Journal of Comparative Economics xxx (xxxx) xxx

Contents lists available at ScienceDirect

Journal of Comparative Economics

journal homepage: www.elsevier.com/locate/jce

# The Washington Consensus Works: Causal Effects of Reform, 1970-2015

Kevin B. Grier [a], Robin M. Grier [b,*]

[a] Department of Political Science, Texas Tech University, 113 Holden Hall, Lubbock, TX, 79409, USA
[b] Free Market Institute, Texas Tech University, Box 45059, Lubbock, TX, 79409, USA

ARTICLE INFO

*JEL Codes:*
O43
O10
P48

*Keywords:*
Reform
Washington Consensus
Rule of law
Property rights
Economic development

ABSTRACT

Traditional policy reforms of the type embodied in the Washington Consensus have been out of academic fashion for decades. However, we are not aware of a paper that convincingly rejects the efficacy of these reforms. In this paper, we define generalized reform as a discrete, sustained jump in an index of economic freedom, whose components map well onto the points of the old consensus. We identify 49 cases of generalized reform in our dataset that spans 141 countries from 1970 to 2015. The average treatment effect associated with these reforms is positive, sizeable, and significant over 5- and 10- year windows. The result is robust to different thresholds for defining reform and different estimation methods. We argue that the policy reform baby was prematurely thrown out with the neoliberal bathwater.

## 1. Introduction

One of the most valuable roles for economists is to inform policymakers about policies or collections of policies that they could adopt which would reliably increase economic growth. Twenty years ago, the so-called Washington Consensus represented a fairly comprehensive list of such policies. [1] However, this view has since fallen far out of fashion. As Rodrik (2006, p. 974) noted more than a decade ago, "it is fair to say that nobody really believes in the Washington Consensus anymore. The debate now is not over whether the Washington Consensus (hereafter WC) is dead or alive, but over what will replace it"

Easterly (2019) argues that the disenchantment over the WC is due to two reasons. The first is the poor overall economic performance of Latin America and Sub-Saharan Africa during and after reforms in the 1980s and 1990s. The second is that published papers showing a link between trade reform and growth were later shown to not be robust.[2] Rodrik (2006, p. 974) provides a third reason, "the reform agenda eventually came to be perceived, at least by its critics, as an overtly ideological effort to impose 'neoliberalism' and 'market fundamentalism' on developing nations."

In this paper, we argue that sustained, generalized policy reform along the lines suggested by the Washington Consensus may have been prematurely discredited. In a sample of 141 countries from 1970 to 2015, we use discrete jumps in the Fraser Institute's Economic Freedom of the World index (hereafter, EFW), which is based on policies such as free trade, secure property rights, lower government

* Corresponding author. Phone: (806) 834-8568, Fax: (806) 742-1854
*E-mail addresses:* kevin.grier@ttu.edu (K.B. Grier), robin.grier@ttu.edu (R.M. Grier).

[1] See Williamson (1989, 1993) for in depth discussions of the components of the Washington Consensus.

[2] Rodríguez and Rodrik (2000) was an influential paper showing the non-robustness of earlier studies. However, Irwin (2019) offers an updated survey that includes newer work demonstrating a positive relationship between trade and growth.

https://doi.org/10.1016/j.jce.2020.09.001
Received 20 July 2019; Received in revised form 31 August 2020; Accepted 1 September 2020

spending, fewer regulations, and sound money.[3] We find 49 usable cases of such generalized, sustained reform. We find a positive, significant, and sizeable short- to medium-term average effect of these reforms on living standards. The result is robust to the size of the jump required to be classified as a reform case and to the technique used to estimate the average treatment effect. Thus, we believe that this set of policies is worth seriously considering for countries that wish to accelerate their growth rates for up to a decade.

The two papers most related to ours are Estevadeordal and Taylor (2013) and Easterly (2019), in that they both argue that the WC, or at least some of its components, has been prematurely discarded. Estevadeordal and Taylor find that "liberalizing tariffs on imported capital and intermediate goods let to faster growth." They find an effect of one percentage point faster annual growth. Easterly shows that while reforms might have been associated with short-run stagnation in Sub-Saharan Africa and Latin America, growth since then has recovered.[4] He goes on to demonstrate that extremely bad policies significantly reduce growth.[5] Our paper differs from theirs in that (a) we use discrete jumps in an index of reform that broadly captures the main ideas of the WC, (b) we study broad-based general reform and not a specific policy, and (c) we incorporate recent work on issues with generalized difference-in-difference panel estimators in our empirical analysis.

There is also a large literature that regresses a range of economic outcomes on the EFW index. Hall and Lawson (2014) survey 198 empirical papers that use the EFW index (in some form) as an explanatory variable and report that in over two-thirds of the studies, economic freedom is positively correlated with a good outcome such as faster growth, better living standards, more happiness, etc. The authors conclude that the "balance of evidence is overwhelming that economic freedom corresponds with a wide variety of positive outcomes with almost no negative tradeoffs."

The vast majority of this empirical literature runs linear regressions with the EFW index on the right-hand side to estimate the treatment effect. This is not ideal, as EFW is a constructed index of a limited range and is unlikely to have a linear effect on outcome variables. That is to say, it is extremely unlikely that a movement in the index from say 1.5 to 2 would have the same effect on GDP as a move from 5.5 to 6. We look for large, discrete, sustained jumps in the index and estimate the average effect of those jumps. As the correct cutoff for defining "large" is not obvious, we experiment with a range of cutoff criteria. For all of these reasons, we believe our results are a substantial contribution the literature on reform and growth.

Specifically, we identify a set of countries that experienced a discrete, sustained, jump in their level of economic freedom. We present estimates of the causal effect of these jumps on growth using matching on differences methods. Even though matching methods are less dependent than regression analysis on the choice of functional form and do not suffer from potential extrapolation bias, they are not favored among economists as a means of identifying causal estimates due to their inability to match on unobservables. To deal with this issue, we follow An and Winship (2017) and match on the first difference of the outcome variable (real per-capita GDP). As they put it, "Using the differenced outcome helps remove the effects of time-invariant factors while matching helps balance covariates and create a more focused causal inference."[6]

Matching also computes the average treatment effect as the simple, unweighted average of the individual treatment effects. In contrast, the more commonly used two-way fixed effects with a treatment dummy model (often referred to as a difference-in-difference model), does *not* do this if there is treatment effect heterogeneity, as De Chaisemartin and D'Haultfoeuille (2019, hereafter CH) and Goodman-Bacon (2018) show. CH put it this way, "Linear regressions with period and group fixed effects are widely used to estimate treatment effects. We show that they identify weighted sums of the average treatment effects in each group and period, with weights that may be negative. Due to the negative weights, the linear regression estimand may for instance be negative while all the (individual) average treatment effects are positive."

When we estimate a two-way fixed effects panel with the reform treatment dummy, we find a positive and significant effect of generalized, sustained reform on real per-capita GDP. However, the estimated effect is implausibly large and reform cases later in the period and in richer countries receive greater weight in determining the coefficient. For these reasons, we prefer to emphasize our matching on the differences results.

In what follows below, Section 2 reviews the Washington Consensus and discusses how the Economic Freedom of the World index matches well with it. Section 3 outlines how we define reform as a jump in EFW and describes which regions and time periods experienced generalized, sustained reform between 1970 and 2014. In Section 4, we discuss our methods, explaining the benefits of using propensity score or covariate matching on the differences methods instead of regression to deal with selection bias. We also discuss the recent literature that raises questions about how appropriate panel difference-in-difference estimators are for estimating average treatment effects when there is staggered adoption and treatment effect heterogeneity. Section 5 presents the results. Section 6 concludes.

## 2. Economic freedom and the Washington Consensus

Williamson (1989) coined the term WC and he used it to define a broad consensus amongst policymakers and academics about what

[3] Gwartney et al. (2018).

[4] Prati et al. (2013) study the effect of a variety of sector-specific reform indices on growth, finding some cases of very large effects. However, their regressions use these individual indices as linear regressors, a practice we critique below.

[5] Easterly's findings are consistent with de Carvalho Filho and Chamon (2012), who show that Brazil and Mexico actually grew faster post-reform than many originally thought. They note that "economic policies are often judged by a handful of statistics, some of which may be biased during periods of change."

[6] See also Heckman et al. (1997, 1998).

types of reforms would be growth-enhancing. The tenets of the WC included policies such as sustainable fiscal and monetary policy, an elimination of public spending on subsidies, tax reform, market-determined interest rates and exchange rates, a liberalization of trade and foreign direct investment, privatization, deregulation, and secure property rights. This is a long list and most of the tests of the WC have focused on the most controversial of these recommendations, trade liberalization, and have paid less attention to the complexity of the rest of the consensus.

One reason for this might be that it is difficult to measure all of the various components of the WC. We argue that the Economic Freedom of the World (hereafter, EFW) Index by the Fraser Institute is a good proxy for the consensus. The EFW is an index reflecting a spectrum of policy on trade, property rights, government spending, and money, with higher values going to countries with freer trade, more secure property, less government spending, and more sound monetary policy.[7] All of these components map nicely onto the WC. As Birdsall et al. (2010) state, "For its advocates, the Consensus reflected a doctrine of economic freedom that was best suited for the political democracies to which many Latin countries had returned after a long spell of military dictatorships."

We believe there is good reason to think these types of policies, on balance, would raise equilibrium GDP and thus temporarily raise growth. The rule of law and secure property rights encourages the long-term investments that are important for economic growth.[8] Freer trade and lower regulation let resources be allocated more efficiently, and stable government policies help reduce uncertainty and encourage investment. [9] We include the phrase "on balance" because we recognize that government regulation and spending can be growth promoting. The effect of smaller government and less regulation on growth depends on the initial level compared to the efficient level of such activities.[10] In addition, at least in the short-term, more economic freedom might not always be better for GDP. Big changes can cause temporary disruptions in existing patterns of economic activity or cause pushback from groups that were beneficiaries of the old way of doing things. Big Bang reforms, for example, do not have a good reputation for effectiveness (Dewatripont and Roland (1995)).[11]

## 3. Measuring jumps in economic freedom

We use the EFW index as our measure of economic freedom, which is available every five years from 1970 to 2000 and then yearly after that. EFW is a broad index of economic freedom that uses over 70 different variables to document the size of a country's government, its legal system and how secure property rights are, whether the country has sound money, whether its citizens can trade freely with other nations, and the level of economic regulation. [12]

As we noted above, the EFW is attractive for our purposes as it tracks the Washington Consensus fairly well. We also like it because it has a wide country coverage over a longer time period than other similar indices. [13] However, with any index like this it is valid to consider whether, consciously or not, the constructors manipulate the values to reflect desired outcomes.[14] It is important to note that the EFW index is based on third party data, not on the subjective evaluation of the Fraser institute's researchers, and its individual components rarely change, making it less likely to suffer from "halo bias."[15,16]

The average EFW score in our sample is 6.51 out of 10, where 10 represents the highest level of economic freedom, and the sample standard deviation is 1.13. In the spirit of Hausmann et al. (2005), we identify large sustained increases and decreases in EFW in a

[7] These policies are consistent with Williamson's (1989, 1993) articulation of the ten main tenets of what he called the "Washington Consensus."

[8] See Clague et al. (1996)

[9] On trade see Ben-David and Loewy (1998), on stability see Alesina et al. (1996) and Dollar and Kraay (2002).

[10] See Barro (1990) and Karras (1996) for more on the theory and empirical evidence of optimal government size.

[11] More recently though, Eicher and Schreiber (2010) find in a panel of transition countries that structural reform has led to significantly higher growth rates, while Kerekes (2012) shows that successful countries have "limited policy volatility" and invest in education, infrastructure, and trade liberalization

[12] "In order to receive a high EFW rating, a country must provide secure protection of privately-owned property, a legal system that treats all equally, even-handed enforcement of contracts, and a stable monetary environment. It also must keep taxes low, refrain from creating barriers to both domestic and international trade, and rely more fully on markets rather than government spending and regulation to allocate goods and resources." https://www.fraserinstitute.org/economic-freedom/approach

[13] There are 3 other potential indices of economic freedom: the Heritage Foundation's Index of Economic Freedom, a competitiveness ranking from The World Economic Forum (WEF), and the International Institute for Management Development (IMD)'s World Competitiveness Ranking. They are fairly highly correlated. Ochel and Röhn (2006) show that the correlation between EFW and the WEF, IMD, and Heritage indices are .68, .83, and .87). We prefer the EFW index because it matches closely to the Washington Consensus, relies exclusively on third party data, and is available for both a long time period and a wide range of countries. In contrast, the IMD index does not start until 1995 and only has 63 countries, less than half the country coverage of EFW or Heritage. The WEF Global Competitiveness Index does not start until 2007, which means there would be even fewer cases to study, and the Heritage Index begins in 1995. For robustness purposes, we run our model on all the Heritage Index jumps (sustained or not). We include this in Appendix B.

[14] See Sandefur and Wadhwa (2018) for a relevant potential example of this phenomenon.

[15] From the Fraser institute website, "Within the five major areas, there are 24 components in the index. Many of those components are themselves made up of several sub-components. In total, the index comprises 42 distinct variables. All variables come from third party sources, such as the International Country Risk Guide, the Global Competitiveness Report, and the World Bank's Doing Business project, so that the subjective judgments of the authors do not influence the index. This also creates transparency and allows researchers to replicate the index. The index for past years is updated with each new edition to take account of revisions in the underlying data." https://www.fraserinstitute.org/economic-freedom/approach

[16] In our work, we are comparing outcomes in the treated units to the outcomes of their matched controls, where the treatment is a large discrete jump in the index. This method further mitigates, but not eliminates, any concerns that manipulation is driving our results.

sample of 141 countries from 1970 to 2014. We define large jumps to be equal to or greater than a 1.0 increase in the EFW index over a 5-year period.[17] These changes in EFW must also be sustained for at least the following ten years to count as a jump in our sample. From 1970 to 2000, we have data for every 5 years. So, if a country's EFW score goes up by one point or more between two 5-year periods and is sustained for the subsequent 5-year period, we code it as a jump. For example, Nicaragua's EFW score in 1990 was 2.92. By 1995, it was 6.05. This increase is larger than our 1.0 threshold but before we code it as a jump, we need to check if the increase is sustained over the subsequent ten years. From 1995 to 2005, Nicaragua's EFW score went up by another 1.14 points, meaning that these gains in economic freedom were sustained over the following ten years and thus we code Nicaragua in 1995 as experiencing a sustained jump in economic freedom.[18]

Table 1 lists the countries that had large positive jumps of 1.0 or greater in the EFW index and the time period they took place.[19] In Table 2, we examine the distribution of these jumps over time and across regions. The majority of these jumps happened in the 1990s (64%) and most of them occurred in Sub-Saharan Africa, Latin America and the Caribbean, and Eastern Europe. There were very few positive jumps in the 1970s and 2000s (just 5 episodes for both combined, or less than 10% of the total), while jumps were more common in the 1980s (14%). Jumps were much less frequent in Asia, Oceania, and Western Europe (the first two regions only experienced 4 of the 53 positive jumps in the sample, while the latter had 6).

Before we begin our empirical analysis, it is worth looking at the raw data on the average before and after performance of these 49 cases of reform. Figure 1 shows the evolution of average real per capita income for these cases 10, 5 and 0 years before reform and 5 and 10 years after.[20]

There is an obvious positive trend break in the post reform data, showing that economic performance on average improves in these countries after reform. In what follows, we will quantify and test for the significance of this effect after accounting for measures of counterfactual performance. In other words, we will investigate whether the reforms caused this improved performance.

## 4. Our Empirical Method

Neoclassical growth theory tells us that institutions affect the level of real GDP per capita (Y), so our model is:

$$\mathrm{Ln}(\mathrm{Y}_{\mathrm{it}}) = \alpha_{\mathrm{i}} + \tau_{\mathrm{t}} + X_{\mathrm{it}}{}'\beta + \gamma * \mathrm{REFORM}_{\mathrm{it}} + \varepsilon_{\mathrm{it}} \quad (1)$$

Here the log of real per capita GDP is explained by a country fixed effect ($\alpha_i$), a period fixed effect ($\tau_t$), a set of control variables ($X_{it}$) and a dummy variable ($REFORM_{it}$) that indicates whether or not the treatment is in place in unit i and time t. Different units adopt the reform at different times (and some abandon it), which is referred to as staggered adoption.

This model (two-way fixed effects (TWFE) with a treatment status dummy), is often referred to as a difference-in-difference model, but recent research has pointed out that this is not really correct.[21] In fact, models like this estimate a weighted average of all the different canonical difference-in-difference models that make up the full set of staggered implementations. If there is any treatment effect heterogeneity over time (e.g. groups treated later have different average effects than those treated earlier), the coefficient on the treatment status dummy in the TWFE model is a biased estimate of the actual average effect of the treatment. In fact, the coefficient can actually have the opposite sign of the true average treatment effect!

Techniques like matching or synthetic control can give unbiased estimates of the average treatment effect under staggered adoptions and treatment effect heterogeneity if the assumptions underlying the model are fulfilled. Here we focus on matching.[22] If the underlying assumptions are met, matching does not have a problem with biased treatment effect estimates in the face of staggered adoption and treatment effect heterogeneity because it does not use a weighted average of the individual treatment effects but a rather a simple average to estimate the average effect. However, in matching, the chief concern is that there are unobservables affecting both selection into treatment and the outcome variable that cannot be controlled in the matching process. The TWFE model has been preferred because it sweeps out unobservables, but we now know that under staggered adoption and treatment effect heterogeneity, it does not produce unbiased estimates of the average treatment effect. If we could control for unobservables, matching would give us such an unbiased estimate.

Returning to equation 1, above, it is clear that differencing will remove the unobservables and thus estimating the average

[17] This is close to the sample standard deviation (1.13). Later in the paper, we experiment with raising and lowering the definition of a jump to 1.25 and .75 increases in the index.

[18] We define a sustained jump to be one where the EFW score does not drop by more than .20 points in the subsequent ten-year period. If a country's score falls by more than .20 points in the 5-year period but then recovers (so that the ten-year difference is less than .20 points), we do not consider this a sustained jump. If a country has more than one jump, we take only the first jump if there is less than 10 years between jumps.

[19] The reader may be struck by the absence of China and India from this list. Between 1980-85, China jumps +1.18, but from 1985-90, it falls by -.73. Thus, the reforms were not sustained. Then in 1990-95, China jumps +.98, just missing our 1.0 threshold. This case is included in our "smaller jumps" results shown in Table 8. India has no jumps at or above 1.0 but they have two distinct smaller jumps, +.82 in 1975-80 and +.87 in 1990-95. Both are in the "smaller jumps" results presented in Table 8.

[20] To create this graph, we line up the reforming countries in event time and then average their individual performances before and after their reforms.

[21] See de Chaisemartin and D'Haultfoeuille (2020) and Goodman-Bacon (2019) for derivations and details.

[22] While matching is more typically used in a micro context, there is a burgeoning literature using matching methods in empirical macro. See for example, Hutchison and Noy (2003), Glick et al. (2006), Lin and Ye (2007), and Persson and Tabellini (2009).

**Table 1**
Onsets of Sustained, Generalized Reform.

| Country | Years | EFW Jump | Country | Years | EFW Jump |
|---|---|---|---|---|---|
| Nicaragua | 1990-1995 | 3.13 | Latvia | 1995-2000 | 1.42 |
| El Salvador | 1990-1995 | 2.73 | Egypt | 1990-1995 | 1.41 |
| Uganda | 1990-1995 | 2.37 | Croatia | 1995-2000 | 1.40 |
| Bolivia | 1985-1990 | 2.11 | Peru | 1985-1990 | 1.40 |
| Kuwait | 1980-1985 | 1.92 | Senegal | 1995-2000 | 1.38 |
| Portugal | 1975-1980 | 1.82 | Madagascar | 1995-2000 | 1.35 |
| Ghana | 1985-1990 | 1.79 | Albania | 1995-2000 | 1.31 |
| Chile | 1975-1980 | 1.77 | Mali | 1995-2000 | 1.28 |
| Poland | 1990-1995 | 1.76 | Hungary | 1990-1995 | 1.28 |
| Iran | 1995-2000 | 1.70 | Togo | 1980-1985 | 1.28 |
| T. & Tobago | 1990-1995 | 1.67 | Lithuania | 1995-2000 | 1.25 |
| Zambia | 1990-1995 | 1.66 | Philippines | 1990-1995 | 1.25 |
| Costa Rica | 1985-1990 | 1.65 | Sri Lanka | 1990-1995 | 1.21 |
| Rwanda | 1995-2000 | 1.65 | Russia | 1990-1995 | 1.20 |
| Romania | 1995-2000 | 1.64 | Portugal | 1990-1995 | 1.20 |
| Dominican Rep. | 1990-1995 | 1.62 | Ireland | 1990-1995 | 1.19 |
| Slovenia* | 1995-2000 | 1.61 | Malta* | 1990-1995 | 1.18 |
| Mexico | 1985-1990 | 1.59 | Niger | 1995-2000 | 1.13 |
| Estonia | 1995-2000 | 1.57 | Namibia* | 1990-1995 | 1.11 |
| New Zealand | 1985-1990 | 1.57 | Guinea-Bissau* | 1995-2000 | 1.11 |
| Ukraine | 1995-2000 | 1.53 | Jamaica | 1980-1985 | 1.11 |
| Iceland* | 1985-1990 | 1.50 | Cyprus | 2000-2005 | 1.06 |
| Nigeria | 1995-2000 | 1.49 | Turkey | 1980-1985 | 1.06 |
| Mauritius | 1980-1985 | 1.48 | Indonesia | 1980-1985 | 1.04 |
| Tanzania | 1990-1995 | 1.45 | Brazil | 1985-1990 | 1.01 |
| Bulgaria | 2001-2005 | 1.44 | France | 1985-1990 | 1.01 |
| Israel | 1990-1995 | 1.43 | Jordan | 1995-2000 | 1.01 |

* signifies cases that lack covariates and are thus not included in the statistical analysis

**Table 2**
Frequency of Reform

| Region | 1970s | 1980s | 1990s | 2000s | % of Total | Episodes |
|---|---|---|---|---|---|---|
| Asia | 0 | 1 | 2 | 0 | *5.7* | *3* |
| Sub-Saharan Africa | 0 | 3 | 11 | 0 | *26.4* | *14* |
| Latin America | 1 | 6 | 4 | 0 | *20.8* | *11* |
| Western Europe | 1 | 2 | 2 | 1 | *11.3* | *6* |
| Eastern Europe | 0 | 0 | 11 | 2 | *22.6* | *13* |
| Middle East/N. Africa | 0 | 1 | 5 | 0 | *11.3* | *6* |
| Oceania | 0 | 1 | 0 | 0 | *1.9* | *1* |
| *% of Total* | *3.8* | *26.4* | *64.2* | *5.7* | *100* | |
| *Episodes* | *2* | *14* | *35* | *3* | | *54* |

treatment effect via matching on the differences should give an unbiased estimate of true average effect. We thus propose using matching the treated units to un-treated at the date of the jump in EFW and studying its effect on the first difference (i.e. *growth*) of the log real per-capita income to generate an unbiased estimate of the average effect of reform. This is the method proposed by An and Winship (2017) who say, "Using the differenced outcome helps remove the effects of time-invariant factors while matching helps balance covariates and create a more focused causal inference"

The reason why matching fell out of favor compared to TWFE with treatment status dummy models is a preoccupation over unobservables, but in fact performing a matching analysis and then using the differenced outcome variable to measure the treatment effect removes that obstacle and produces an unbiased average effect estimate in a way that the TWFE plus treatment dummy does not, in cases where there is staggered adoption and treatment effect heterogeneity.

Matching has other advantages over regression analysis. First and foremost, it does not extrapolate to estimate a treatment effect. If there are no controls with propensity scores sufficiently close to a treated unit, that unit is dropped from the analysis.[23] In other words, we restrict our attention to the region of common support. Second, matching is less sensitive to functional form choices than is regression. Finally, matching specifically creates and reports the counterfactuals for each unit. It is possible to manipulate regression results to show the counterfactuals and the degree of extrapolation involved, but that is rarely done, and matching makes this

[23] However, the Mahalanobis matching algorithm we use (*"teffects"* in Stata) uses all treated cases and does not impose common support. Our propensity score matching models, even the kernel matching ones, only use treated cases in the range of common support.

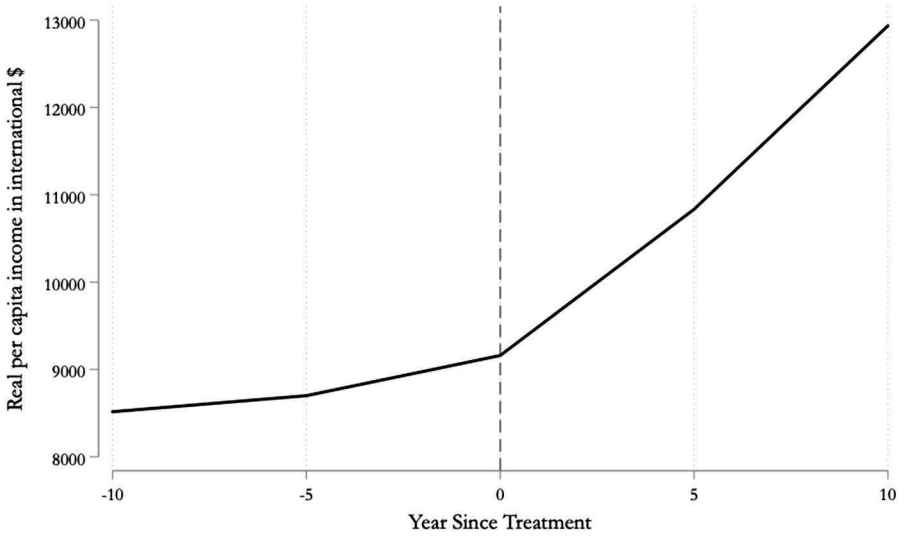


**Fig. 1.** Average Real Per Capita Income in the Treated Units.

explicit.[24]

Matching creates for each treated unit a control unit that is as similar as possible to the treated unit. It then estimates the average treatment effect on the treated by the difference between the average outcome among the treated and the average outcome among the set of matched controls. Exact matching means finding a control that has exactly the same characteristics. With a large number of covariates, or with continuous covariates, exact matching is not feasible. There are two main alternatives. The first is propensity score matching, where a logit or probit equation that models the probability of receiving the treatment is estimated and then the treated units are matched to the control that has the most similar propensity score. The second is, for lack of a better term, inexact covariate matching, where the chosen control is the one with the smallest weighted average of differences from the treated unit.

There are many distinct methods of implementing these matching methods. In this paper, we will report four types of propensity score matching: matching to the nearest neighbor, to the average of the two nearest neighbors, the average of the three nearest neighbors, and kernel matching where all control units are used to construct the match with weights related to the closeness of their propensity score (we use a normal kernel to implement this matching). For our covariate matching we use the Mahalanobis measure of distance to construct the matched controls and present three cases: nearest, nearest 2, and nearest 3 neighbors.

Thus, in our models, we are studying the effect of discrete jumps in EFW and not trying to estimate a linear relationship between economic growth and EFW. Further, we are explicitly modeling the non-random selection of the jumping countries by creating a matched control for each jumper. Our estimate of the treatment effect of the EFW jump is just the difference between the average growth rate over the subsequent 5 or 10 years for the jumping countries minus the average growth rate over the same period for the matched controls.

All attempts to isolate causal inference with observational data rely on at least one untestable assumption. We recognize that economists' emphasis on the importance of unobservables makes matching suspect. But we argue those concerns are mitigated by matching on the differences and that matching has advantages over the more commonly used fixed effects panel regressions.

## 5. Results

### *5.1. Logit to Estimate Propensity Scores*

The first step in matching is to choose a set of covariates to match on. For propensity score models, they are used in a logit or probit model to estimate the propensity scores, while in covariate matching they are used slightly more directly.[25] We address this initial step in this sub-section.

We use the Penn World Table (version 9) for most of our economic data (except for inflation, which comes from World Development Indicators) and also use their human capital index. We use the Polity 4 database for a regime type variable (polity2).[26] Table A1 of Appendix A lists all the countries in our sample, Table A2 describes the variables and their sources, and Table A3 provides summary statistics.

The biggest question for the researcher in matching analysis is what covariates should be used to match on. The literature is mixed

[24] See King and Zeng (2006) and Aronow and Samii (2016)

[25] Feenstra et al. (2015). See Lawson, Murphy, and Powell (2020) for an excellent overview of the literature on the determinants of EFW scores.

[26] Marshall et al. (2017); World Bank (2017).

on what is recommended. Here we match both on variables that determine the probability of jumping and variables that help determine the outcome, which is our reading of the optimal approach. Of course, some variables may well do both. Specifically, we match on seven covariates: the lagged investment rate, the lagged EFW score, the lagged Polity2 score, the lagged human capital index, the lagged share of government consumption in GDP, the lagged export share, and lagged inflation. Investment, human capital, government spending and exports are likely to be important determinants of income as well as determinants of reform, while regime type and inflation are also included as likely determinants of the probability of reform.

In the micro matching literature where data often come from individual level surveys, the number of covariates used in matching may be quite large. However, in the macro matching literature, it is common to match on less than 10 covariates.[27]

Table 3 contains the estimated logit model that we use to generate the propensity scores that underlie our propensity score analysis. Country years with higher investment rates and higher existing levels of economic freedom are significantly less likely to experience a jump in EFW. More democratic and better-educated polities are significantly more likely to see an EFW jump. Lagged, government consumption, exports and inflation are insignificantly related to the probability of an EFW jump.

Figure A1 shows the distribution of the propensity scores for the treated and the untreated. There are a large number of untreated units whose score is lower than the lowest score of any treated units. They will not be used in the analysis. There are only a few treated units with a score higher than the highest score of any untreated unit. They are dropped as well.

### 5.2. Main Results

Table 4 presents our main results. We use both propensity score matching and covariate matching (with the Mahalanobis metric). In both cases we match to the one, two, and three nearest neighbors, and in the propensity score case we also use kernel matching with a normal kernel. This gives us seven estimates per outcome and we study two outcomes: growth over the 5 years and 10 years subsequent to the jump.

For our propensity score matching (PSM) results, we restrict the comparisons to the region of common support, which eliminates six of our treated units in the five-year growth models and eight in the 10-year growth models. Again, we are doing this to avoid using extrapolation to construct the estimated treatment effect. The Mahalanobis covariate matching models use all 49 treated cases. In the PSM case, we construct our standard errors by bootstrapping, using 250 replications in each case. For the covariate matching results, we use bias-adjusted standard errors as developed by Abadie and Imbens (2011).

As shown in the second column of the table, all but one of the coefficients on the 5-year growth rate are positive and significant at the 0.10 level or better. The estimated effect sizes of the statistically significant coefficients range from 2.87 to 2.07 percentage points, with an average estimated treatment effect of 2.51.[28] These results point to a large short-run causal effect of EFW jumps on growth. For the 10-year growth effect, as shown in the fourth column of the table, the propensity score matching coefficients are positive and significant in three of the four cases. The values range from 1.93 to 1.03, implying a strong, long-term effect of an EFW jump. The 10-year effects estimated by covariate matching are smaller, but significant at least at the .10 level in two of the three cases.

These effects are substantial. Assuming an initial per-capita income of $6,000 and a baseline growth rate of 2%, a 5-year treatment effect of 2.51 additional percentage points of growth would leave a country 13% richer than it otherwise would have been (per-capita income after 5 years of $7,481 with the treatment as opposed to $6,624 in the baseline case). Assuming the same baseline, a 10-year treatment effect of 1.54 additional percentage points of growth leaves the treated country 16% richer than it would have been under the baseline (per-capita income after 10 years of $8,496 with the treatment versus $7,314 in the baseline scenario). Our results show that an investment in economic freedom can pay substantial dividends over the short to medium run.[29]

We have also estimated these same models on the sample of all jumps of 1.0 or greater in the index, whether the jump was sustained or not. Using all jumps gives us 75 cases in the 5-year results and 72 in the 10-year. Given that our treatment is a sustained jump, we can consider this exercise as sort of an "intent to treat" outcome. These results are presented in Table 5.

While the coefficients on average are roughly 20% smaller, the significance patterns are the same as what we report in Table 4. This result raises the economic payoffs to reform, in that the average treatment effect of all reforms, no matter how long they last is positive, sizeable, and significant.

### 5.3. Covariate Balance

In PSM, besides restricting the estimates to the region of common support, it is also important to check whether the treated units and their matched controls are sufficiently similar in terms of average values of the covariates. Columns 2 and 4 of Table 4 present a portmanteau test for covariate balance and show that we have no issues with covariate imbalance. Here the null hypothesis is that the matched controls and the treated do not differ in the values of their covariates, and we fail to reject at any reasonable significance level.

[27] Persson and Tabelini (2008) use 6 covariates in their matching equation, Hutchison and Noy (2003) use 14, Glick et al. (2006) use 8 and Lin and Ye (2007) use 7.

[28] These effects are notably larger that the effects Estevadeordal and Taylor (2013) find for tariff reforms.

[29] Our matching algorithm allows the treated unit to be matched to units measured either before, during, or after the treated unit's reform period. We can restrict the matching to be only on untreated units measured during the treated unit's reform period, at least for the propensity score matching models. These results are presented in Table A4. The five-year results are, if anything, stronger than what we report in Table 4, while the 10-year results are a bit weaker.

**Table 3**
Determinants of the Initiation of Reform.

| Variable | Coefficient | p-values | elasticity |
|---|---|---|---|
| Lagged Investment Share of GDP | -4.52* | 0.06 | -0.92 |
| Lagged EFW Score | -1.27*** | 0.01 | -7.35 |
| Lagged Polity 2 Score | 0.073** | 0.02 | 0.22 |
| Lagged Human Capital Index | 1.08*** | 0.01 | 2.25 |
| Lagged Government Share of GDP | 0.102 | 0.95 | 0.018 |
| Lagged Export Share of GDP | -0.52 | 0.66 | -0.117 |
| Lagged Inflation | 0.0001 | 0.67 | 0.002 |
| Intercept | 2.44*** | 0.01 | n.a. |

N=846; Pseudo $R^2 = .255$
p-values are in parentheses. ***, **, and * indicate significance at the .01, .05, and .10 levels, respectively. Estimates are obtained using a Logit model. Elasticities are calculated at the means of the covariates.

**Table 4**
The Effects of Reform on Economic Performance.

| Matching Method | 5-yr growth coefficient | $X^2$ covar. balance | 10-yr growth coefficient | $X^2$ covar. balance |
|---|---|---|---|---|
| Propensity Score: Nearest Neighbor | 2.32 | 1.19 | 1.74 | 3.13 |
| | [.13] | [.99] | [.17] | [.87] |
| Propensity Score: Nearest 2 Neighbors | 2.65** | 2.72 | 1.8* | 1.40 |
| | [.05] | [.91] | [.10] | [.98] |
| Propensity Score: Nearest 3 Neighbors | 2.79** | 2.69 | 1.93** | 0.64 |
| | [.04] | [.91] | [.04] | [.99] |
| Propensity Score: Normal Kernel | 2.87*** | 1.90 | 1.71*** | 1.64 |
| | [.01] | [.96] | [.01] | [.98] |
| Mahalanobis: Nearest Neighbor | 2.07*** | n.a. | 0.39 | n.a. |
| | [.01] | | [.55] | |
| Mahalanobis: Nearest 2 Neighbors | 2.26*** | n.a. | 1.03* | n.a. |
| | [.01] | | [.10] | |
| Mahalanobis: Nearest 3 Neighbors | 2.61*** | n.a. | 1.23** | n.a. |
| | [.01] | | [.05] | |

p-values are in parentheses. ***, **, and * indicate significiance at the .01, .05, and .10 levels, respectively.
Propensity score matching uses 46 of the 49 treated cases in both the 5 and 10 year results.. The other three lie outside the region of common support. The Mahalanobis matching uses all 49 treated cases.
Columns 3 and 5 report the Chi-square statistic testing the null hypothesis that the covariates are, on average, balanced between the treated cases and their matched controls.

In Table 6 we present an example of a detailed breakdown of the testing for covariate balance that helps to show exactly what matching is accomplishing. The table shows covariate balance before and after matching for equation 2 of Table 4.

Before matching, the treated and the untreated differed significantly on 5 of the 7 covariates (investment, EFW, government consumption, exports, and inflation). However, after matching, the treated units inside the range of common support and their matched controls do not significantly differ on any of the covariates, even if we were to choose the 0.25 significance level for the tests.

### *5.4. Panel Regression*

Despite the potential shortcomings of the two-way fixed effects with treatment dummy approach, we implement it here for comparison. Table 7 reports these results. We regress real-per capita GDP on its first lag and the variables we use for matching along with a set of country and time dummies that are not reported. Our treatment dummy, labeled reform, goes to 1 when the EFW jump occurs and stays there for the rest of the time periods, unless the EFW score falls below 80% of the post jump level. The minimum number of periods that reform can be equal to one is 3. Using standard errors clustered at the country level, we find that the coefficient on reform is 0.121 and its p-value is 0.003.

We also implemented the diagnostic checks from CH (2020). None of the individual treatments receive a negative weight (which is good), but there only needs to be a small amount of treatment effect heterogeneity in them for the true average treatment effect to be zero given our estimated coefficient on reform (which is not so good). If the standard deviation (variance) of the individual treatment effects is 0.2126 (0.045) or greater, then the true average treatment effect is indistinguishable from zero, despite our highly significant reform coefficient. When we check what factors are correlated with the weights on the treatment effects we find, they are significantly correlated with time and income. Specifically, treatments occurring later in the sample period and in richer countries are getting greater weight in determining the regression coefficient.

Compared to the unweighted average of the individual treatment effect given in our matching exercises, the TWFE model gives an extremely large effect of reform. The coefficient on reform is 0.121, but there is a lag of the dependent variable in the model with a

**Table 5**
The Effects of Any Reform on Economic Performance.

| Matching Method | 5-yr growth coefficient | $X^2$ covar. balance | 10-yr growth coefficient | $X^2$ covar. balance |
|---|---|---|---|---|
| Propensity Score: Nearest Neighbor | 2.56** | 2.83 | 1.9* | 4.03 |
| | [.04] | [.90] | [.06] | [.78] |
| Propensity Score: Nearest 2 Neighbors | 2.89** | 1.04 | 2.07** | 2.63 |
| | [.02] | [.99] | [.02] | [.92] |
| Propensity Score: Nearest 3 Neighbors | 2.85*** | 1.69 | 1.83** | 3.18 |
| | [.01] | [.99] | [.05] | [.87] |
| Propensity Score: Normal Kernel | 2.45*** | 1.71 | 1.53*** | 1.88 |
| | [.01] | [.97] | [.01] | [.97] |
| Mahalanobis: Nearest Neighbor | 2.09*** | n.a. | 0.83 | n.a. |
| | [.01] | | [.15] | |
| Mahalanobis: Nearest 2 Neighbors | 2.12*** | n.a. | 0.58 | n.a. |
| | [.01] | | [.29] | |
| Mahalanobis: Nearest 3 Neighbors | 2.15*** | n.a. | 0.75 | n.a. |
| | [.01] | | [.15] | |

p-values are in parentheses. ***, **, and * indicate significiance at the .01, .05, and .10 levels, respectively.
Propensity score matching uses 70 (66) of the 75 (72) treated cases the 5 (10) year results. The other 5 (6) lie outside the region of common support. The Mahalanobis matching uses all treated cases.
Columns 3 and 5 report the Chi-square statistic testing the null hypothesis that the covariates are, on average, balanced between the treated cases and their matched controls.

**Table 6**
Detailed Example of Covariate Balance Achieved by Matching.

| Variable | Unmatched/ Matched | Mean Treated | Control | t-test t | p-value |
|---|---|---|---|---|---|
| Lagged Investment | U | 0.162 | 0.213 | -3.60*** | 0.01 |
| | M | 0.164 | 0.149 | 0.94 | 0.35 |
| Lagged EFW | U | 4.58 | 6.04 | -7.61*** | 0.01 |
| | M | 4.68 | 4.62 | 0.29 | 0.77 |
| Lagged Polity 2 | U | 2.94 | 2.87 | 0.07 | 0.95 |
| | M | 2.63 | 2.31 | 0.22 | 0.82 |
| Lagged Human Capital Index | U | 2.10 | 2.15 | -0.43 | 0.66 |
| | M | 2.09 | 1.99 | 0.78 | 0.44 |
| Lagged Govt. Consumption | U | 0.224 | 0.183 | 3.19*** | 0.01 |
| | M | 0.219 | 0.232 | -0.60 | 0.55 |
| Lagged Exports | U | 0.156 | 0.230 | -2.10** | 0.04 |
| | M | 0.161 | 0.180 | -0.60 | 0.55 |
| Lagged Inflation | U | 398.4 | 34.6 | 4.37*** | 0.01 |
| | M | 143.5 | 108.9 | 0.28 | 0.78 |

***, **, and * indicate significiance at the .01, .05, and .10 levels, respectively.
Results are from the PSM nearest 2 neighbors, 5-year growth equation shown in Table 4.

coefficient of 0.686. Thus, the equilibrium impact of reform is 0.121 / (1 - 0.686) or 0.385. Taken literally, this implies that a country would experience a 47% increase in real per-capita GDP from undertaking reform compared to what they otherwise would have experienced. This is a huge number and more than twice as big as any effect we find from our matching experiments.[30]

Because the TWFE results are not robust to modest degrees of treatment effect heterogeneity, because they give more weight to reforms later in the sample and in richer countries, and because the estimated effect taken at face value is implausibly large, we prefer our matching results and will focus on them for the rest of the paper. However, we wanted to show that the TWFE model taken at face value, supports the finding that generalized reform, as measured by a large and sustained increase in the EFW index, significantly raises real incomes.

### 5.5. Robustness Tests

Having established a baseline result, a significant and sizeable causal effect of EFW jumps on economic growth, we now redefine a jump in two ways: first as a movement of .75 or more in a country's EFW score and later as a movement of 1.25 or more. Table 8 presents our results using 0.75 as the threshold defining a jump in EFW. We now have 75 cases with data for the 5-year PSM models. 72 of the 75 cases are in the region of common support and used to construct the estimated treatment effect. As before, the covariate

[30] Our equation is in logs, so to get the effect on levels we take $e^{0.38} = 1.47$, which is where we get the 47% figure.

**Table 7**
Two-Way Fixed Effects Model of the Effect of Reform on Real Per-Capita GDP.

| Variable | Coefficient | Clustered S.E. | P-value |
|---|---|---|---|
| Reform | 0.121 | 0.04 | 0.003 |
| Lagged Real Per-Capita Income | 0.686 | 0.03 | 0.001 |
| Human Capital Index | 0.218 | 0.09 | 0.016 |
| Investment Share of GDP | 0.495 | 0.18 | 0.007 |
| Government Share of GDP | 0.107 | 0.205 | 0.602 |
| Export Share of GDP | -0.027 | 0.08 | 0.74 |
| Polity 2 Score | -0.004 | 0.003 | 0.12 |
| Inflation | -0.0001 | 0.0002 | 0.42 |

N=859; Number of country clusters = 124; Overall $R^2 = 0.97$
Unit and year dummies suppressed.

**Table 8**
The Effects of More Broadly Defined Reform on Economic Performance.

| Matching Method | 5-yr growth coefficient | $X^2$ covar. balance | 10-yr growth coefficient | $X^2$ covar. balance |
|---|---|---|---|---|
| Propensity Score: Nearest Neighbor | 3.48*** | 3.14 | 1.26 | 5.32 |
| | [.01] | [.87] | [.18] | [.62] |
| Propensity Score: Nearest 2 Neighbors | 2.69*** | 2.04 | 1.15 | 2.89 |
| | [.01] | [.96] | [0.19] | [.89] |
| Propensity Score: Nearest 3 Neighbors | 2.60*** | 1.44 | 1.22 | 2.56 |
| | [.01] | [.98] | [.13] | [.93] |
| Propensity Score: Normal Kernel | 2.12*** | 2.76 | 1.62*** | 2.16 |
| | [.01] | [.91] | [.01] | [.95] |
| Mahalanobis: Nearest Neighbor | 1.25** | n.a. | 0.78* | n.a. |
| | [.05] | | [.10] | |
| Mahalanobis: Nearest 2 Neighbors | 1.45** | n.a. | 1.08** | n.a. |
| | [.02] | | [.02] | |
| Mahalanobis: Nearest 3 Neighbors | 1.90*** | n.a. | 1.32*** | n.a. |
| | [.01] | | [.01] | |

p-values are in parentheses. ***, **, and * indicate significiance at the .01, .05, and .10 levels, respectively.
Propensity score matching uses 72 of the 75 treated cases in the 5-year growth results. The other 3 lie outside the region of common support. The Mahalanobis matching uses all 75 treated cases. For the 10-year results, propensity score matching uses 69 of the 74 treated cases, with 5 outside support. The Mahalanobis matching uses all 74 treated cases.
Columns 3 and 5 report the Chi-square statistic testing the null hypothesis that the covariates are, on average, balanced between the treated cases and their matched controls.

**Table 9**
The Effects of More Narrowly Defined Reform on Economic Performance.

| Matching Method | 5-yr growth coefficient | $X^2$ covar. balance | 10-yr growth coefficient | $X^2$ covar. balance |
|---|---|---|---|---|
| Propensity Score: Nearest Neighbor | 1.45 | 5.04 | 1.62 | 1.49 |
| | [.42] | [.66] | [.27] | [.98] |
| Propensity Score: Nearest 2 Neighbors | 1.01 | 0.83 | 2.07* | 2.54 |
| | [.57] | [.99] | [.10] | [.92] |
| Propensity Score: Nearest 3 Neighbors | 1.47 | 0.49 | 2.03* | 2.12 |
| | [.31] | [.99] | [.10] | [.95] |
| Propensity Score: Normal Kernel | 2.57*** | 3.31 | 1.75** | 3.29 |
| | [.01] | [.86] | [.03] | [.86] |
| Mahalanobis: Nearest Neighbor | 1.57* | n.a. | 0.26 | n.a. |
| | [.10] | | [.74] | |
| Mahalanobis: Nearest 2 Neighbors | 1.81** | n.a. | 0.84 | n.a. |
| | [.03] | | [.24] | |
| Mahalanobis: Nearest 3 Neighbors | 2.37** | n.a. | 1.18 | n.a. |
| | [.02] | | [.12] | |

p-values are in parentheses. ***, **, and * indicate significiance at the .01, .05, and .10 levels, respectively.
Propensity score matching uses 36 of the 39 treated cases for both the 5 and 10 year results. The other three lie outside the region of common support. The Mahalanobis matching uses all 39 treated cases.
Columns 3 and 5 report the Chi-square statistic testing the null hypothesis that the covariates are, on average, balanced between the treated cases and their matched controls.

matching models use all 75 cases.

All seven of the estimated 5-year growth coefficients are significant at the 0.10 level or better and the average value of these coefficients is 2.21, which is smaller than the average using 1.00 as the threshold (2.51). For the 10-year effect, 4 of the 7 coefficients are significant at 0.10 level or better and the average value for the five significant effects is 1.20.

Table 9 presents the results when we use a threshold of 1.25 to define a jump in EFW. Using this criterion, we have 39 cases with data, of which 36 are in the region of common support and used to estimate the treatment effect. In this case, 4 of the 7 estimated 5-year treatment effects are significant at the 0.10 level or better and the average value of the statistically significant effects is 2.08. For the 10-year growth effects, only 3 of the 7 estimated effects are significant at 0.10 or better, and the average value of the significant effects is 1.95. As in the previous tables, our PSM method achieves covariate balance in all 8 models.

In general, then, the estimated causal effect of EFW jumps on growth is quite robust to the threshold for defining a jump. There is overwhelming evidence of a large short-run effect and mixed but substantial evidence for a significant long-run effect as well.

## 6. Conclusion

The Washington Consensus has fallen out of favor in the last few decades, but we believe that the ideas behind it have been prematurely discarded. In fact, we are not aware of a paper that convincingly rejects the efficacy of this set of reforms. We defined generalized reform as a discrete, sustained jump in an index of economic freedom, whose components are consistent with the main tenets of the Washington Consensus.

We identify 49 cases of generalized reform in a sample of 141 countries from 1970 to 2015. The average treatment effect associated with these reforms is positive, sizeable, and significant over 5- and 10- year windows. Specifically, over a 5-year horizon, we find growth is 2.07 to 2.87 percentage points higher in the treated countries compared to their matched controls. Over a 10-year horizon, we find a 1.03 – 1.93 percentage point increase compared to our counterfactual group. These effects are substantial and the results are robust to different thresholds for defining reform and different estimation methods. For these reasons, we believe it is worth giving the ideas behind the Washington Consensus a second chance.

## Appendix A

**Table A1**
Countries in the Sample

| | | | |
|---|---|---|---|
| Albania | Algeria | Angola | Argentina |
| Armenia | Australia | Austria | Azerbaijan |
| Bahamas | Bahrain | Bangladesh | Barbados |
| Belgium | Belize | Benin | Bolivia |
| Bosnia & Herzegovina | Botswana | Brazil | Bulgaria |
| Burkina Faso | Burundi | Cameroon | Canada |
| Central African Rep. | Chad | Chile | China |
| Colombia | Congo D.R. | Congo, Republic of | Costa Rica |
| Cote d'Ivoire | Croatia | Cyprus | Czech Republic |
| Denmark | Dominican Republic | Ecuador | Egypt |
| El Salvador | Estonia | Ethiopia | Fiji |
| Finland | France | Gabon | Georgia |
| Germany | Ghana | Greece | Guatemala |
| Guinea-Bissau | Guyana | Haiti | Honduras |
| Hong Kong | Hungary | Iceland | India |
| Indonesia | Iran | Ireland | Israel |
| Italy | Jamaica | Japan | Jordan |
| Kazakhstan | Kenya | Korea, South | Kuwait |
| Kyrgyz Republic | Latvia | Lesotho | Lithuania |
| Luxembourg | Macedonia | Madagascar | Malawi |
| Malaysia | Mali | Malta | Mauritania |
| Mauritius | Mexico | Moldova | Mongolia |
| Montenegro | Morocco | Mozambique | Myanmar |
| Namibia | Nepal | Netherlands | New Zealand |
| Nicaragua | Niger | Nigeria | Norway |
| Oman | Pakistan | Panama | Papua New Guinea |
| Paraguay | Peru | Philippines | Poland |
| Portugal | Romania | Russia | Rwanda |
| Senegal | Serbia | Sierra Leone | Singapore |
| Slovak Republic | Slovenia | South Africa | Spain |
| Sri Lanka | Sweden | Switzerland | Syria |
| Taiwan | Tanzania | Thailand | Togo |
| Trinidad & Tobago | Tunisia | Turkey | Uganda |
| Ukraine | United Arab Emirates | United Kingdom | United States |
| Uruguay | Venezuela | Vietnam | Zambia |
| Zimbabwe | | | |

**Table A2**
Variable Description and Sources.

| Name | Definition | Source |
|---|---|---|
| EFW | Economic freedom | Fraser Inst. (2017) |
| Income Growth | Annual % change of real per-capita income | Feenstra et al. (2015) |
| Inflation | Annual % change of the GDP deflator | World Bank (2017) |
| Investment (%) | Share of gross capital formation at current PPP | Feenstra et al. (2015) |
| Gov. Consumption (%) | Share of gov. consumption at current PPP | Feenstra et al. (2015) |
| Exports (%) | Share of merchandise exports at current PPP | Feenstra et al. (2015) |
| Human Capital | Index of avg. years of education and an assumed rate of return | Feenstra et al. (2015) |
| Polity2 | An index ranging from +10 (highly democratic) to -10 (highly autocratic). | Polity IV (2017) |

**Table A3**
Summary Statistics.

| Variable | Average | Std. Dev | N |
|---|---|---|---|
| EFW Index | 6.18 | 1.29 | 1,118 |
| Real per-capita income | 13,748 | 16,103 | 1,105 |
| Inflation | 42 | 467 | 1,077 |
| Investment (%) | 0.21 | 0.10 | 1,105 |
| Gov. Consumption (%) | 0.19 | 0.10 | 1,105 |
| Exports (%) | 0.26 | 0.26 | 1,105 |
| Human Capital Index | 2.28 | 0.71 | 1,061 |
| Polity2 Score | 3.35 | 6.90 | 1,045 |

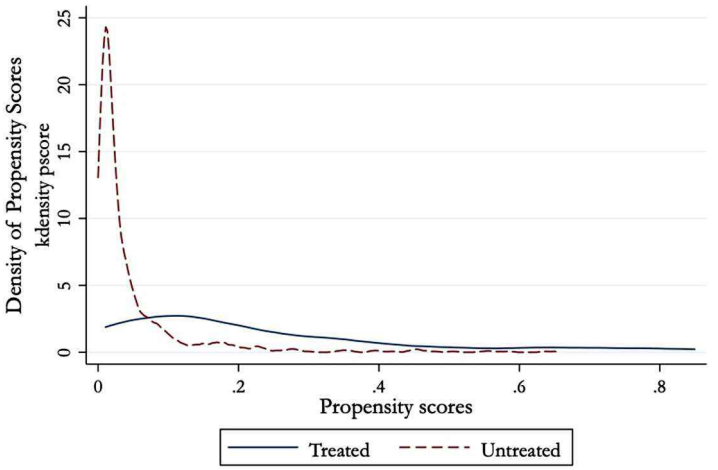


**Figure A1.** Propensity Scores by Treatment Status.

**Table A4**
The Effects of Reform on Economic Performance with Exact Matching on Year.

| Matching Method | 5-yr growth coefficient | $X^2$ covar. balance | 10-yr growth coefficient | $X^2$ covar. balance |
|---|---|---|---|---|
| Propensity Score: Nearest Neighbor | 2.81** | 4.68 | 1.50* | 4.68 |
| | [.02] | [.69] | [.10] | [.70] |
| Propensity Score: Nearest 2 Neighbors | 2.22** | 1.77 | 0.75 | 1.85 |
| | [.05] | [.97] | [.39] | [.97] |
| Propensity Score: Nearest 3 Neighbors | 2.02** | 1.58 | 0.62 | 1.59 |
| | [.05] | [.98] | [.50] | [.98] |
| Propensity Score: Normal Kernel | 2.36*** | 4.25 | 0.98 | 4.19 |
| | [.01] | [.75] | [.16] | [.76] |

p-values are in parentheses. ***, **, and * indicate significiance at the .01, .05, and .10 levels, respectively.
Columns 3 and 5 report the Chi-square statistic testing the null hypothesis that the covariates are, on average, balanced between the treated cases and their matched controls.

## Appendix B

**Table B1**
The Effects of Reform on Economic Performance using the Heritage Index

| Matching Method | 5-yr growth coefficient | $X^2$ covar. balance | 10-yr growth coefficient | $X^2$ covar. balance |
|---|---|---|---|---|
| Propensity Score: Nearest Neighbor | 1.37 | 19.41*** | 0.524 | 13.9* |
| | [.57] | [.01] | [.83] | [.06] |
| Propensity Score: Nearest 2 Neighbors | 0.90 | 2.34 | 0.281 | 13.9* |
| | [.65] | [.94] | [.89] | [.06] |
| Propensity Score: Nearest 3 Neighbors | 1.47 | 2.26 | 0.378 | 2.66 |
| | [.43] | [.94] | [.84] | [.92] |
| Propensity Score: Normal Kernel | 1.49 | 0.93 | 1.01 | 0.59 |
| | [.29] | [.99] | [.50] | [.99] |
| Mahalanobis: Nearest Neighbor | 2.23* | n.a. | 1.54 | n.a. |
| | [.10] | | [.36] | |
| Mahalanobis: Nearest 2 Neighbors | 2.70** | n.a. | 1.00 | n.a. |
| | [.02] | | [.34] | |
| Mahalanobis: Nearest 3 Neighbors | 1.68* | n.a. | 1.44 | n.a. |
| | [.08] | | [.19] | |

p-values are in parentheses. ***, **, and * indicate significiance at the .01, .05, and .10 levels, respectively.
Propensity score matching uses 7 of the 10 treated cases in the 5-year results and 5 of the 8 in the 10-year results.. The other three lie outside the region of common support. The Mahalanobis matching uses all treated cases.
Columns 3 and 5 report the Chi-square statistic testing the null hypothesis that the covariates are, on average, balanced between the treated cases and their matched controls.

Results Using the Heritage Foundation Index.

As we discussed in footnote 13, the Heritage Foundation also has an index on economic freedom. It begins in 1995, covers 186 countries, and is composed of 12 sub-components: property rights, judicial strength, government integrity, tax burden, government spending, fiscal health, and freedom of business, labor, money, trade, investment, and financial. The correlation between the EFW and Heritage Index and EFW for our sample is .83. In Table 4 of our paper (our main results), we have 49 cases of sustained jumps and in Table 5, we have 75 cases of any jumps.

Because the Heritage Index does not start until 1995, the first jump we can record is not until 2000. We find 15 jumps (most are not sustained or are too near the end of the data to tell). Of those 15, four match exactly with EFW jumps: Bulgaria (2005), Myanmar (2015), Slovak Republic (2005), and Zimbabwe (2015). However, the 2015 jumps cannot be evaluated in either case because we do not have outcome data for 2020 or 2025.

In another five cases, the EFW index also shows a jump, but one period different than the Heritage index (Guinea-Bissau (2005 v 2000), Iran (2005 v 2000), Nicaragua (2000 v 1995), Peru (2000 v 1995) Romania (2010 v 2005)). Of these five, Guinea-Bissau cannot be used in the analysis in either case because of missing covariates. In another pair of cases, the jumps are further than five years apart (Malta (2005 v 1995) and Turkey (2010 v 1995)). Malta cannot be used in the analysis in either case because of missing covariates.

For the cases of Mauritania, Moldova, Mongolia, the Heritage index shows a jump but EFW has not yet started reporting data for the country. Finally, Heritage shows a jump for Georgia in 2010 that is not reflected anywhere in EFW. This case is not used in the analysis because of missing covariates.

In sum, of the 15 Heritage Jumps, EFW shows the same pattern in nine cases, is further away in two others, is not available to be compared in three others, and disagrees with one.

Below we report the results of estimating our model on all the Heritage Index jumps (sustained or not) that have covariates for the matching. There are only 10 such cases in the 5-year results and 8 for the 10-year results. The analog to this experiment using EFW is Table 5 with over 70 cases. All the coefficients are positive and the 5-year results using Mahalanobis matching are significant. The 10-year coefficients are notably smaller than those in Table 5, but again we only have 8 cases using the Heritage index versus 70 using EFW.

## References

Abadie, Alberto, Imbens, Guido W., 2011. Bias-corrected matching estimators for average treatment effects. Journal of Business and Economic Statistics 29, 1–11.
An, Weihua, Winship, Christopher, 2017. Causal inference in panel data with application to estimating race-of-interviewer effects in the general social survey. Sociological Methods and Research 46, 68–102.
Aronow, Peter M, Samii, Cyrus, 2016. Does regression produce representative estimates of causal effects? American Journal of Political Science 60, 250–267.
Alesina, Alberto, Özler, Sule, Roubini, Nouriel, Swagel, Phillip, 1996. Political instability and economic growth. Journal of Economic Growth 1, 189–211.
Barro, Robert J., 1990. Government spending in a simple model of endogeneous growth. Journal of Political Economy 98, S103–S125.
Ben-David, Dan, Loewy, Michael B., 1998. Free trade, growth, and convergence. Journal of Economic Growth 3, 143–170.
Birdsall, Nancy, Valencia Caicedo, Felipe, De la Torre, Augusto, 2010. The Washington Consensus: Assessing a damaged brand. Working Paper No. 5316. World Bank, Washington D.C.

Clague, Christopher, Keefer, Philip, Knack, Stephen, Olson, Mancur, 1996. Property and contract rights in autocracies and democracies. Journal of Economic Growth 1, 243–276.
de Carvalho Filho, Irineu, Chamon, Marcos, 2012. The myth of post-reform income stagnation: Evidence from Brazil and Mexico. Journal of Development Economics 97, 368–386.
de Chaisemartin, Clément, D'Haultfoeuille, Xavier, 2020. Two-way fixed effects estimators with heterogeneous treatment effects. American Economic Review, forthcoming.
Dewatripont, Mathias, Roland, Gérard, 1995. The design of reform packages under uncertainty. The American Economic Review 85, 1207–1223.
Dollar, David, Kraay, Aart, 2002. Growth is good for the poor. Journal of Economic Growth 7, 195–225.
Easterly, William, 2019. In search of reforms for growth: new stylized facts on policy and growth outcomes. Working Paper No. 26318. NBER.
Eicher, Theo S., Schreiber, Till, 2010. Structural policies and growth: Time series evidence from a natural experiment. Journal of Development Economics 91, 169–179.
Estevadeordal, Antoni, Taylor, Alan M., 2013. Is the Washington consensus dead? Growth, openness, and the great liberalization, 1970s–2000s. Review of Economics and Statistics 95, 1669–1690.
Feenstra, Robert C., Inklaar, Robert, Timmer, Marcel P., 2015. The next generation of the Penn World Table. American Economic Review 105, 3150–3182.
Glick, Reuven, Guo, Xueyan, Hutchison, Michael, 2006. Currency crises, capital-account liberalization, and selection bias. The Review of Economics and Statistics 88, 698–714.
Goodman-Bacon, Andrew, 2019. Difference in differences with variation in treatment timing. Working Paper No. 25018. NBER.
Gwartney, James, Lawson, Robert A., Hall, Joshua C., Murphy, Ryan, Czeglédi, Pál, Fike, Rosemarie, McMahon, Fred, Newland, Carlos, 2018. Economic Freedom of the World: 2018 Annual Report. Fraser Institute, Vancouver, BC.
Hall, Joshua C., Lawson, Robert A., 2014. Economic freedom of the world: An accounting of the literature. Contemporary Economic Policy 32, 1–19.
Hausmann, Ricardo, Pritchett, Lant, Rodrik, Dani, 2005. Growth accelerations. Journal of Economic Growth 10, 303–329.
Heckman, James J., Ichimura, Hidehiko, Todd, Petra E., 1997. Matching as an econometric evaluation estimator: evidence from evaluating a job training programme. Review of Economic Studies 64, 605–654.
Heckman, James J., Ichimura, Hidehiko, Todd, Petra E., 1998. Matching as an econometric evaluation estimator. Review of Economic Studies 65, 261–294.
Hutchison, Michael M., Noy, Ilan, 2003. Macroeconomic effects of IMF-sponsored programs in Latin America: output costs, program recidivism and the vicious cycle of failed stabilizations. Journal of International Money and Finance 22, 991–1014.
Irwin, Douglas A., 2019. Does Trade Reform Promote Economic Growth? A Review of Recent Evidence. Working Paper No. 25927. NBER.
Karras, Georgios, 1996. The optimal government size: further international evidence on the productivity of government services. Economic Inquiry 34, 193–203.
Kerekes, Monika, 2012. Growth miracles and failures in a Markov switching classification model of growth. Journal of Development Economics 98, 167–177.
King, Gary, Zeng, Langche, 2006. The dangers of extreme counterfactuals. Political Analysis 14, 131–159.
Lawson, Robert A., Murphy, Ryan, Powell, Benjamin, 2020. The determinants of economic freedom: A survey. Contemporary Economic Policy, forthcoming.
Lin, Shu, Ye, Haichun, 2007. Does inflation targeting really make a difference? Evaluating the treatment effect of inflation targeting in seven industrial countries. Journal of Monetary Economics 54, 2521–2533.
Marshall, Monty G., Gurr, Ted R., Jaggers, Keith, 2017. Polity IV Project: Political regime characteristics and transitions, 1800-2015, Dataset Users' Manual. Available at http://www.systemicpeace.org/inscr/p4manualv2015.pdf. Centre for Systemic Peace, Vienna, VA.
Ochel, Wolfgang, Röhn, Oliver, 2006. Ranking of countries-the WEF, IMD, Fraser and Heritage indices. Working Paper No. 4. CESifo DICE.
Persson, Torsten, Tabellini, Guido, 2009. The growth effect of democracy: Is it heterogeneous and how can it be estimated? In: Helpman, Elhanan (Ed.), Institutions and Economic Performance. Harvard Univ. Press, Cambridge, MA.
Prati, Alessandro, Onorato, Massimiliano G., Papageorgiou, Chris, 2013. Which reforms work and under what institutional environment? Evidence from a new data set on structural reforms. Review of Economics and Statistics 95, 946–968.
Rodrik, Dani, 2006. Goodbye Washington consensus, hello Washington confusion? A review of the World Bank's economic growth in the 1990s: learning from a decade of reform. Journal of Economic Literature 44, 973–987.
Rodríguez, Francisco, Rodrik, Dani, 2000. Trade policy and economic growth: a skeptic's guide to the cross-national evidence. NBER Macroeconomics Annual 15, 261–325.
Sandefur, Justin, Wadhwa, Divyanshi, 2018. Why the World Bank Should Ditch the "Doing Business" Rankings—in One Embarrassing Chart. Available at https://www.cgdev.org/blog/chart-week-3-why-world-bank-should-ditch-doing-business-rankings-one-embarrassing-chart. Center for Global Development, Washington D.C.
Williamson, John, 1989. What Washington Means by Policy Reform. In: Williamson, John (Ed.), Latin American readjustment: How much has happened. Peterson Institute for International Economics, Washington D.C., pp. 90–120
Williamson, John, 1993. Democracy and the "Washington consensus." World Development 21, 1329-1336.
World Bank, 2017. World Development Indicators, Available at http://databank.worldbank.org/. Washington D.C.