting of the uncertainty of our model estimates. In particular, we are interested in exploring the effect of high-stakes testing on the difference-in-differences at various levels of per capita education spending. The results of the two such sets of simulations are presented in Figure 1, below.

Figure 1: Predictions from the Panel DiD Model with Interactions



Note: Vertical bars represent 95% confidence intervals. Horizontal axis represents zero expected change in scores.

In these simulations, we estimate the impact of high-stakes testing on the difference in test scores for a hypothetical state with average per capita spending of $4000 (approximately the low end of the sample) versus one with spending of $10000 (the high end). The results are quite interesting. The graph on the left shows the differences for the high-stakes and no high-stakes conditions at the $4000 spending level—graphically, the difference-in-differences is thus the right point minus the left point. As is clear from the overlap of the plotted 95% confidence intervals, there is insufficient precision to distinguish the high-stakes effect from zero at this level of significance. Nevertheless, the results are suggestive that high-stakes testing might predict an increase in student achievement.

As the right graph shows, however, this is clearly not the case at the $10000 per pupil spending level. Here, the confidence intervals overlap almost totally, and the means are virtually indistinguishable. In short, at this level of spending the DiD prediction is essentially zero: high-stakes testing does not appear to improve student scores when spending is at this level. Interestingly, we can conclude at the 95% level that states with this amount of spending do have an expected gain in their average NAEP 4^{th} grade math score during this period, but high-stakes testing appears to provide no additional benefit.

Conclusion

The DiD estimator is a useful tool for applied quantitative education and public policy researchers confronted with observational data in which self-selection to treatment may be confounded with the outcome measure. The model is extremely flexible, and allows for the inclusion of additional covariates that are hypothesized to influence either the baseline change common to all units of observation or the amount of change predicted by the treatment. Moreover, the models are simple to estimate with "off-the-shelf" technology and are reasonably easy to interpret.

Appendix: SPSS Syntax for DiD Models

SPSS Syntax for DiD Models Based on 4^{th} Grade Math NAEP Data (Panel Data)

The example included in the article uses true panel data, and only includes states that report scores for both the 1996 and 2000 NAEP tests. All three models presented above can be estimated in SPSS using the following syntax:

```
REGRESSION

    /MISSING LISTWISE
    /STATISTICS COEFF OUTS R ANOVA
    /CRITERIA=PIN(.05) POUT(.10)
    /NOORIGIN
    /DEPENDENT d4math
    /METHOD=ENTER histakes  /METHOD=ENTER spend  /METHOD=ENTER hsspend  .
```

Note: the coefficients of the variables *histakes* (the high-stakes testing dummy), *spend* (the level of state educational

spending per pupil (in 000's of $), and *hsspend* (the interaction between the two) correspond respectively to γ, π, and λ.

SPSS Syntax for DiD Models Based on 4th Grade Math NAEP Data (Cross-section Data)

We also include, for completeness, syntax and results of the repeated cross-section DiD model discussed in the paper. Here we assume that we do not have true panel data, and thus we ignore the names of the states and simply model the data as two cross-sections. An interesting consequence of this change is that we can now include states that only have a reported NAEP score in one of the two years.

```
REGRESSION

    /MISSING LISTWISE
    /STATISTICS COEFF OUTS R ANOVA
    /CRITERIA=PIN(.05) POUT(.10)
    /NOORIGIN
    /DEPENDENT y
    /METHOD=ENTER time  /METHOD=ENTER histakes  /METHOD=ENTER hstime
    /METHOD=ENTER spend  /METHOD=ENTER hsspend  .
```

Note: *time* is coded 0 for 1996 and 1 for 2000; *histakes* is coded 1 for high-stakes states and 0 for non-high-stakes states; *hstime* is coded 1 for high-stakes states in 2000 and 0 otherwise; *spend* is *hsspend* is again the interaction between *histakes* and *spend*).

Table A1: Results of the Cross-Section DiD Models:

	Model 1	*Model 2*	*Model 3*
α (Common constant for all observations	224.857 (1.412)	208.908 (3.480)	212.710 (4.830)
δ time (Time effect for all states)	1.943 (2.022)	2.393 (1.781)	2.286 (1.781)
γ hstime (High-stakes testing treatment effect)	2.694 (2.844)	2.994 (2.503)	3.269 (2.510)
β histakes (Constant for states that choose to have high-stakes)	-4.994 (1.975)	-5.167 (1.738)	1.879 (.722)
π spend (Per pupil spending effect, 000's of $)		2.467 (.503)	-12.562 (6.755)
λ hsspend (High-stakes testing and per pupil spending interaction effect)			1.138 (1.004)
Root mean squared error of the model	6.472	5.694	5.684

Note the large increase in root mean squared errors (i.e. average prediction error) of these models versus those reported in the body of the article above. Clearly the panel approach, when possible, is to be preferred; the presumed benefit of adding states with observed NAEP scores only in either 1996 or 2000 is overwhelmed by the additional uncertainty of the cross-section model.

References

Abadie. A. (in press). Semiparametric Differences-in-Differences Estimators. *Review of Economic Studies*.

Amrein, A. L., & Berliner, D. C. (2002). High-stakes testing, Uncertainty, and Student Learning. *Education Policy Analysis Archives*, *10(18)*. Retrieved Nov 10, 2003 from http://epaa.asu.edu/epaa/v10n18 .

Amrein-Beardsley, A., & Berliner, D. C. (2003). Re-analysis of NAEP Math and Reading Scores in States with and without High-stakes Tests: Response to Rosenshine. *Education Policy Analysis Archives*, *11(25)*, Retrieved Nov 10, 2003 from http://epaa.asu.edu/epaa/v11n25/ .

Ashenfelter, O. (1978). Estimating the Effect of Training Programs on Earnings. *Review of Economics and Statistics, 60*(1), 47-57.

Ashenfelter, O., & Card, D. (1985). Using the Longitudinal Structure of Earnings to Estimate the Effect of Training Programs. *Review of Economics and Statistics, 67*(4), 648-660.

Brundy, J. and D. Jorgenson. (1971). Consistent and Efficient Estimation of Systems of Simultaneous Equations by Means of Instrumental Variables. *Review of Economics and Statistics*, 53, 207-224.

Dee, T. S., & Fu, H. (2003). Do Charter Schools Skim Students or Drain Resources? Working paper. Retrieved Nov 10, 2003 from http://www.swarthmore.edu/socsci/tdee1/Research/charter0503.pdf.

Goldberger, A. (1972). *Selection Bias in Evaluating Treatment Effects: Some Formal Illustrations.* Madison, Wisconsin: University of Wisconsin Press.

Heckman, J. (1976). The Common Structure of Statistical Models of Truncation, Sample Selection, and Limited Dependent Variables, and a Simple Estimator for Such Models. *Annals of Economic and Social Measurement*, 5, 475-492.

King, G., Tomz, M., & Wittenberg, J. (2000). Making the Most of Statistical Analyses: Improving Interpretation and Presentation. *American Journal of Political Science, 44*(2), 347-61.

Maddala, G. S. (1983). *Limited-Dependent and Qualitative Variables in Econometrics.* Cambridge, U.K.: Cambridge University Press.

Meyer, D. B. (1995). Natural and Quasi-Experiments in Economics. *Journal of Business and Economic Statistics, 13*, 151-161.

Rosenbaum, P. and Rubin, D. (1983). The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika*, *70*, 41-55.

Rosenshine, B. (2003). High-stakes Testing: Another Analysis. *Education Policy Analysis Archives*, *11(24)*. Retrieved Nov 10, 2003 from http://epaa.asu.edu/epaa/v11n24/.

Schneider, M. and Buckley, J. (2003). Making the Grade: Comparing DC Charter Schools to Other DC Public Schools. *Educational Evaluation and Policy Analysis*, *25(2)*, 203-215.

Tomz, M., Wittenberg, J. and King, G. (2000). *CLARIFY: Software for Interpreting and Presenting Statistical Results.*

White, H. (1980). A Heteroscedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroscedasticity. *Econometrica*, *48*, 817-838.

Authors:

Jack Buckley & Yi Shang
Department of Educational Research, Measurement, and Evaluation
Boston College
Lynch School of Education, Campion Hall 336E
Chestnut Hill, MA 02467

(617) 552-8089

bucklesj@bc.edu
www2.bc.edu/~bucklesj

Descriptors: Decision Making; Evaluation Methods; Program Effects; Statistical Methods

Citation: Buckley, Jack & Yi Shang (2003). Estimating policy and program effects with observational data: the "differences-in-differences" estimator. *Practical Assessment, Research & Evaluation*, 8(24). Available online: http://PAREonline.net/getvn.asp?v=8&n=24.