Difference-in-Differences when the Treatment Status is Observed in Only One Period[*]

Irene Botosaru[†]and Federico Gutierrez[‡]

July 7, 2015

Abstract

This paper considers the difference-in-differences (DID) method when the data comes from repeated cross-sections and the treatment status is observed either before or after the implementation of a program. Empirical research that has faced this missing data issue has replaced the unobserved treatment status appearing in the DID estimand by a proxy, without accounting for the fact that the proxy is an imperfect measure of the true treatment status. We show that this method obtains biased results for the average treatment effect on the treated (ATT). Our contribution is at least four-fold: (1) We show that when the proxy satisfies an exclusion restriction and the propensity score is stationary, the ATT is identified via a DID analysis; (2) We propose two GMM estimators; (3) We provide an overidentification test for our assumptions; and (4) We provide both an empirical illustration and an application of our method.

JEL: C1, C4, I3, O1

[*] We are grateful to Sylvain Chabé-Ferret, Yanquin Fan, Jean-Pierre Florens, Xavier D'Haultfoeuille, Yuichi Kitamura, Pascal Lavergne, and Chris Muris for helpful discussions, and to seminar audiences at TSE, Bristol Econometric Study Group 2013, AMES Singapore, Wisconsin-Madison, UW Seattle, UBC, SFU, Vanderbilt, Yale, LAMES Mexico City, CEA Toronto, and the 5th French Econometrics Conference (Toulouse) for comments and suggestions.

[†] Simon Fraser University, Economics Department, 8888 University Drive, Burnaby, BC, Canada.

[‡] Corresponding author: federico.gutierrez@vanderbilt.edu, Department of Economics, 415 Calhoun Hall, Vanderbilt University, TN, USA.

1 Introduction

Researchers in the social sciences often have to rely on data from observational studies in order to evaluate causal effects of social programs. Since observational studies do not require randomization of participation to the program, statistical methods used to evaluate the effects of the program have to account for the possibility of selection bias.[1] Difference-in-differences (DID) is a heavily used method that accounts for selection bias due to time-invariant unobserved covariates. DID obtains the average treatment effect on the treated (ATT) by comparing the difference in post- and pre-program outcomes between two distinct groups: a group that participates in the program, known as the treated group, and a group that does not participate, known as the control group. DID can be applied with either panel data or repeated cross-sectional data as long as the observed outcomes can be classified into treated and control *both* before and after the program.

The use of repeated cross-sectional data, however, raises issues related to data availability and the identification of the ATT. In repeated cross-section surveys, it is generally the case that the treatment status is observed *only after* the implementation of the program. In such cases, it is not possible to classify pre-program outcomes into treated and control, resulting in lack of identification of the ATT. For example, many of the nationally-representative cross-sectional household surveys usually include questions regarding participation in active social programs. However, they rarely contain retrospective information on *pre-program* outcomes. Information about pre-program outcomes is only available in pre-program surveys in which the treatment status is not observed, since pre-program surveys do not contain information on future participation of programs not yet implemented.[2]

Abadie (2005) notes that one way around this missing data issue is to determine the treatment status of the pre-program sample from "some individual characteristic observed in both periods." Essentially, Abadie (2005) is suggesting using a proxy for the latent treatment status in order to group the pre-program outcomes into control and treated. To our knowledge, we are the first to show that, in order for this strategy to identify the ATT, additional assumptions must be imposed and the expression for the DID estimand must be changed accordingly.

Researchers have used the strategy of conditioning on a proxy for the unobserved treatment status quite often, albeit incorrectly. As such, it is important to work out the precise conditions needed for this procedure to identify the ATT. For example, Groen and Polivka (2008) studies the effect of hurricane Katrina on the labor market outcome of evacuees using data from the Current Population Survey (CPS). The treatment status is being an evacuee due to hurricane Katrina. Since pre-Katrina rounds of the CPS do not contain

[1] Selection bias arises whenever the decision to participate to the program is correlated with the outcome.

[2] Examples of such social programs are conditional and unconditional cash transfers implemented during the last couple of decades in many countries around the world, such as Ecuador, Peru, Brazil, Nicaragua, etc. In only few exceptions, e.g. Progresa in Mexico, the implementation was made with the explicit objective to be evaluated. The evaluation of most social programs has to be made using non-experimental methods. The non-observability of the treatment status in pre-program periods is the rule rather than the exception in these cases.

identifiers of (future) evacuees, the authors proxy the unobserved treatment status with an indicator for living in the affected areas. Although this proxy is highly correlated with the latent treatment status, the authors do not account for the fact that the proxy is not the same as the true treatment status resulting, as we show in the paper, in a biased estimate for the ATT.[3]

The problem of unobserved treatment status is not exclusive to the pre-program period. To assess long-term effects, researchers have to relay on surveys conducted several years after the treatment took place. For example, Galasso and Ravallion (2004) study the impact of Argentina's social program, JEFES. The authors use the fact that the Argentinian household survey they use consists of rotating panels, with a sub-sample of households that can be followed for one year and a half. Using the households that can be observed both before and after the implementation of the program, the authors use DID to estimate the short-run effects of the program. Our method allows for the estimation of the long-term effects of JEFES since our analysis is not restricted to the subsample that forms a panel.

In this paper, we show the identification of the ATT via the DID method when the only available data comes from repeated cross-sections with missing group membership in one of the two time periods. Our identification strategy is based on formulating the problem as a finite mixture problem,[4] and our main assumptions are (1) the existence of a proxy for the latent treatment status, where the proxy satisfies an exclusion restriction conditional on the true treatment status, and (2) the stationarity of the propensity score conditional on the proxy. In the literature on nondifferential measurement error, when proxies satisfy exclusion restrictions conditional on the true variables that they measure, the proxies are known as surrogates, see e.g. Carroll et al. (2006) and Hu (2008). Examples of such proxies are (i) covariates known to have been used by the program organizers to determine eligibility,[5] (ii) coarse measurements of the treatment status such as treatment status observed with error due to either coding or recalling error, (iii) surrogate measures of exposure,[6] or (iv) measures taken at a fixed point in time when the treatment status is supposed to measure a long time trend.[7] We note here that the propensity score is likely to be stationary when the cross-sections come from the same population and anticipatory behavior is ruled out, both of which are standard assumptions in the DID literature, see e.g. Blundell and Costa Dias (2009).

Our identification strategy gives rise to an expression for the ATT that allows us to cast our estimation

[3] The common (and incorrect) strategy of replacing the treatment indicator with a proxy can be done only when there is a single proxy strongly correlated with the treatment status. In contrast, our method can allow for a set of proxy variables where each proxy may be weakly correlated with the treatment status but such that they may be jointly relevant in predicting it.

[4] We are not the first ones to connect mixture models to treatment effects models. To our knowledge, one of the first papers to do this was Cross and Manski (2002).

[5] We illustrate this particular case in our applications.

[6] E.g. an indicator for smoking when the true treatment status is supposed to measure exposure to both first and second hand smoke.

[7] E.g. average cell phone usage last week when the true treatment status is supposed to measure long term average cell phone usage.

procedure as a grouped-data estimator. Using insights from Angrist (1991), we first show that the sample equivalent of our ATT estimand is equivalent to a two-stage least-squares estimator. Then we express the estimator as a GMM estimator and estimate all parameters of the model simultaneously. We show that this estimator is always just-identified. However, from our identification proof it is clear that when the proxy is a discrete random variable taking on more than two different values, the ATT is overidentified. To take advantage of this possible overindentification, we introduce another GMM estimator by parametrizing the outcome equations. We then present an overidentification GMM test for our main identification assumptions - the stationarity of the propensity score and the exclusion restriction of the proxy. We show the performance of both estimators via Monte Carlo simulations.

To illustrate our results on real data, we provide both an empirical illustration and an application. First, for our empirical illustration, we aim to compare the results of our method to those of the standard DID. We do this by working with a panel data set with which we first estimate the standard DID and second, after we delete the treatment indicator from the pre-treatment period, we estimate the ATT via the method we propose in this paper. For this purpose, we estimate the effect of the Mexican conditional-cash transfer program PROGRESA on the school attendance of teenagers.

Second, for our application, we evaluate the effect of the Peruvian social program JUNTOS on the demand for health care using data from the *Encuesta Nacional de Hogares* (EHANO). This data set contains information on program participation and outcomes after the program was implemented. But it only contains information on the outcomes and not on program participation before the program was implemented, which is exactly our setting.

Our paper is related to the literature on treatment effects with misclassification, see Mahajan (2006), Lewbel (2007), and Hu (2008).[8] All these papers deal with cross-sections and they focus on the identification of the average treatment effect (ATE) by requiring the existence of both an instrumental variable and a proxy for the mismeasured treatment status. In contrast, our paper takes advantage of the time dimension of the DID set-up, which allows us to require just one proxy for the treatment status. Other related papers are Chen, Hu, and Lewbel (2009), which identifies the ATE by imposing a symmetry assumption on the density of the measurement error rather than relying on additional information such as proxies and instrumental variables, and Molinari (2010), which provides partial identification results for the ATE.

Our identification strategy is closely related to the identification of finite mixtures, with the most relevant paper being Henry, Kitamura, and Salanié (2014). Although Henry, Kitamura, and Salanié (2014) do not make the connection to DID, their results can be seen as complementary to ours, in the sense that their paper would obtain partial identification of the ATT when (i) either the treatment status is not observed in any of the time periods, or (ii) the stationarity of the propensity score does not hold. We are able to obtain

[8] Notice that in our set-up we have "extreme" misclassification, in the sense that the treatment status is completely missing in one of the time periods.

point identification by making use of both the observability of the treatment status in at least one of the periods and of the stationarity of the propensity score.

Finally, our first estimator is a two-stage least squares estimator. A similar estimator has been studied in a different context by Angrist (1991). Our second estimator uses a GMM framework to combine the pre- and post-treatment samples in order to estimate the ATT. There is a rich literature in econometrics and statistics on data combination, see Ridder and Moffitt (2007) for a review.

The paper is organized as follows. Section 2 shows the identification of the ATT via DID. Subsection 2.1 considers the standard set-up, in which the treatment status is observed in both time periods, while Subsection 2.2 considers identification of the ATT when the treatment status is missing in the pre-treatment period and a proxy for the latent status is available. In Section 3 we introduce our estimators. Subsection 3.1 introduces our two-stage least squares estimator, which is then shown to be equivalent to a GMM estimator in Subsection 3.2. Subsection 3.3 introduces our second GMM estimator and our overidentification test for model specification. The small sample performance of our estimators is shown in Section 4. Section 5 shows our empirical illustration to the Mexican conditional cash transfer program, Progresa, while Section 6 presents our application to the Peruvian cash transfer program, Juntos. Finally, Section 7 concludes. In the Appendix, we include identification assumptions when covariates are present, as well as additional simulation results.

2 Identification

In this section, we first discuss the standard DID estimand and show that it identifies the ATT when individual group membership is observed in both time periods. Then we introduce our solution to doing a DID analysis when the treatment status is observed in only one period. We consider the case without observed covariates for simplicity of exposition. We show identification of the ATT with covariates in the Appendix.

2.1 Standard DID

The standard DID framework is as follows. A fraction of a population receives a treatment between two time periods, $t \in \{0, 1\}$. At each time period t, each individual has two potential outcomes: $Y_t(1)$ if the individual is exposed to the treatment and $Y_t(0)$ if the individual is not exposed to the treatment. Letting $D \in \{0, 1\}$ represent the treatment indicator, the realized outcome at time t is:

$$Y_t = DY_t(1) + (1 - D)Y_t(0), \ t = 0, 1 \tag{1}$$

The parameter of interest is the ATT defined as the difference in post-program potential outcomes:

$$ATT \equiv E\left(Y_1\left(1\right) - Y_1\left(0\right)|D = 1\right) \tag{2}$$

while the standard DID estimand is defined as:

$$\theta \equiv \left[E\left(Y_1|D = 1\right) - E\left(Y_0|D = 1\right)\right] - \left[E\left(Y_1|D = 0\right) - E\left(Y_0|D = 0\right)\right] \tag{3}$$

We show below that θ equals the ATT under the following standard assumptions.

For what follows, let $F_t\left(Y_t, D\right)$ denote the joint distribution function of Y_t and D at period $t = 0, 1$.

Assumption A1 (parallel paths) In the absence of the treatment, the average outcomes for the treated would have followed the same trend as that for the control:

$$E(Y_1(0)|D = 1) - E\left(Y_0\left(0\right)|D = 1\right) = E\left(Y_1(0)|D = 0\right) - E\left(Y_0(0)|D = 0\right)$$

Assumption A1 allows for selection on time-invariant unobservables. This assumption may be violated if, for example, the treatment and the control groups do not respond in the same way to macro shocks.

Assumption A2 (no anticipation) There is no anticipatory response for those in the treatment group:

$$E(Y_0(0)|D = 1) = E(Y_0(1)|D = 1)$$

Assumption A2 may be violated if individuals, anticipating participation in a program, change their behavior before the program is implemented.

Assumption A3 (observability) Both $F_0\left(Y_0, D\right)$ and $F_1\left(Y_1, D\right)$ are observed.

Under Assumption A3, individuals can be classified at each time period into treated or control. This assumption is needed in order to be able to identify each of the four conditional means in equation (3).

The theorem below is not unique to our analysis, but we present it here for completeness.

Theorem 1. *Let assumptions A1, A2, and A3 hold. The DID estimand defined in* (3) *identifies the ATT defined in* (2).

Proof. Rewriting expression (3), plugging in definition (1), and by assumptions A1, A2, and A3 obtains:

$$\begin{aligned} \theta &= \left[E\left(Y_1|D = 1\right) - E\left(Y_0|D = 1\right)\right] - \left[E\left(Y_1|D = 0\right) - E\left(Y_0|D = 0\right)\right] \qquad (4) \\ &= \left[E\left(Y_1\left(1\right)|D = 1\right) - E\left(Y_0\left(1\right)|D = 1\right)\right] - \left[E\left(Y_1\left(0\right)|D = 0\right) - E\left(Y_0\left(0\right)|D = 0\right)\right] \\ &= \left[E\left(Y_1\left(1\right)|D = 1\right) - E\left(Y_0\left(1\right)|D = 1\right)\right] - \left[E\left(Y_1\left(0\right)|D = 1\right) - E\left(Y_0\left(0\right)|D = 1\right)\right] \\ &= \left[E\left(Y_1\left(1\right)|D = 1\right) - E\left(Y_1\left(0\right)|D = 1\right)\right] - \left[E\left(Y_0\left(1\right)|D = 1\right) - E\left(Y_0\left(0\right)|D = 1\right)\right] \\ &= E\left(Y_1\left(1\right)|D = 1\right) - E\left(Y_1\left(0\right)|D = 1\right) \\ &= ATT \end{aligned}$$

□

2.2 DID with Missing Status

The setting in this paper is such that assumption A3 does not hold. In our set-up, we observe repeated-cross sections drawn from the same population such that we observe $F_1(Y_1, D)$ and the marginal distribution $F_0(Y_0)$ but *not* the joint distribution $F_0(Y_0, D)$. Because of this data-limitation, the second term in the square brackets entering expression (3) is not identified without further assumptions.

Let Z be a time-invariant random variable with support $\mathcal{Z}$ that is observed in both periods. Define the propensity score at time t as

$$e_t(Z) \equiv F_t(D = 1|Z), \ t = 0, 1$$

Assumption A3' (missing group membership) Distribution functions $F_0(Y_0, Z)$ and $F_1(Y_1, D, Z)$ are observed, while $F_0(Y_0, Z, D)$ is not observed.

Assumption A4 (stationarity) The propensity score is stationary, i.e. for all $z \in \mathcal{Z}$

$$e_0(z) = e_1(z) \equiv e(z)$$

Assumption A4 holds automatically if the population of reference remains unchanged over time since D and Z are time-invariant. A4 is violated if the population changes significantly between the two time periods due to migration, births or deaths. For example, assume that the government implements a social program that targets an ethnic minority in a particular region. If the program induces a disproportionate immigration of members of this ethnic minority to the region, ethnicity cannot be used as a proxy for participation in the social program since the correlation structure between Z and D changes between the pre- and post-program periods.[9] This assumption is not specific to our setting however. Even when the treatment status is observed in all periods, migration, births and deaths affect the composition of both the treatment and the control groups threatening identification of the ATT via the standard DID method.

Assumption A5 (relevance) Z is informative about D, i.e. for $z_1 \neq z_2 \in \mathcal{Z}$

$$e(z_1) \neq e(z_2)$$

Assumption A5 requires the proxy variable Z and the treatment status D to be correlated. This assumption is testable by using information from the post-program period when $F_1(Y_1, Z, D)$ is observed.

Assumption A6 (conditional mean independence in changes) For all D and Z, the change over time in mean potential outcomes is independent of Z conditional on D:

$$E(Y_1(D)|D, Z) - E(Y_0(D)|D, Z) = E(Y_1(D)|D) - E(Y_0(D)|D)$$

[9] If the proportion of immigrants is known and one wants to make the strong assumption that immigrants are drawn from the same distribution as the residents, the method can still be applied because the change in the propensity score over time can be inferred.

Assumption A6 implies that, once we condition on the treatment status, Z contains no more information about the *change* over time in mean outcomes, i.e.

$$\begin{aligned} E(Y_1|D &= 1, Z) - E\left(Y_0|D=1,Z\right) = E(Y_1\left(1\right)|D=1,Z) - E\left(Y_0\left(1\right)|D=1,Z\right) \\ &= E(Y_1\left(1\right)|D=1) - E\left(Y_0\left(1\right)|D=1\right) = E(Y_1|D=1) - E\left(Y_0|D=1\right) \end{aligned}$$

Assumption A6 allows Z to affect the level of potential outcomes. For example, by letting the potential outcome be additively linear in Z, assumption A6 allows Z to have a time homogeneous time effect on the potential outcome, e.g.:

$$Y_t\left(D\right) = h_t\left(D\right) + g\left(Z\right) + \varepsilon_t,\ t = 0,1$$

where $h_t\left(.\right)$ is a function that can vary with time, while $g\left(.\right)$ is a time-invariant function. For a similar assumption see Lewbel (2007). The restriction here is that the effect of the proxy on the outcome does not change over time. If it did, it would not be possible to disentangle the effectiveness of the treatment from the effect of the proxy on the changes in outcomes over time. We illustrate this in the following section via a simple linear model.

Before continuing, we note a common misconception: Z need not be an instrumental variable. Importantly, Z need not be a cause of the treatment; in fact, Z may possibly be *driven* by the treatment. Notice that Z needs only be correlated with D and satisfy an exclusion restriction *conditional* on D. In contrast, an instrumental variable needs to satisfy a *joint* independence assumption, see e.g. condition 1 in Imbens and Angrist (1994). If Z were an instrumental variable, then for all Z and at each time period t, $\left(Y_t\left(0\right),\ Y_t\left(1\right),\ D\right)$ should be *jointly* independent of Z. Joint independence implies that, once we fix $Z = z$, the randomness that remains in both the potential outcomes and D is independent of Z. In contrast, assumption A6 allows Z to be correlated with both the unobservables in D and $Y_t\left(D\right)$. More importantly, any instrumental variable will satisfy assumption A6 but not all variables satisfying A6 may be instrumental variables. Assumption A6 is equivalent to the statement that Z is a surrogate of D as defined in the literature on measurement error, see e.g. Carroll et al. (2006) and Hu (2008).

Theorem 2 shows identification when treatment status is observed in only one period.

Theorem 2. *Suppose that Z takes on $K \geq 2$ different values, $\{z_k\}_{k=1}^{K} \in \mathcal{Z}$.*[10] *Additionally, let P be a $K \times 2$ matrix and Δ be a $K \times 1$ vector defined as, respectively:*

$$P \equiv \begin{bmatrix} 1 - e\left(z_1\right) & e\left(z_1\right) \\ ... & ... \\ 1 - e\left(z_K\right) & e\left(z_K\right) \end{bmatrix},\ \Delta \equiv \begin{bmatrix} E\left(Y_1|Z=z_1\right) - E\left(Y_0|Z=z_1\right) \\ ... \\ E\left(Y_1|Z=z_K\right) - E\left(Y_0|Z=z_K\right) \end{bmatrix} \tag{5}$$

[10] Our analysis allows for Z to be a continuous random variable. In this case, it would suffice to pick K different values from the support of Z.

Let assumptions A1, A2, A3', A4, A5, and A6 hold. Then the ATT defined by (3) *is identified, with the two differences of conditional means entering* (3) *given by:*

$$\begin{bmatrix} E\left(Y_1|D=0\right)-E\left(Y_0|D=0\right) \\ E\left(Y_1|D=1\right)-E\left(Y_0|D=1\right) \end{bmatrix} = \left(P'P\right)^{-1} P'\Delta \tag{6}$$

Proof. Define

$$\begin{aligned} \Delta E\left(Y|Z=z\right) &\equiv E\left(Y_1|Z=z\right)-E\left(Y_0|Z=z\right) \\ \Delta E\left(Y|D=d\right) &\equiv E\left(Y_1|D=d\right)-E\left(Y_0|D=d\right) \end{aligned}$$

Applying the law of total probability and using assumption A6 obtains the following mixture representation:

$$\begin{aligned} \Delta E\left(Y|Z\right) &= \Delta E\left(Y|Z,D=1\right)F\left(D=1|Z\right)+\Delta E\left(Y|Z,D=0\right)F\left(D=0|Z\right) \\ &= \Delta E\left(Y|D=1\right)e\left(Z\right)+\Delta E\left(Y|D=0\right)\left[1-e\left(Z\right)\right] \end{aligned}$$

Evaluating the expression above at $\{z_k\}_{k=1}^{K}$ obtains the following system of equations

$$\begin{bmatrix} \Delta E\left(Y|Z=z_1\right) \\ \dots \\ \Delta E\left(Y|Z=z_K\right) \end{bmatrix} = \begin{bmatrix} 1-e\left(z_1\right) & e\left(z_1\right) \\ \dots & \dots \\ 1-e\left(z_K\right) & e\left(z_K\right) \end{bmatrix} \begin{bmatrix} E\left(Y_1|D=0\right)-E\left(Y_0|D=0\right) \\ E\left(Y_1|D=1\right)-E\left(Y_0|D=1\right) \end{bmatrix} \tag{7}$$

Premultiplying (7) by P' obtains:

$$P'\Delta = P'P \begin{bmatrix} E\left(Y_1|D=0\right)-E\left(Y_0|D=0\right) \\ E\left(Y_1|D=1\right)-E\left(Y_0|D=1\right) \end{bmatrix}$$

Since $P'P$ is invertible, we obtain the following solution:

$$\begin{bmatrix} E\left(Y_1|D=0\right)-E\left(Y_0|D=0\right) \\ E\left(Y_1|D=1\right)-E\left(Y_0|D=1\right) \end{bmatrix} = \left(P'P\right)^{-1} P'\Delta$$

□

Remark 1. *Suppose that Z is discrete with two different values in its support, $z_1 \neq z_2 \in \mathcal{Z}$. Theorem 2 obtains an expression for the ATT given by:*

$$\theta = \frac{E\left(Y_1|Z=z_2\right)-E\left(Y_0|Z=z_2\right)-\left[E\left(Y_1|Z=z_1\right)-E\left(Y_0|Z=z_1\right)\right]}{e\left(z_2\right)-e\left(z_1\right)} \tag{8}$$

Notice that this expression is similar to that for the standard DID estimand where the treatment status has been replaced by the proxy. The DID term appearing in the numerator is weighted by the difference in propensity scores. It is this weight that accounts for the fact that Z is a proxy rather than the true treatment status. Without this weight, the DID estimand would be biased.

Remark 2. *When the proxy takes on two different values and when it is perfectly correlated with the treatment status, expression* (8) *collapses to that of the standard DID. Suppose that* $z \in \{0,1\}$ *such that*

$$P(D = d|Z = z) = \begin{cases} 1 \text{ \textit{if} } d = z \\ 0 \text{ \textit{if} } d \neq z \end{cases}$$

Then (8) *becomes*

$$\begin{aligned} & E(Y_1|Z=1) - E(Y_0|Z=1) - [E(Y_1|Z=0) - E(Y_0|Z=0)] \\ = \;& E[E(Y_1|D=1,Z=1) - E(Y_0|D=1,Z=1)] - E[E(Y_1|D=0,Z=0) - E(Y_0|D=0,Z=0)] \\ = \;& E(Y_1|D=1) - E(Y_0|D=1) - [E(Y_1|D=0) - E(Y_0|D=0)] \end{aligned}$$

This shows that in our set-up, we do not need an overlap condition to hold with respect to Z.[11]

If one is interested in the distributions $F_0(Y_0|D = d)$, $d = 0, 1$, the following theorem that replaces assumption A6 by A6' below shows that the distribution functions are identified.

Assumption A6' (independence) For all $z \in \mathcal{Z}$ and $d \in \{0,1\}$:

$$F(Y_1|D = d, Z = z) - F(Y_0|D = d, Z = z) = F(Y_1|D = d) - F(Y_0|D = d)$$

Theorem 3. *Let assumptions A1, A2, A4, A5, and A6' hold, and suppose that* Z *takes on* K *different values,* $\{z_k\}_{k=1}^{K} \in \mathcal{Z}$. *Then the distribution functions* $F_0(Y_0|D = d)$, $d \in \{0,1\}$ *are identified and given by*

$$\begin{bmatrix} F_1(Y_1|D=0) - F_0(Y_0|D=0) \\ F_1(Y_1|D=1) - F_0(Y_0|D=1) \end{bmatrix} = (P'P)^{-1} P' \begin{bmatrix} F_1(Y_1|Z=z_1) - F_0(Y_0|Z=z_1) \\ \ldots \\ F_1(Y_1|Z=z_K) - F_0(Y_0|Z=z_K) \end{bmatrix}$$

Proof. The proof parallels that of Theorem 2 and is not repeated here.

□

Remark 3. *In applied work, researchers usually condition on observed covariates,* X. *Our identification strategy carries through when conditioning on such covariates. In our framework, the main differences between proxies,* Z, *and covariates,* X, *are that (1) covariates need to satisfy a common support assumption, i.e.* $F_t(D = 1|X, Z) < 1$, *for* $t = 0, 1$ *and for all* X, Z, *and (2) covariates do not need to satisfy an exclusion restriction such as assumption A6. Formal assumptions needed for identification in the presence of observed covariates are included in the Appendix.*

[11] Due to the fact that Z is a proxy, $E(Y_1 - Y_0|D, Z)$ does not depend on Z. As such, proxies do not require overlap assumptions. In contast, letting X be a covariate (but not a proxy), then $E(Y_1 - Y_0|D, X)$ would depend on X, and the ATT would require an overlap assumption.

2.3 A Simple Example

To build intuition for our identification assumptions, as well as for why the ATT is not identified when the proxy is not accounted for, consider the following linear equation for the observed outcome:

$$Y_t = \alpha + \beta D + \gamma t + \tau Dt + \xi_t Z + \varepsilon_t,\ t = 0, 1 \tag{9}$$

where D and Z are binary random variables. The DID is given by τ.

We further assume the existence of a linear projection of D onto Z:

$$D = \delta Z + u,\ E\left(uZ\right) = 0 \tag{10}$$

which can be computed only with data from the post-treatment period.

Notice that we imposed the stationarity assumption on the propensity score (δ is time invariant), but that we did not impose the time-homogeneity of Z on the potential outcomes. We have not imposed the latter assumption since we want to illustrate why it is needed for our identification strategy.

Suppose that the treatment status is observed only in the post-program period, so that $E\left(Y_0|D = 1\right)$ and $E\left(Y_0|D = 0\right)$ are not observed. Evaluating expression (8) by using (9) and (10), obtains:

$$ATT_{proxy} = \frac{\left[E\left(Y_1|Z = 1\right) - E\left(Y_1|Z = 0\right)\right] - \left[E\left(Y_0|Z = 1\right) - E\left(Y_0|Z = 0\right)\right]}{E\left(D|Z = 1\right) - E\left(D|Z = 0\right)} = \tau + \frac{\xi_1 - \xi_0}{\delta}$$

since

$$\begin{aligned} E\left(Y_1|Z = 1\right) &= \alpha + \beta E\left(D|Z = 1\right) + \gamma + \tau E\left(D|Z = 1\right) + \xi_1 = \alpha + \beta\delta + \gamma + \tau\delta + \xi_1 \\ E\left(Y_1|Z = 0\right) &= \alpha + \beta E\left(D|Z = 0\right) + \gamma + \tau E\left(D|Z = 0\right) + \xi_1 = \alpha + \gamma \end{aligned}$$

Imposing the time homogeneity of Z, i.e. $\xi_1 = \xi_0$, obtains that $ATT_{proxy} = ATT$. Notice that when assumption A6 does not hold, the true ATT and the effect of the proxy over time on the potential outcomes cannot be disentangled.

On the other hand, what is currently done in practice is to use a proxy for D without accounting for the fact that the proxy is an imperfect measure of the treatment status. That is, researchers compute:

$$\begin{aligned} ATT_{bias} &= \left[E\left(Y_1|D = 1\right) - E\left(Y_0|Z = 1\right)\right] - \left[E\left(Y_1|D = 0\right) - E\left(Y_0|Z = 0\right)\right] \\ &= \tau + \beta\left(1 - \delta\right) + \xi_1\frac{1}{\delta} - \xi_0 \end{aligned}$$

Assuming that $\xi_0 = \xi_1$ is not enough for ATT$_{bias}$ to be unbiased. $ATT_{bias} = ATT$ if the proxy is a perfect measurement of the true treatment status, i.e. $\delta = 1$. This is a strong restriction, which is not likely to hold for all possible set-ups.

3 Estimation

In this section, we consider inference for (6) when the proxy is discrete with $K \geq 2$ points of support. First, our identification is constructive, in the sense that expression (6) can be estimated directly from the data. To this end, we use the insights of Angrist (1991) to show that the sample counterpart of (6) is a grouped-data estimator, which is equivalent to a two-stage least-squares estimator, which is further equivalent to a GMM estimator. Second, following the standard DID formulation, we also propose a GMM estimator based on a linear parametric model. We then show that these model-based moment conditions lead to a GMM overidentification test for our two main identifying assumptions: the stationarity of the propensity score and the exclusion restriction on the proxy.

3.1 A Two-Stage Least-Squares Estimator

Let there be two random samples: one in the first period, $\{Y_{0i}, Z_i\}_{i=1}^{n_0}$, and one in the second period, $\{Y_{1j}, Z_j, D_j\}_{j=1}^{n_1}$. The samples are drawn from the joint distribution of (Y_0, Y_1, Z, D).

Suppose Z is a discrete random variable with $K \geq 2$ points of support, i.e. $\mathcal{Z} = \{z_1...z_K\}$, and let n_{tk} be the number of individuals in cross-section t for whom $Z_i = z_k$, $k = 1, ..., K$, $i = 1, ..., n_t$. That is, letting $1(.)$ be the indicator function, n_{tk} is defined as:

$$n_{0k} \equiv \sum_{i=1}^{n_0} 1(Z_i = z_k), \; n_{1k} \equiv \sum_{j=1}^{n_1} 1(Z_j = z_k)$$

Then the estimators of $e(z)$ and of $E(Y_t|Z = z)$ are cell averages defined as, respectively:

$$\bar{e}_k \equiv \frac{1}{n_{1k}} \sum_{j=1}^{n_1} 1(Z_j = z_k) D_j \tag{11}$$

$$\overline{Y}_{1k} \equiv \frac{1}{n_{1k}} \sum_{j=1}^{n_1} 1(Z_j = z_k) Y_{1j} \tag{12}$$

$$\overline{Y}_{0k} \equiv \frac{1}{n_{0k}} \sum_{i=1}^{n_0} 1(Z_i = z_k) Y_{0i}$$

Define now the following matrix and vectors:

$$\widehat{P} \equiv \begin{bmatrix} 1-\bar{e}_1 & \bar{e}_1 \\ ... & ... \\ 1-\bar{e}_k & \bar{e}_k \end{bmatrix}, \; \overline{Y}_t \equiv \begin{bmatrix} \overline{Y}_{t1} \\ ... \\ \overline{Y}_{tk} \end{bmatrix}, \; \widehat{\Delta} \equiv \overline{Y}_1 - \overline{Y}_0$$

Then the sample counterpart of (6) is

$$\left(\widehat{P}'\widehat{P}\right)^{-1} \widehat{P}'\widehat{\Delta} \tag{13}$$

When Z takes on more than two different values, any two values can be used to obtain an estimate for the ATT. When Z takes on K different values, it is possible to obtain $K - 1$ independent estimators (13) for the ATT. We then introduce a $K \times K$ positive definite matrix W that allows us to combine alternative estimates of the same parameter into one estimate, that is:

$$\left(\widehat{P}'W\widehat{P}\right)^{-1}\widehat{P}'W\widehat{\Delta} \tag{14}$$

Estimator (14) represents a grouped-data estimator. In fact, using the language of Angrist (1991), estimator (14) is a linear combination of independent pairwise Wald estimators of the same parameter. Angrist (1991) shows that when W is a $K \times K$ diagonal matrix with the element on row k representing the total number of observations for which $Z = z_k$.

In our set-up, the number of individuals with $Z = z_k$ may be different at each time period. Thus, we introduce two weight matrices, W_1 and W_0, where W_t is a $K \times K$ diagonal matrix with the element on row k representing the total number of observations in sample t with $Z = z_k$, i.e. the diagonal elements of W_t are $\{n_{tk}\}_{k=1}^{K}$, $t = 0, 1$. Then our sample counterpart of (6) is given by:

$$\begin{bmatrix} \widehat{m}_0 \\ \widehat{m}_1 \end{bmatrix} = \left(\widehat{P}'W_1\widehat{P}\right)^{-1}\widehat{P}'W_1\overline{Y}_1 - \left(\widehat{P}'W_0\widehat{P}\right)^{-1}\widehat{P}'W_0\overline{Y}_0 \tag{15}$$

where $\widehat{m}_d$ is the estimator of $E(Y_1|D = d) - E(Y_0|D = d)$ for $d \in \{0, 1\}$. The right hand side of (15) represents the difference of two grouped-data estimators, one for each cross-section.

By the same arguments as in Angrist (1991), it can be shown that the grouped-data estimator (15) is equivalent to the following two-stage estimator:

$$\begin{bmatrix} \widehat{m}_0 \\ \widehat{m}_1 \end{bmatrix} = \left(\widehat{\Pi}_1'\widehat{\Pi}_1\right)^{-1}\widehat{\Pi}_1'Y_1 - \left(\widehat{\Pi}_0'\widehat{\Pi}_0\right)^{-1}\widehat{\Pi}_0'Y_0 \tag{16}$$

where $\widehat{\Pi}_t$ is a matrix of estimated propensity scores for the n_t individuals observed in period t.[12]

$$\widehat{\Pi}_t \equiv \begin{bmatrix} 1 - \widehat{e}_1(z_1) & \widehat{e}_1(z_1) \\ ... & ... \\ 1 - \widehat{e}_{n_t}(z_{n_t}) & \widehat{e}_{n_t}(z_{n_t}) \end{bmatrix} \tag{17}$$

Finally, the estimator for the DID estimand in (3) is given by

$$\widehat{\theta} = \widehat{m}_1 - \widehat{m}_0 \tag{18}$$

The first term on the right hand side of (16) is the formula for the standard two-step least squares estimator using the post-treatment sample. The first stage consists of estimating the propensity scores to form $\widehat{\Pi}_1$ and the second stage consists of projecting Y_t onto the space generated by the estimated propensity

[12] The propensity score (first stage) is computed regressing D on K-1 regressors (plus a constant). The regressors are dummy variable indicators for each value that Z takes. If Z takes on K values, then only K-1 linearly independent variables can be included.

scores, $\widehat{\Pi}_t$. The second term on the right hand side of (16) is similar to the first term with the caveat that the first stage and the second stage are computed using different samples. Since the treatment status is observed only at $t = 1$, the propensity scores are estimated using information exclusively from that period. Thus, $\widehat{\Pi}_0$ is a matrix of generated regressors containing the predicted values of the propensity score. The second term in (16) is the formula for the two-sample two-stage least squares estimator, see e.g. Angrist and Krueger (1992) and Inoue and Solon (2010).[13] The two-stage procedure ignores that $\widehat{\Pi}_t$ are generated regressors, which gives incorrect standard errors. We introduce below a GMM characterization of the two-stage procedure which does not run into this issue.

3.2 A Just-Identified GMM Model

As in the literature on data combination and on GMM estimation with missing data,[14] we define T_i to be an indicator for sample membership:

$$T_i = \begin{cases} 1 \text{ if individual i is observed at } t = 1 \\ 0 \text{ if individual i is observed at } t = 0 \end{cases} \tag{19}$$

such that this dummy variable satisfies the assumption below:

Assumption A7 (missing completely at random) For each i, T_i is independent of

$$Y_{1i}\left(D_i\right), Y_{0i}\left(D_i\right), D_i$$

Assumption A7 is easily justified in our set-up due to the cross-sectional design of data collection.

Introducing the sample membership indicator allows us to combine the pre- and post-treatment samples and to treat the combined data as i.i.d. draws from $\left(T_i Y_{1i}, \left(1 - T_i\right) Y_{0i}, T_i D_i, Z_i\right),\ i = 1, ..., n = n_0 + n_1$. The sample size allocated to period 1, i.e. $\sum_{i=1}^{n} T_i$ is now a random variable, introducing T_i allows us to express our moment conditions in a way that is consistent with GMM estimation and to apply standard arguments showing our asymptotic results.

Note that now, Z_i is a vector of $K - 1$ dummy variables and a constant. As explained in section 3.1 and following Angrist (1991), the regressors in Z_i are dummy variables indicators for each value that Z takes

[13] In summary, expression (16) can be computed via the following two-stage procedure:

Step 1: Estimate a model for the propensity score using the sample from $t = 1$. For example, if a probit model is used for the first stage, estimate γ from the model $P(D = 1|Z) = \Phi(Z\gamma)$. Then generate the propensity score for each observation i at $t = 1$ and $t = 0$ using the model previously estimated. e.g., for each observation in $t = 1$ generate $\widehat{e}(z_i) = \Phi(Z_i\widehat{\gamma})$, and for each observation in $t = 0$ generate $\widehat{e}(z_i) = \Phi(Z_i\widehat{\gamma})$. Then generate the matrices $\widehat{\Pi}_t$, $t = 0, 1$.

Step 2: Regress Y_t on $\widehat{\Pi}_t$ (no constant), $t = 0, 1$. The regression for $t = 1$ yields the first term on the right hand side of (16) and the regression for $t = 0$ yields the second term. Subtract the estimated coefficients to obtain the vector $(\widehat{m_0}, \widehat{m_1})'$ and then compute $\widehat{\theta} = \widehat{m_1} - \widehat{m_0}$.

[14] See for example Chen, Hong, and Tarozzi (2008), Devereux and Tripathi (2009), Graham (2011), Abrevaya and Donald (2011), Muris (2013).

(minus one to avoid multicollinearity). For example, if the proxy is birth year, Z_i is defined as the set of dummy variables indicating the year when the person was born. Nonetheless, it may be possible to estimate a more parsimonious model and include year of birth in a linear fashion, i.e. $Z_i = [1, \text{year of birth}]$. In this case the dimension of Z_i may be much lower than the number of values in the proxy, $dim(Z) < K$.

We now parametrize the propensity score $e(Z;\gamma)$ by assuming that it is known up to a finite set of parameters, $\gamma \in \Gamma \subset \mathbb{R}^{d_\gamma}$. For each individual i, let

$$\pi(Z_i, \gamma) \equiv [1 - e(Z_i, \gamma), e(Z_i, \gamma)] \tag{20}$$

and let $e_\gamma(Z, \gamma) = \frac{\partial e(Z,\gamma)}{\partial \gamma}$.

Let $(\gamma^*, \eta_1^*, \eta_2^*)'$ be a $d_\gamma + 4$ vector of true parameters. Our estimator can be obtained from the following moment conditions:

$$E\left[g\left(Y_{1i}, Y_{0i}, Z_i, D_i, T_i; \gamma^*, \eta_1^*, \eta_0^*\right)\right] \equiv E\begin{bmatrix} T_i e_\gamma(Z_i, \gamma^*)\frac{D_i - e(Z_i,\gamma^*)}{e(Z_i,\gamma^*)(1-e(Z_i,\gamma^*))} \\ T_i \pi(Z_i, \gamma^*)'(Y_{1i} - \pi(\gamma^*, Z_i)\eta_1^*) \\ \left(1 - T_i\right)\pi(Z_i, \gamma^*)'(Y_{0i} - \pi(\gamma^*, Z_i)\eta_0^*) \end{bmatrix} = 0 \tag{21}$$

where the first moment condition is the score function of a likelihood function corresponding to the first stage, while the remaining two moment conditions correspond to the second stage, derived as the linear projections of Y_t onto the space generated by the estimated propensity scores.

The estimates of interest used to compute the DID, see equation (18), are obtained as:

$$\begin{bmatrix} \widehat{m}_0 \\ \widehat{m}_1 \end{bmatrix} = \widehat{\eta}_1 - \widehat{\eta}_0$$

The estimator based on the moment conditions in (21) is just-identified since there are $d_\gamma + 4$ moment conditions and parameters. This estimator is numerically identical to expression (16).

We introduce next a second GMM estimator, which may be used to perform a GMM overidentification test for model specification.

3.3 An Over-Identified GMM Model

In this section, we propose a GMM estimator based on a parametric model for the outcome equations. This estimator can be used to perform a GMM overidentification test for our main identification assumptions - the stationarity of the propensity score and the exclusion restriction on the proxy. Since the GMM overidentification test is a test of correct model specification, a rejection would indicate that some identification assumptions or orthogonality conditions do not hold.

The GMM estimator derived from the moment conditions in (21) is just-identified. By projecting the outcomes onto the space of proxies, we can take advantage of the dimension of $Z_i = [1, Z_{i1}, Z_{i2}, ..., Z_{iK-1}]$, and obtain an estimator that may be overidentified. For this procedure, the dependence between Y and Z is modelled directly.

As it is standard in the DID literature, we specify the following linear model for the outcomes in the two time periods:[15]

$$Y_{it} = \beta_{0t}\left(1 - D_i\right) + \beta_{1t} D_i + Z_i \delta + \varepsilon_{it}, \qquad \varepsilon_{it} \sim iid,\ i = 1, ..., n_t,\ t = 0, 1$$

where, without loss of generality, we assume that $\delta_0 = 0$ in the linear expression $Z_i\delta = \delta_0 + \delta_{i1} Z_{i1} + ... + \delta_{ik-1} Z_{ik-1}$. This assumption is innocuous.[16]

The DID estimand implied by this model is

$$\theta = (\beta_{11} - \beta_{01}) - (\beta_{10} - \beta_{00})$$

Let $d_i = (1 - D_i, D_i)$ and define the vector

$$\beta_t \equiv \begin{bmatrix} \beta_{0t} \\ \beta_{1t} \end{bmatrix}, t = 0, 1$$

For notational convenience define $\alpha^* = (\gamma^*, \delta^*, \beta_1^*, \beta_0^*)'$ be the $4 + d_\gamma + K$ vector of true parameters, where K is the dimension of Z_i, and let $\mathcal{A} = \Gamma \times \mathcal{D} \times \mathcal{B} \subset \mathbb{R}^{d_\gamma} \times \mathbb{R}^K \times \mathbb{R}^4$ be the parameter space, where $\Gamma, \mathcal{D}, \mathcal{B}$ are compact.

The moment conditions for our over-identification test are:

$$\begin{aligned} E\left[g\left(Y_{0i}, Y_{1i}, Z_i; \alpha^*\right)\right] \quad &\equiv \quad E\begin{bmatrix} T_i e_\gamma(Z_i, \gamma^*) \frac{D_i - e(Z_i,\gamma^*)}{e(Z_i,\gamma^*)(1 - e(Z_i,\gamma^*))} \\ T_i (Z_i, D_i)' \left(Y_{1i} - d_i' \beta_1^* - Z_i \delta^*\right)) \\ \left(1 - T_i\right) Z_i \left(Y_{0i} - \pi(Z_i, \gamma^*) \beta_0^* - Z_i \delta^*\right)) \end{bmatrix} \qquad (22) \\ &\equiv \quad E\begin{bmatrix} g_1\left(T, Z, D; \gamma^*\right) \\ g_2\left(T, Z, D, Y_1; \delta^*, \beta_1^*\right) \\ g_3\left(T, Z, Y_0; \gamma^*, \delta^*, \beta_0^*\right) \end{bmatrix} = 0 \end{aligned}$$

where T_i is the sample indicator defined in (19) and $\pi(Z_i, \gamma)$ is defined in (20) . There are $2K + 2 + d_\gamma$ moment conditions and $4 + d_\gamma + K$ parameters. Thus, when $K > 2$ (Z includes the unity as the first component)

[15] We note here that Z can affect Y in a non-linear way, e.g.

$$Y_{it} = \beta_{0t}\left(1 - D_i\right) + \beta_{1t} D_i + g\left(Z, \delta\right) + \varepsilon_{it}$$

where $g\left(Z, \delta\right)$ is any function known up to finite dimensional parameter, δ.

[16] Doing some algebra, the outcome equation can be written as $Y_{it} = (\beta_{0t} + \delta_0) + (\beta_{1t} - \beta_{0t}) D_i + \delta_1 Z_{i1} + ... + \delta_{k-1} Z_{ik-1} + \varepsilon_{it}$ Then, it is evident that β_{0t} and δ_0 are not separately identified. Nonetheless, even when $\delta_0 \neq 0$, the estimate of the difference $(\beta_{1t} - \beta_{0t})$ used to compute the DID is not affected.

there are $K-2$ overidentifying moment conditions, which can be used to test jointly our stationarity and exclusion restriction assumptions.[17]

Define the sample analogue of the population moments as:

$$\overline{g}\left(\alpha\right)=\frac{1}{n}\sum_{i=1}^{n}g\left(Y_{0i},Y_{1i},Z_{i};\alpha\right)$$

The GMM estimator of α^{*} is then given by

$$\widehat{\alpha}=\arg\min_{\alpha\in\mathcal{A}}\overline{g}'\left(\alpha\right)\widehat{V}\overline{g}\left(\alpha\right)$$

where $\widehat{V}$ converges in probability to V, the optimal positive semidefinite weighting matrix, which we define below.

Let $||.||$ denote the Euclidean norm, and let $B_{r}\left(\alpha\right)\subset\mathcal{A}$ be the $4+d_{\gamma}+K$ open ball of radius r with center at a.

Assumption A8

(i) $\mathcal{A}$ is compact; (ii) $\alpha^{*}\in int\left(\mathcal{A}\right)$;

(iii) $e\left(Z,\gamma\right),\ e_{\gamma}\left(Z,\gamma\right),\pi\left(Z,\gamma\right)$ are continuous at each $\gamma\in\mathcal{A}$ w.p.1;

(iv) $e\left(Z,\gamma\right),\ e_{\gamma}\left(Z,\gamma\right),\pi\left(Z,\gamma\right)$ are continuously differentiable on $B_{r}\left(\alpha^{*}\right)$ for some $r>0$ w.p.1;

(v) $E\left(\sup_{\alpha\in\mathcal{A}}||g\left(Y_{0i},Y_{1i},Z_{i};\alpha\right)||^{2}\right)<\infty$;

(vi) $E\left(g\left(Y_{0i},Y_{1i},Z_{i};\alpha^{*}\right)g'\left(Y_{0i},Y_{1i},Z_{i};\alpha^{*}\right)\right)$ is nonsingular;

(vii) $E\left(\frac{\partial}{\partial\alpha}g\left(Y_{0i},Y_{1i},Z_{i};\alpha^{*}\right)\right)$ is of full-column rank;

(viii) $E\left(\sup_{\alpha\in B_{r}\left(\alpha^{*}\right)}\left|\left|\frac{\partial}{\partial\alpha}g\left(Y_{0i},Y_{1i},Z_{i};\alpha\right)\right|\right|\right)<\infty$ for some r.

Assumptions A7 and A8 are sufficiently strong to guarantee consistency and asymptotic normality of $\widehat{\alpha}$, see Theorems 2.6 and 3.4 in Newey and McFadden (1994). The optimal weighting matrix, V, is the inverse of the following covariance matrix:

$$V=\begin{bmatrix}V_{11} & V_{12} & 0\\ V_{21} & V_{22} & 0\\ 0 & 0 & V_{33}\end{bmatrix}$$

[17] Given the stationarity assumption, if the exclusion restriction does not hold, δ (or the function $h\left(Z,\delta\right)$ when the model is not linear) would not be time-invariant so that $E\left[Z_{i}\left(Y_{0i}-\pi_{0i}'(\gamma)\beta_{0}-Z_{i}\delta\right)\right]\neq0$.

where

$$\begin{aligned}
V_{11} &= Eg_1\left(T,Z,D;\gamma^*\right)g_1'\left(T,Z,D;\gamma^*\right) \\
V_{12} &= Eg_1\left(T,Z,D;\gamma^*\right)g_2'\left(T,Z,d,Y_1;\delta^*,\beta_1^*\right) \\
V_{21} &= Eg_2\left(T,Z,d,Y_1;\delta^*,\beta_1^*\right)g_1'\left(T,Z,D;\gamma^*\right) \\
V_{22} &= Eg_2\left(T,Z,d,Y_1;\delta^*,\beta_1^*\right)g_2'\left(T,Z,d,Y_1;\delta^*,\beta_1^*\right) \\
V_{33} &= Eg_3\left(T,Z,Y_0;\gamma^*,\delta^*,\beta_0^*\right)g_3'\left(T,Z,Y_0;\gamma^*,\delta^*,\beta_0^*\right)
\end{aligned}$$

Finally, the overidentification test associated with the GMM estimator presented in this subsection is the standard J test. Under the null hypothesis of correct model specification, the test statistic has a χ^2 distribution with $K-2$ degrees of freedom.

4 Monte Carlo Simulations

The goal of this section is two-fold. First, we evaluate the small-sample performance of our GMM estimators based on two different sets of moment conditions, those in (21) and those in (22). Second, we aim to compare our estimators to (1) the standard DID estimator, which is infeasible in practice under assumption A3', but which can be computed in simulation studies, and (2) the DID estimator that uses the proxy without accounting for the fact that it is not the true status.[18]

We consider the following simple data generating process, where 1(.) is the indicator function.

For $i=1,...,n=n_0+n_1$, we first generate two binary random variables, Z_1 and Z_2 :

$$\begin{aligned}
Z_{1i} &= 1\left(\omega_{1i}>0\right) \\
Z_{2i} &= 1\left(\omega_{2i}>0\right) \\
\left(\omega_1,\omega_2\right)' &\sim N\left(\left(0,0\right)',\begin{bmatrix}1 & 0.5\\ 0.5 & 1\end{bmatrix}\right)
\end{aligned}$$

The number of observations, n_0 and n_1, is random and it is determined by the indicator T_i. For $i=1,...,n=n_0+n_1$:

$$T_i \sim Bernoulli\left(\kappa\right),\ \kappa\in\{0.2,0.5,0.8\}$$

which obtains that $E\left(n_1\right)=\kappa n$ and $E\left(n_0\right)=\left(1-\kappa\right)n.$

The data generating process for the pre-treatment period, $t=0$, is the following. For $i=1,...,n_0$:

$$\begin{aligned}
D_i^* &= 1\left(\alpha\left(Z_{1i}+Z_{2i}\right)+\left(1-\alpha\right)\epsilon_{i0}^D>0\right),\ \alpha\in(0,1] && (23)\\
Y_{0i}\left(D^*\right) &= 1+D_i^*+\gamma Z_{2i}+\epsilon_{i0}^Y && (24)\\
\left(\epsilon_{i0}^D,\epsilon_{i0}^Y\right)' &\sim N\left(\left(0,0\right)',I\right) && (25)
\end{aligned}$$

[18] As explained in the introduction, this is a common practice in the applied literature dealing with the case of repeated-cross sections with unobserved treatment status, see e.g. Groen and Polivka (2008).

where (23) specifies the model for the treatment status and (24) specifies the model for the pre-program potential outcomes. Our relevance assumption on the proxy is captured by the restriction on α. That is, when $\alpha = 0$, the relevance assumption does not hold since Z_1, Z_2, and D are independent. Since γ is time-invariant, the proxy satisfies our exclusion restriction, assumption A6. D_i^* is not observed by the econometrician under assumption A3'.

The data generating process for the post-treatment period, $t = 1$, is given by the model below. For $i = 1, ..., n_1$:

$$\begin{aligned} D_i &= 1\left(\alpha\left(Z_{1i} + Z_{2i}\right) + \left(1 - \alpha\right)\epsilon_{i1}^D > 0\right),\ \alpha \in (0, 1] && (26) \\ Y_{1i}\left(D\right) &= 2 + 2D_i + \gamma Z_{2i} + \epsilon_{i1}^Y && (27) \\ \left(\epsilon_{i1}^D, \epsilon_{i1}^Y\right)' &\sim N\left(\left(0, 0\right)', I\right) && (28) \end{aligned}$$

where (26) models the treatment status and (27) models the post-program potential outcomes. Notice that our stationarity assumption is satisfied, see expressions (23) and (26). The model for the post-treatment period is similar to the pre-treatment period but the treatment status D_i is observed.

The DID status implicitly defined by the equations for the potential outcomes is $\tau = 1$.

We present results for different values of $\alpha \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$, for different values of $\kappa \in \{0.2, 0.5, 0.8\}$, and for $n = n_0 + n_1 = 2000$.[19] We fix $\gamma = 10$ in all our simulation designs (we found that our results are not sensitive to the value of γ). We perform 2000 replications for each design.

The results of our simulations are reported in Table 7 in the Appendix, and are summarized in the figures below. The estimators we consider are:

1. Infeasible DID, which is the sample analogue of (3). This estimator can be computed in simulation studies, since we observe D*. However, the estimator is infeasible in the type of applications we have in mind, since D* is unobserved.

2. GMM 1, which is the estimator based on moment conditions (21).

3. GMM 2, which is the DID estimand based on the moment conditions in (22).

4. Naive DID - the estimator that has been used in practice, which uses a proxy for the latent treatment status but which does not account for the proxy, i.e. instead of computing the sample analogue of (3), we compute the sample analogue of:

$$\theta_{naive} = \left[E\left(Y_1|D=1\right) - E\left(Y_0|Z_2=1\right)\right] - \left[E\left(Y_1|D=0\right) - E\left(Y_0|Z_2=0\right)\right]$$

[19] We also considered a much larger sample size, $n = 5000$. The results are the same as those we present, the only difference being that we obtain even smaller RMSEs for our two GMM estimators.

For each estimator, we compute the average bias, average standard deviation, and root mean square error (RMSE). Table 7 in the Appendix shows all these different statistics, while figures 1, 2, and 3 below show the RMSE of the estimators as a function of α for each of the three different values of κ.

All conditional means for both the standard DID estimand and for θ_{naive} are computed as cell averages. For example, the estimators of $E\left(Y_1|D=1\right)$ and $E\left(Y_0|Z_2=1\right)$ are given by, respectively:

$$\begin{aligned}\widehat{E}\left(Y_1|D=1\right) &= \frac{1}{n_1}\sum_{j=1}^{n_1} 1\left(D_j=1\right)Y_{1j}\\ \widehat{E}\left(Y_0|Z_2=1\right) &= \frac{1}{n_0}\sum_{i=1}^{n_0} 1\left(Z_{2i}=1\right)Y_{0i}\end{aligned}$$

Additionally, when computing the infeasible DID estimator, we include the two proxies, Z_1 and Z_2, in the outcome equations.

Our results show that (1) the naive estimator performs very poorly. It is heavily biased across different specifications of α and κ, although the bias decreases as α increases. And (2) our proposed estimators, GMM 1 and GMM2, are robust in terms of RMSE across different specifications of α and κ, with the RMSE quickly approaching that of the infeasible DID as α increases.

The two estimators GMM1 and GMM2 are comparable in terms of RMSEs, with GMM2 having a slightly smaller RMSE for smaller values of α and of κ. For values of $\alpha \geq 0.3$, the RMSE of the two estimators approach the RMSE of the infeasible estimator. When $\alpha \in \{0.1, 0.3\}$, the R^2 of a first stage regression of the treatment status observed in the post-program period on the two proxies, Z_1 and Z_2, is between 2% and 18%, and the correlation between the true treatment status and the predicted treatment status is between 14% and 40%.[20] For $\alpha \geq 0.7$, the first-stage R^2 is greater than 60%, when the RMSEs of our estimators approach the RMSE of the infeasible estimator. For example, when $\alpha = 0.9$, the first-stage R^2 is approximately 66%, and the correlation between the true treatment status and the predicted status is approximately 80%. For this case, our proposed estimators, GMM1 and GMM2, recover the RMSE of the infeasible estimator.

5 Empirical Illustration: PROGRESA

In this section, we estimate the impact of the Mexican conditional cash transfer program PROGRESA on the school attendance of eligible children between the ages of 13-15. The aim of this section is twofold. First, since the PROGRESA data set is a panel, our goal is to illustrate the performance of our proposed estimators relative to the standard DID. Second, we aim to show that the proxy variables can be chosen in a simple way when there is either partial or full information about the program's eligibility criteria. For detailed studies about PROGRESA, see Attanasio, Meghir, and Santiago (2012), Schultz (2004), Skoufias

[20] In empirical work, applied researchers should select proxies that have higher correlations with the treatment status that is observed in the post-program period.

Figure 1: Simulation Results: RMSE for n = 2000 and $\kappa = 0.2$

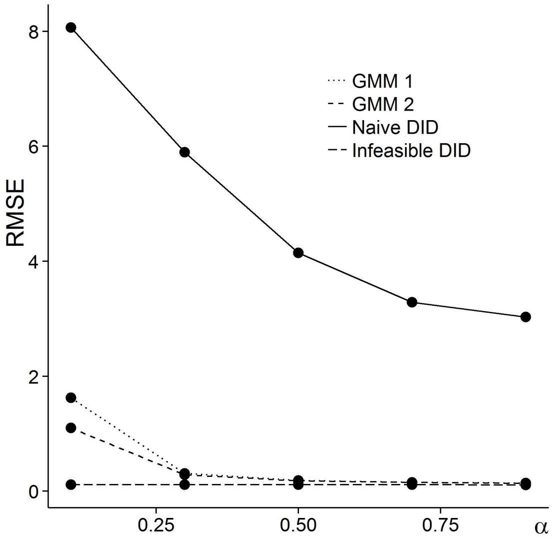


Figure 2: Simulation Results: RMSE for n = 2000 and $\kappa = 0.5$

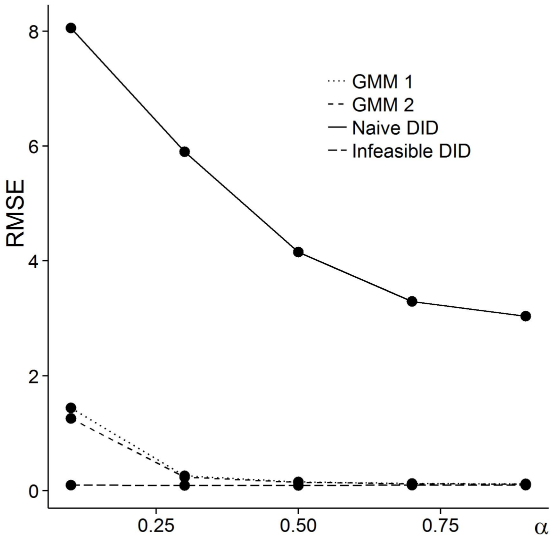
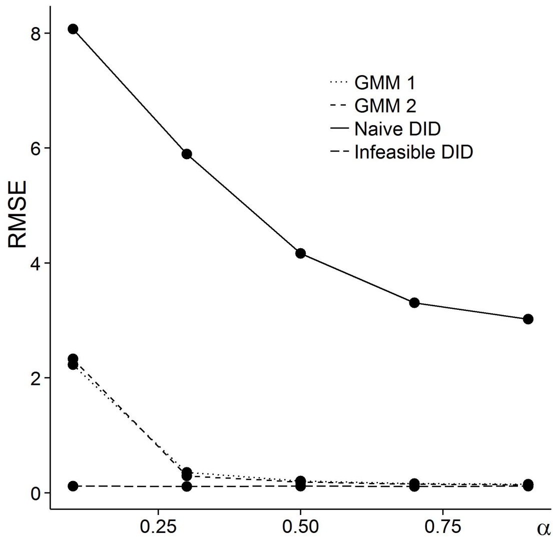


Figure 3: Simulation Results: RMSE for n = 2000 and $\kappa = 0.8$



(2001), Skoufias et al. (2001), Todd and Wolpin (2006), and to some extent Lee and Shaikh (2014).

5.1 Description of PROGRESA

The PROGRESA program, later renamed Oportunidades, was an anti-poverty program introduced in Mexico in May 1998. The program was aimed at improving the educational and health status of poor families. The program consisted of cash transfers for each grade-eligible child who attended at least 85% of the school days. The cash transfers targeted children in the last four grades of primary school and the first three grades of secondary school. The amount of the grant increased with the grade, being larger for girls in secondary school.

Eligibility for the program was determined in two phases. The first phase consisted of geographical targeting: a national census was used in order to create a marginality index, which was then used to select the villages where the extremely poor were more likely to be found. Based on this index, 506 villages were deemed "sufficiently poor." Each of the 506 villages was then selected for treatment with a probability of two-thirds, resulting in 320 treatment villages and 186 control villages.[21] In the second phase, a poverty index was created by proxy means testing using data on household income and housing characteristics. On the basis of this index, households within both treatment and control villages were classified as eligible or not for the program. Almost all eligible households in treatment villages participated. In the fall of 2000, the program was extended to control villages.

[21] There is random allocation at the village level.

The data set We use the set of surveys conducted by the government of Mexico, Sedesol (2006). Following Schultz (2004), we use the two waves prior to the intervention, ENCASEH1997 and ENCEL1998 March, and the three waves post-intervention, ENCEL1998 October, ENCEL1999 March, and ENCEL1999 November. The fives waves form a panel of individuals residing in both treatment and control villages.

The sample is partitioned in four groups based on place of residence - treatment or control village, and eligibility status in the program. Our sample consists of 38,984 teenager-year observations. Table 1 presents summary statistics. The fraction of children in treatment localities is 60% and the school attendance rate is 67%. In our sample, 65% of the children belong to an eligible household, 38% have access to piped water at home, 64% have a restroom exclusive for household members and 76% have electricity. The houses where they live have on average 2 rooms and in 70% of the cases one of the residents own land for cultivation.

Table 1: Descriptive statistics: Average and standard deviation (in parentheses).

Variable	Treatment villages	Control villages
	N = 23,508	N = 15,476
Attendance	0.7	0.64
	(0.461)	(0.48)
Elligible	0.665	0.625
	(0.472)	(0.484)
Post-treatment observations	0.63	0.62
	(0.484)	(0.485)
Water access	0.41	0.34
	(0.492)	(0.473)
Restroom access	0.63	0.66
	(0.484)	(0.474)
Electricity	0.75	0.79
	(0.434)	(0.407)
Number of rooms	1.98	1.97
	(1.143)	(1.101)
Own land	0.71	0.69
	(0.453)	(0.463)

5.2 Difference-in-differences

The way the program was implemented lends itself naturally to a DID evaluation. First, we use household information in treatment villages to (1) calculate the standard DID estimand and (2) calculate our proposed

DID estimand when the panel structure is ignored. Second, we use household information in control villages to test the parallel trend assumption required by our DID analysis. Control and treatment villages are expected to be ex-ante similar due to the randomization at the village level.

Standard DID Treatment among households within treatment villages was not randomly allocated. The post-intervention difference between eligible and ineligible children captures both the impact of PROGRESA and selection bias. The DID estimator eliminates the selection bias by subtracting the pre-intervention difference.

Table 2 shows the pre- and post-intervention difference between eligible and non-eligible children in treatment and control villages. This table can be used to calculate the estimate for the standard DID:

$$\widehat{DID}_P \equiv 0.0175 - (-0.0742) = 0.0917 \ (0.013)$$

The standard DID estimate indicates that PROGRESA increased school attendance by 9.2 percentage points among teenagers 13 to 15 years old.

Table 2 can also be used to verify the parallel trends assumption. Since villages were randomly allocated to receive the program, we can use control villages to analyze the counterfactual of what would have happened had treatment villages not received the program. As it can be seen, the gap between eligible and ineligible children in control villages is about 3.9% points both before and after the PROGRESA was implemented. The DID estimate computed in control villages is very small (0.0007) and not statistically different from zero, which is consistent with the parallel-trends assumption.

5.3 DID when eligibility is not observed

In order to apply our method, we delete the eligibility indicator in the pre-intervention period. Our method relies on a set of proxies that is observed in both time periods and that correlates with the eligibility indicator.

Proxy for latent eligibility status The PROGRESA methodological note indicates that a poverty index was used in order to determine eligibility. The information provided by the note is that the poverty index was generated by proxy means testing using household income and household characteristics that were correlated with poverty. The note does not contain any information about which exact method and characteristics were used in the generation of the poverty index, making it impossible to replicate the index.[22] Then we use as proxies variables that are generally believed to be correlated with poverty, such as: the number of rooms

[22] Although the poverty index is available in the dataset, we are not using it for our analysis since in the vast majority of cases there is no such information. Our goal is to follow a general approach that can be used to evaluate other social programs when longitudinal data are not available.

Table 2: Average school attendance

	Treatment villages		
	Eligible household	Ineligible household	Difference
Pre-program	0.62	0.695	-0.0742
	(0.0065)	(0.0083)	(0.0106)
Post-program	0.729	0.712	0.0175
	(0.0045)	(0.0066)	(0.0079)

	Control villages		
	Eligible household	Ineligible household	Difference
Pre-program	0.607	0.647	-0.0397
	(0.0083)	(0.0098)	(0.0129)
Post-program	0.636	0.675	-0.0390
	(0.0061)	(0.0079)	(0.0101)

Standard error in parentheses.

in the household, and indicator variables for household access to piped water, restroom, electricity, and for whether the household owns land.

Table 3 shows descriptive statistics of these variables. The selected proxies are good predictors of eligibility: eligible children live in houses that are less likely to have access to piped water, sewage system, and electricity, and they live in households that own smaller homes in terms of number of rooms and that are less likely to own land.

Table 4 shows the results of our estimation procedure. We use different specifications to test the robustness of our method. Column 1 indicates the proxy variable excluded when computing our GMM estimators, e.g. "none" indicates that all five proxies were used in the estimation. Column 2 indicates the method we used to compute the propensity score. Columns 3 and 4 show the results when moment conditions (21) were used, while columns 5 and 6 show the results with moment conditions (22). Finally, columns 7 and 8 show the results of our overidentification test.

A few results are worth mentioning. As column 3 suggests, the functional form of the first stage does not affect our results significantly. Using a linear probability model, a probit or a logit model in the first stage yields very similar results in the second stage. However, our estimators are sensitive to the choice of proxy variables. For example, when the number of rooms in the house is excluded, the ATT obtained by our method is much lower than the standard DID. This result is not surprising given that the number of rooms in the house is a strong proxy for program eligibility. Notice also that when we exclude this variable,

Table 3: Proxies: Descriptive statistics

Variable	Eligible N = 15621	Ineligible N = 7863	Difference
Piped water	0.33	0.48	-0.15
	(0.003)	(0.004)	(0.0051)
Sewage	0.6	0.71	-0.10
	(0.003)	(0.004)	(0.0051)
Electricity	0.68	0.91	-0.23
	(0.003)	(0.002)	(0.0043)
Number of rooms	1.75	2.4	-0.65
	(0.006)	(0.011)	(0.0011)
Own land	0.68	0.74	-0.06
	(0.003)	(0.004)	(0.0048)

Standard error in parentheses.

the F-statistic of the first stage is the lowest. When all proxies are included, the results of our estimation procedure are statistically equal to those obtained via the standard DID. For example, our method indicates that PROGRESA increased school attendance by 9.2 percentage points, while the standard DID indicates a 9.17 percentage point increase. This difference is not statistically significant.

Comparing columns 3 and 5, we observe that the results are very similar regardless of which moment conditions we use. The GMM estimator based on moment conditions (22) requires the estimation of more parameters than the estimator based on moment conditions (21). This extra set of parameters does not affect the precision of our DID estimator, see columns 4 and 6. Finally, the p-values from the test indicate that we cannot reject the assumption of correct model specification.

Overall, our two proposed estimators are able to recover the standard DID estimand when the proxies used are strongly correlated with the eligibility status.

6 Application: JUNTOS

In this section, we evaluate the impact of the Peruvian social program JUNTOS on the demand for health inputs among women of reproductive age and among children younger than five years old. Contrary to PROGRESA, but similar to most social programs in developing countries, JUNTOS did not contemplate the evaluation of the program in its implementation. Despite the fact that data is available before and after

Table 4: PROGRESA estimation results

Excluded proxy	Propensity score	GMM estimator (21)		GMM estimator (22)		Overidentification test		F-statistic first stage
		coeff	s.e.	coeff	s.e.	Chi2 stat	p-value	
none								
	ols	0.092	0.039	0.101	0.039	5.72	0.221	417.19
	probit	0.091	0.039	0.100	0.039	5.80	0.215	
	logit	0.090	0.038	0.100	0.039	5.73	0.220	
water								
	ols	0.108	0.041	0.115	0.040	4.01	0.260	468.40
	probit	0.112	0.041	0.116	0.041	4.07	0.254	
	logit	0.110	0.040	0.114	0.040	4.03	0.258	
restroom								
	ols	0.092	0.040	0.098	0.039	5.77	0.123	510.60
	probit	0.091	0.040	0.098	0.040	5.89	0.117	
	logit	0.090	0.039	0.098	0.039	5.79	0.122	
electricity								
	ols	0.097	0.045	0.103	0.044	6.06	0.109	380.33
	probit	0.104	0.046	0.104	0.046	6.16	0.104	
	logit	0.103	0.043	0.102	0.044	6.05	0.109	
nrooms								
	ols	0.070	0.049	0.072	0.049	4.79	0.188	325.92
	probit	0.060	0.048	0.072	0.049	4.79	0.188	
	logit	0.058	0.048	0.071	0.049	4.79	0.188	
land								
	ols	0.088	0.039	0.096	0.039	2.43	0.4887	518.95
	probit	0.087	0.039	0.096	0.039	2.55	0.4664	
	logit	0.086	0.038	0.096	0.039	2.44	0.4864	

the program, the identity of beneficiaries and non-beneficiaries of JUNTOS can only be observed after the program was implemented.

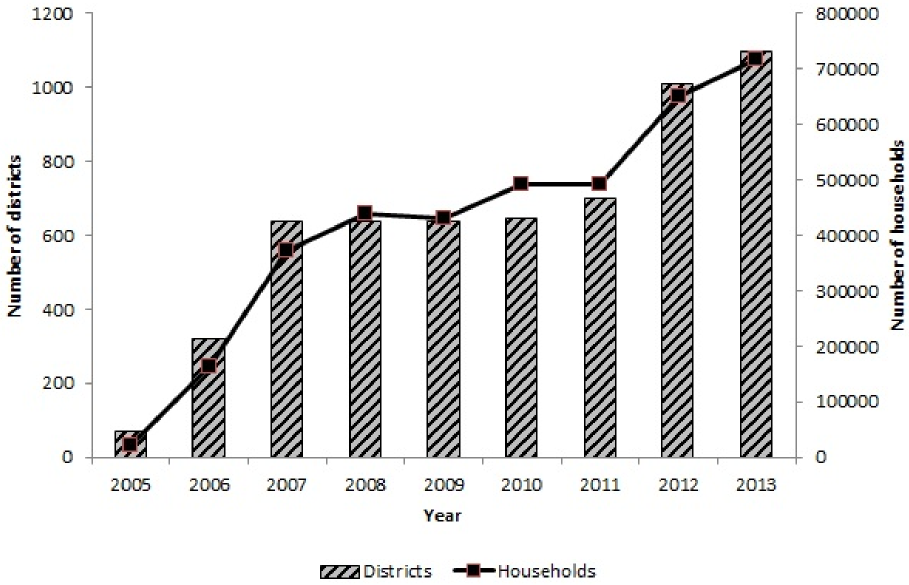
6.1 Description of JUNTOS

JUNTOS is a conditional cash transfer program implemented in Perú in 2005. The goal of the program is poverty reduction in both the short- and the long-run via human capital accumulation.

JUNTOS consists of a monthly monetary transfer of 100 soles (35 USD) per household. The money is transferred to the mother with the condition that (i) children younger than 5 years old attend health controls, (ii) children who are 6 to 14 years old attend formal education at least 85% of the academic year, and (iii) pregnant women and lactating mothers receive pre- and postpartum health services, see Perova and Vakis (2012).[23]

Figure 4 shows the evolution of JUNTOS from 2005 to 2013. The program was first implemented in 70 Peruvian districts in 2005. During 2006 and 2007 the program expanded to 638 districts reaching more than 420,000 households. From 2008 to 2011 the districts and households enrolled in the program remained relatively constant. In 2012, JUNTOS incorporated new regions and changed the conditionality.

Figure 4: Evolution of JUNTOS



[23] Details of JUNTOS can be found at www.juntos.gob.pe.

6.2 The Dataset

The dataset we use is the *Encuesta Nacional de Hogares* (ENAHO). It is a nationally representative survey regularly conducted by the National Institute of Statistics of Peru (INEI). Since 2004, approximately 90,000 individuals have been interviewed annually. All rounds of ENAHO are comparable, i.e. the interviewing methodology and the framing of questions have not changed since 2004. Before 2004, the ENAHO had a different methodology with a much reduced set of questions. To avoid comparability problems due to methodological changes, we do not use the previous version of ENAHO.

Our sample contains pre-treatment years 2004 and 2005 and post-treatment years 2008 to 2011. We restrict the sample to districts where JUNTOS was introduced between 2006 and 2007, excluding the 70 districts where the program was operating by the end of 2005. The ENAHO survey conducted in 2008 was the first round asking household members if they received JUNTOS payments. That is, 2008 is the first year when treatment status is observed. The last post-treatment year we use is 2011, which is the last year before the program conditionality changed.

Table 5 shows summary statistics for women of reproductive age and for children. The variable JUNTOS is an indicator variable for whether the person lives in a household that is a beneficiary of the program. This variable is observed only in 2008 to 2011. Half of the women and approximately two-thirds of the children in the districts where the program was implemented were living in a household where at least one person was enrolled in JUNTOS.

For women, the outcomes we analyze are (i) whether they sought medical care when sick, (ii) whether they received advice in relation to family planning, and (iii) whether they used contraceptives. For children, we study (i) whether they sought medical care when sick, and (ii) whether they received vaccines.

Table 5: Descriptive statistics for women and children

Women 15 to 45 years old			
	Obs	Mean	S.d.
JUNTOS	13911	0.53	0.50
did not seek medical care when sick	11805	0.63	0.48
family planning advice (last 3 months)	18836	0.16	0.37
contraceptives (last 3 months)	19888	0.14	0.35
Children 0 to 5 years old			
	Obs.	Mean	S.d.
JUNTOS	8829	0.64	0.48
did not seek medical care when sick	7341	0.44	0.50
vaccines	12586	0.38	0.48

6.3 DID when participation is not observed in pre-treatment periods

Just as in the PROGRESA case, the households eligible for JUNTOS are those considered poor by the government. Poverty status was determined according to an index weighting housing and household member characteristics correlated with poverty. The complete list of variables used to determine eligibility is not publicly known, neither is the method for calculating the poverty index. We use observed housing characteristics and the level of education of adults in the household as proxies for treatment status.

Table 6 shows the difference in proxy variables for beneficiaries and non-beneficiaries of JUNTOS in post-treatment years. Housing characteristics and education of adults are highly correlated with treatment status.

Estimation with multiple periods When data for multiple periods are available, it is common practice to analyze pre-treatment trends to explore the validity of the parallel trends assumption. Our method allows for this possibility by modifying the moment conditions (22) in the following straightforward way. To estimate the impact of JUNTOS on the variables of interest we use the following moment conditions:

$$E\begin{bmatrix} Z'_{post}(D - Z_{post}\gamma) \\ (Z_{si}, D_{si})'(Y_{si} - d'_{si}\beta_s - Z_{si}\delta) \\ Z_{fi}(Y_{fi} - \pi'_{fi}(Z_{fi},\gamma)\beta_f - Z_{fi}\delta)) \end{bmatrix} = 0 \tag{29}$$

$$f \in \{2004, 2005\},\ \ s \in \{2008, 2009, 2010, 2011\} \tag{30}$$

where *post* refers to the fact that we pool the data from all post-treatment periods, 2008 to 2011, to estimate the first stage via a linear probabilistic model. Pooling the data is valid under the assumption of stationary propensity score. In case this assumption is violated, our overidentification test will reject the null hypothesis. The second set of moment conditions:

$$E(Z_{si}, D_{si})'(Y_{si} - d'_{si}\beta_s - Z_{si}\delta) = 0$$

is used to estimate the second stage with data from the post-treatment periods. The third set of conditions in (29) is used to estimate the second stage with data for the pre-treatment periods when the treatment status is not observed. This set of conditions is identical to that in (22) allowing for β_{2004} to be different from β_{2005}. Conditions (29) simplify to conditions (22) if $\beta_s = \beta_{2008}$ for $s = 2009, 2010, 2011$ and $\beta_{2004} = \beta_{2005}$.

Each pair of betas can be used to estimate a DID. For instance

$$\widehat{\beta}_{2008} - \widehat{\beta}_{2005} = (\widehat{m}_{2008}, \widehat{m}_{2005})'$$

is used to compute the DID for 2005 and 2008 since

$$\widehat{\theta}_{2008,2005} = \widehat{m}_{2008} - \widehat{m}_{2005}$$

Pre-treatment trends are analyzed using $\widehat{\beta}_{2004}$ and $\widehat{\beta}_{2005}$ in a similar way.

Table 6: Descripive statistics for the JUNTOS proxy variables

Women 15 to 45 years old					
		D=1	D=0	difference	(s.e.)
observations		7743	6168		
roof material					
	reed, palm tree leaves, mud	0.183	0.087	0.096	(0.006)
	concrete	0.461	0.510	-0.049	(0.008)
floor material					
	concrete	0.034	0.203	-0.168	(0.0052)
	soil	0.908	0.661	0.247	(0.0066)
wall material					
	bricks, concrete	0.005	0.082	-0.076	(0.0033)
	mud	0.890	0.775	0.115	(0.0062)
	wood	0.036	0.059	-0.023	(0.0036)
max years of education among adults		8.587	10.589	-2.002	(0.0561)
Children 0 to 5 years old					
		D=1	D=0	difference	s.e.
observations		5779	3050		
roof material					
	reed, palm tree leaves, mud	0.202	0.124	0.078	(0.0084)
	concrete	0.453	0.499	-0.046	(0.011)
floor material					
	concrete	0.030	0.162	-0.131	(0.0058)
	soil	0.911	0.689	0.222	(0.008)
wall material					
	bricks, concrete	0.005	0.065	-0.061	(0.0035)
	mud	0.864	0.737	0.127	(0.0084)
	wood	0.042	0.086	-0.044	(0.0052)
max years of education among adults		7.957	9.522	-1.564	(0.0724)

Results Figure 5 shows the results for women of reproductive age. The graphs show the impact of JUNTOS and the 90% confidence interval for each year in relation to year 2005. In 2005, the DID is zero by construction. The pre-treatment trend is analyzed for the years 2004 and 2005. In each graph, we include the results of our over-identification test.

JUNTOS reduced the probability that women did not seek medical care when sick by 18 percentage points. The impact is similar for all post-treatment years. JUNTOS increased both the probability of receiving advice in relation to family planning and the use of contraceptives.

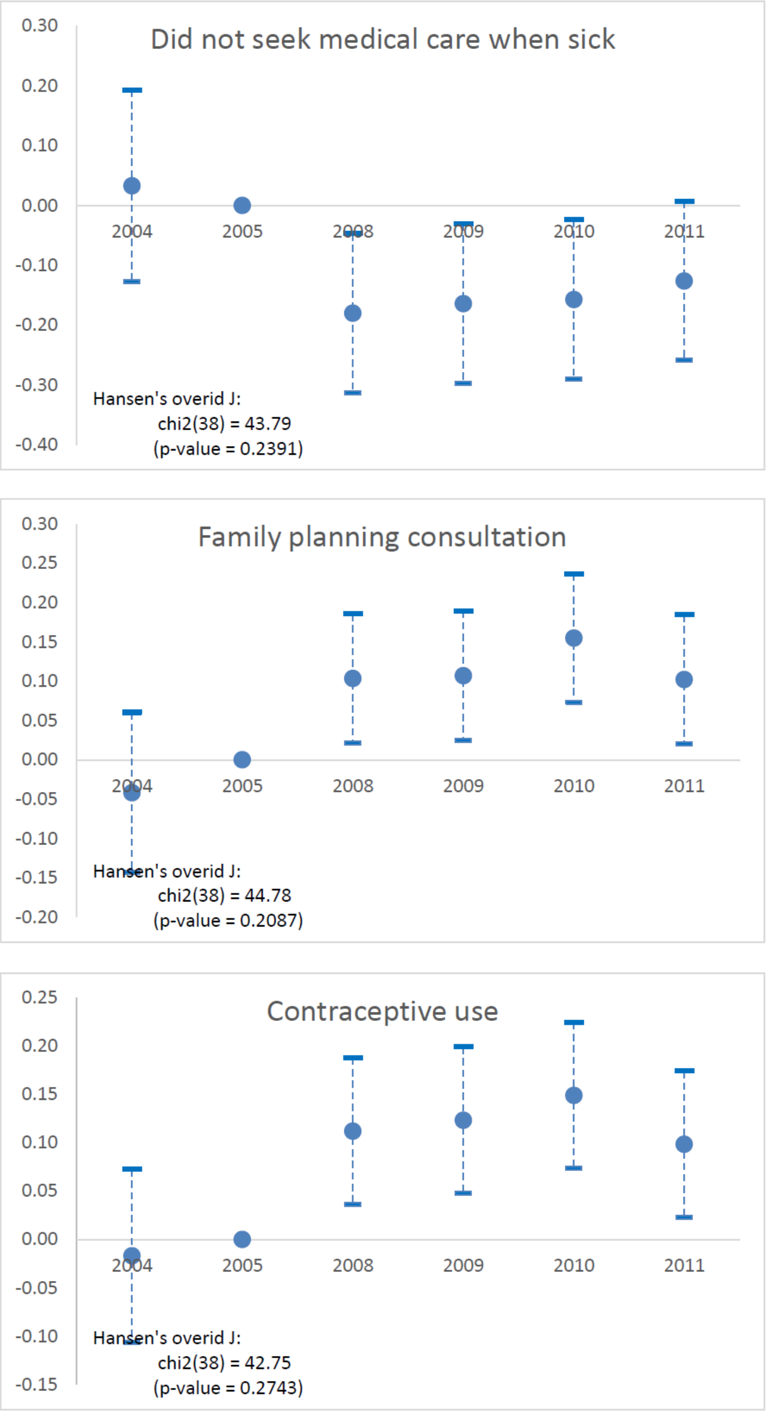
The program had a positive impact on children as well, see Figure 6. The probability that the mother did not seek medical care for the child when the child was sick declined by more than 40 percentage points. The impact of JUNTOS on vaccination was positive and particularly strong for 2008. For the rest of the years it was around 8 percentage points although the estimate is not statistically different from zero at conventional levels.

Consistent with the parallel trends assumption, the DID using the 2004 and 2005 samples is not statistically different from zero. Additionally, our over-identification test fails to reject the assumption of correct model specification.

7 Conclusion

In this paper, we consider difference-in-differences with repeated cross sections when the treatment status is observed in only one period. We show that the average treatment effect on the treated is identified when a proxy for the latent treatment status exists and is observed in both time periods. Our main identifying assumptions are: (i) the propensity score is stationarity conditional on the proxy, and (ii) the proxy is not correlated with the change over time in potential outcomes conditional on the true treatment status. We propose two GMM estimators, whose small sample performance is evaluated via Monte Carlo simulations. We also provide Monte Carlo results for how our estimator compares to the standard DID estimator and to other estimators that have been used in applied work. Additionally, we propose an overidentification test for model specification. Finally, we provide an empirical illustration to data from the Mexican conditional cash-transfer program PROGRESA, and an application to data from the Peruvian cash-transfer program JUNTOS.

Figure 5: Women



Did not seek medical care when sick

Hansen's overid J:
chi2(38) = 43.79
(p-value = 0.2391)

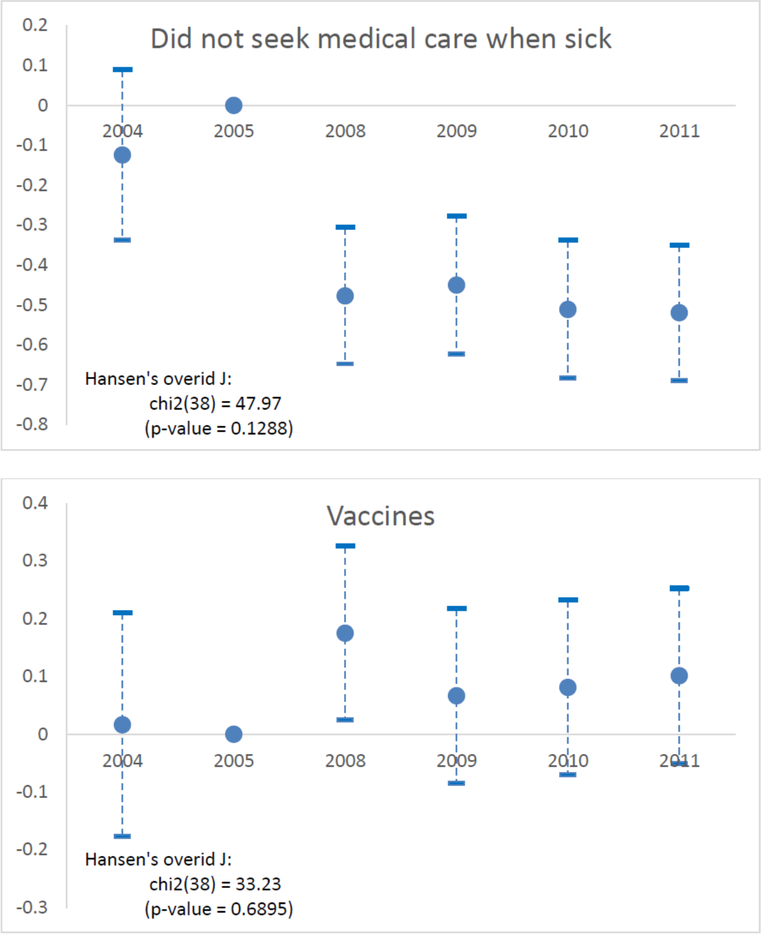
Family planning consultation

Hansen's overid J:
chi2(38) = 44.78
(p-value = 0.2087)

Contraceptive use

Hansen's overid J:
chi2(38) = 42.75
(p-value = 0.2743)

Figure 6: Children



References

Abadie, A. (2005) Semiparametric Difference-in-Difference Estimators. *Review of Economic Studies,* 72: 1-19.

Abrevaya, J. and Donald, S.G. (2011) A GMM Approach for Dealing with Missing Data on Regressors and Instruments. Working paper. Department of Economics, University of Texas.

Angrist, J.D. (1991) Grouped-Data Estimation and Testing in Simple Labor-Supply Models. *Journal of Econometrics,* 47: 243-266.

Angrist, J.D., and Krueger, A.B. (1992) The Effect of Age at School Entry on Educational Attainment: An Application of Instrumental Variables with Moments from Two Samples. *Journal of the American Statistical Association*, 87: 328–336

Attanasio, O., Meghir, C., and Santiago, A. (2012) Education Choices in Mexico: Using a Structural

Model and a Randomization Experiment to Evaluate PROGRESA. *Review of Economic Studies,* 79: 1495-1526.

Blundell, R. and Costa Dias, M. (2009) Alternative Approaches to Evaluation in Empirical Microeconomics. *J. Human Resources*, 44(3): 565-640.

Carroll, R. J., Ruppert, D., Stefanski, L. A., and Crainiceanu, C. M. (2006) Measurement Error in Nonlinear Models: A Modern Perspective. 2nd ed. London: Chapman & Hall.

Chen, X., Hong, H., and Tarozzi, A. (2008) Semiparametric Efficiency in GMM Models with Auxiliary Data. *Annals of Statistics,* 36(2): 808-843.

Chen, X., Hu, Y., and Lewbel, A. (2009) Nonparametric Identification and Estimation of Nonclassical Errors-in-Variables Models Without Additional Information. *Statistica Sinica,* 19: 949-968.

Cross, P.J., and Manski, C.F. (2002) Regressions, Short and Long. *Econometrica,* 70(1): 357-368.

Devereux, P, and Tripathi, G. (2009) Optimally Combining Censored and Uncensored Datasets. *Journal of Econometrics*, 151: 17-32.

Galasso, E. and Ravallion, M. (2004) Social Protection in a Crisis: Argentina's Plan Jefes y Jefas. *World Bank Economic Review,* 18(3): 367-399.

Graham, B.S. (2011) Efficiency Bounds for Missing Data Models with Semiparametric Restrictions. *Econometrica*, 79(2): 437-452.

Groen, J.A. and Polivka, E. (2008) The Effect of Hurricane Katrina on the Labor Market Outcomes of Evacuees. *The American Economic Review*, 98(2): 43-48.

Henry, M., Kitamura, Y., and Salanié, B. (2014) Partial Identification of Finite Mixtures in Econometric Models. *Quantitative Economics*, 5(1): 123-144.

Hu, Y. (2008) Identification and Estimation of Nonlinear Models with Misclassification Error Using Instrumental Variables: A General Solution. *Journal of Econometrics,* 144: 27-61.

Imbens, G.W., and Angrist, J.D. (1994) Identification and Estimation of Local Average Treatment Effects. *Econometrica,* 62(2): 467-475.

Inoue, A., and Solon, G. (2010) Two-Sample Instrumental Variables Estimators. *The Review of Economics and Statistics,* 92: 557-561.

Lee, S. and Shaikh, A. (2014) Multiple Testing and Heterogeneous Treatment Effects: Re-evaluating the Effect of Progresa on School Enrolment. *Journal of Applied Econometrics,* 29: 612–626.

Lewbel, A. (2007) Estimation of Average Treatment Effects with Misclassification. *Econometrica,* 75: 537-551.

Mahajan, A. (2006) Identification and Estimation of Regression Models with Misclassification. *Econometrica,* 74: 631-665.

Molinari, F. (2010) Missing Treatments. Journal of Business and Economic Statistics, 28(2): 82-95.

Muris, C. (2013) Efficient GMM Estimation with General Missing Data Patterns. Working paper, Department of Economics, Simon Fraser University.

Newey, W.K. and McFadden, D. (1994) Large Sample Estimation and Hypothesis Testing. In: Engle, R., McFadden, D. (Eds.), Handbook of Econometrics, vol. IV. Elsevier Science B.V., 2111-2245.

Perova, E. and Vakis, R. (2012) 5 Years in Juntos: New Evidence on the Program's Short and Long-Term Impacts. *Revista Economía,* Departamento de Economía - Pontificia Universidad Católica del Perú, 35(69): 53-82.

Ridder, G. and Moffitt, R. (2007) The Econometrics of Data Combination. *Handbook of Econometrics*, Vol. 6B, Chapter 75. New York: North-Holland.

Schultz, T.P. (2004) School Subsidies for the Poor: Evaluating the Mexican Progresa Poverty Program. *Journal of Development Economics,* 74: 199-250.

SEDESOL (2006) Nota Metodológica General Rural. *Instituto Nacional de Salud Pública. Coordinacion Nacional de Programa de Desarrollo Humano Oportunidades*

Skoufias, E. (2001) PROGRESA and Its Impacts on the Human Capital and Welfare of Households in Rural Mexico: A Synthesis of the Results of an Evaluation. International Food Policy Research Institute, Washington, DC.

Skoufias, E., Parker, S., Behrman, J., and Pessino, C. (2001) Conditional Cash Transfers and Their Impact on Child Work and Schooling: Evidence from the PROGRESA Program in Mexico. *Economia,* 2: 45-96.

Todd, P., and Wolpin, K. (2006) Assessing the Impact of a School Subsidy Program in Mexico: Using a Social Experiment to Validate a Dynamic Behavioral Model of Child Schooling and Fertility. *American Economic Review,* 96: 1384-1417.

A DID with Missing Status and Covariates

In the presence of observed covariates, the following assumptions obtain the identification of the ATT.

Let X denote covariates that are observed in both time periods, and let $\mathcal{X}$ denote the support of X.

Assumption C1 (parallel paths)

$$E(Y_1(0)|X, D=1) - E\left(Y_0\left(0\right)|X, D=1\right) = E\left(Y_1(0)|X, D=0\right) - E\left(Y_0(0)|X, D=0\right)$$

Assumption C2 (no anticipation)

$$E(Y_0(0)|X, D=1) = E(Y_0(1)|X, D=1)$$

Assumption C3 (common support) $F_t(D=1|X) < 1$, for $t = 0, 1$ and for all X.

Let Z be a time-invariant random variable with support $\mathcal{Z}$ and define the propensity score at time t as

$$e_t\left(Z, X\right) \equiv F_t\left(D=1|Z, X\right), \ t = 0, 1$$

Assumption C4 (missing group membership) Distribution functions $F_0\left(Y_0, Z, X\right)$ and $F_1\left(Y_1, D, Z, X\right)$ are observed, while $F_0\left(Y_0, D, Z, X\right)$ is not observed.

Assumption C5 (stationarity) The propensity score is stationary, i.e. for all $z \in \mathcal{Z}$ and $x \in \mathcal{X}$

$$e_0\left(z, x\right) = e_1\left(z, x\right) \equiv e\left(z, x\right)$$

Assumption C6 (relevance) For all $x \in \mathcal{X}$, Z is informative about D, i.e. for $z_1 \neq z_2 \in \mathcal{Z}$

$$e\left(z_1, x\right) \neq e\left(z_2, x\right)$$

Assumption C7 (conditional mean independence in changes) For all D, Z, and X, the change over time in mean potential outcomes is independent of Z conditional on D and X:

$$E(Y_1\left(D\right)|D, Z, X) - E\left(Y_0\left(D\right)|D, Z, X\right) = E(Y_1\left(D\right)|D, X) - E\left(Y_0\left(D\right)|D, X\right)$$

Theorem 4. *Suppose that Z takes on $K \geq 2$ different values, $\{z_k\}_{k=1}^{K} \in \mathcal{Z}$. Additionally, let P_x be a $K \times 2$ matrix and Δ_x be a $K \times 1$ vector defined as, respectively:*

$$P_x \equiv \begin{bmatrix} 1 - e\left(z_1, x\right) & e\left(z_1, x\right) \\ ... & ... \\ 1 - e\left(z_K, x\right) & e\left(z_K, x\right) \end{bmatrix}, \ \Delta_x \equiv \begin{bmatrix} E\left(Y_1|Z=z_1, X=x\right) - E\left(Y_0|Z=z_1, X=x\right) \\ ... \\ E\left(Y_1|Z=z_K, X=x\right) - E\left(Y_0|Z=z_K, X=x\right) \end{bmatrix} \tag{31}$$

Let assumptions C1 to C7 hold. Then the ATT is identified, and given by the difference between two conditional means given by:

$$\begin{bmatrix} E\left(Y_1|D=0, X=x\right) - E\left(Y_0|D=0, X=x\right) \\ E\left(Y_1|D=1, X=x\right) - E\left(Y_0|D=1, X=x\right) \end{bmatrix} = \left(P_x' P_x\right)^{-1} P_x' \Delta_x \tag{32}$$

Proof. The proof is similar to that of Theorem 2 and it is not repeated here. □

B Simulation Results

Table 7: Simulation Results with n = 2000

	Infeasible DID			Naive DID			GMM 1			GMM 2		
$\kappa = 0.2$	bias	s.e.	RMSE	bias	s.e.	RMSE	bias	s.e.	RMSE	bias	s.e.	RMSE
$\alpha = 0.9$	0.0025	0.1096	0.1096	-3.0192	0.2219	3.0273	0.0056	0.1401	0.1402	0.0038	0.1361	0.1362
$\alpha = 0.7$	0.0061	0.1110	0.1112	-3.2762	0.2275	3.2841	0.0073	0.1526	0.1527	0.0069	0.1493	0.1494
$\alpha = 0.5$	0.0068	0.1109	0.1111	-4.1364	0.2437	4.1435	0.0065	0.1872	0.1873	0.0039	0.1804	0.1804
$\alpha = 0.3$	0.0000	0.1138	0.1138	-5.8922	0.2633	5.8981	0.0084	0.3086	0.3087	0.0021	0.2855	0.2855
$\alpha = 0.1$	-0.0021	0.1145	0.1145	-8.0622	0.2802	8.0670	-0.0190	1.6243	1.6244	-0.0655	1.0971	1.0991
$\kappa = 0.5$	bias	s.e.	RMSE	bias	s.e.	RMSE	bias	s.e.	RMSE	bias	s.e.	RMSE
$\alpha = 0.9$	-0.0025	0.0903	0.0903	-3.0271	0.2543	3.0378	-0.0013	0.1128	0.1128	-0.0020	0.1088	0.1088
$\alpha = 0.7$	-0.0011	0.0921	0.0921	-3.2138	0.2631	3.2943	-0.0004	0.1208	0.1208	-0.0024	0.1169	0.1169
$\alpha = 0.5$	0.0001	0.0894	0.0894	-4.1418	0.2853	4.1516	0.0001	0.1477	0.1477	-0.0009	0.1422	0.1422
$\alpha = 0.3$	0.0006	0.0888	0.0888	-5.8940	0.3142	5.9020	0.0109	0.2524	0.2526	-0.0001	0.2324	0.2324
$\alpha = 0.1$	-0.0013	0.0898	0.0898	-8.0524	0.3304	8.0592	0.0338	1.4391	1.4395	-0.0825	1.2524	1.2551
$\kappa = 0.8$	bias	s.e.	RMSE	bias	s.e.	RMSE	bias	s.e.	RMSE	bias	s.e.	RMSE
$\alpha = 0.9$	0.0003	0.1131	0.1131	-2.9983	0.3916	3.0238	0.0039	0.1482	0.1483	0.0021	0.1385	0.1386
$\alpha = 0.7$	-0.0035	0.1119	0.1119	-3.2788	0.4092	3.3042	0.0031	0.1589	0.1589	-0.0008	0.1471	0.1471
$\alpha = 0.5$	-0.0036	0.1143	0.1143	-4.1421	0.4460	4.1660	-0.0022	0.2028	0.2028	-0.0069	0.1850	0.1851
$\alpha = 0.3$	-0.0039	0.1114	0.1115	-5.8771	0.4853	5.8971	-0.0087	0.3544	0.3545	-0.0257	0.2900	0.2911
$\alpha = 0.1$	0.0005	0.1142	0.1142	-8.0554	0.5053	8.0713	0.1122	2.2277	2.2305	-0.2073	2.3183	2.3275