NBER WORKING PAPER SERIES

THE TIME FOR AUSTERITY: ESTIMATING THE AVERAGE TREATMENT EFFECT OF FISCAL POLICY

Òscar Jordà
Alan M. Taylor

Working Paper 19414
http://www.nber.org/papers/w19414

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
September 2013

The views expressed herein are solely the responsibility of the authors and should not be interpreted as reflecting the views of the Federal Reserve Bank of San Francisco, the Board of Governors of the Federal Reserve System , or the National Bureau of Economic Research. We thank seminar participants at the Federal Reserve Bank of San Francisco, the Swiss National Bank, the NBER Summer Institute, the Bank for International Settlements, the European Commission, and HM Treasury for helpful comments and suggestions. We are particularly grateful to Daniel Leigh for sharing data and Early Elias for outstanding research assistance. All errors are ours.

At least one co-author has disclosed a financial relationship of potential relevance for this research. Further information is available online at http://www.nber.org/papers/w19414.ack

The Time for Austerity: Estimating the Average Treatment Effect of Fiscal Policy
Òscar Jordà and Alan M. Taylor
NBER Working Paper No. 19414
September 2013, Revised April 2014
JEL No. C54,C99,E32,E62,H20,H5,N10

ABSTRACT

After the Global Financial Crisis a controversial rush to fiscal austerity followed in many countries. Yet research on the effects of austerity on macroeconomic aggregates was and still is unsettled, mired by the difficulty of identifying multipliers from observational data. This paper reconciles seemingly disparate estimates of multipliers within a unified and state-contingent framework. We achieve identification of causal effects with new propensity-score based methods for time series data. Using this novel approach, we show that austerity is always a drag on growth, and especially so in depressed economies: a one percent of GDP fiscal consolidation translates into 4 percent lower real GDP after five years when implemented in the slump rather than the boom. We illustrate our findings with a counterfactual evaluation of the impact of the U.K. government's shift to austerity policies in 2010 on subsequent growth.

Òscar Jordà
Economic Research, MS 1130
Federal Reserve Bank of San Francisco
101 Market St.
San Francisco, CA 94105
and University of California, Davis
oscar.jorda@sf.frb.org

Alan M. Taylor
Department of Economics and
Graduate School of Management
University of California
One Shields Ave
Davis, CA 95616-8578
and NBER
amtaylor@ucdavis.edu

I solemnly affirm and believe, if a hundred or a thousand men of the same age, same temperament and habits, together with the same surroundings, were attacked at the same time by the same disease, that if one half followed the prescriptions of the doctors of the variety of those practicing at the present day, and that the other half took no medicine but relied on Nature's instincts, I have no doubt as to which half would escape. — Petrarch, letter to Boccaccio, 1364

The boom, not the slump, is the right time for austerity at the Treasury. — J. M. Keynes, 1937

In 1809 on a battlefield in Portugal, in a defining experiment in epidemiological history, a Scottish surgeon and his colleagues attempted what some believe to have been the first recognizable medical trial, a test of the effectiveness of bloodletting on a sample of 366 soldiers allocated into treatment and control groups by alternation. The cure was shown to be bogus. Tests of this sort heralded the beginning of the end of premodern medicine, vindicating skeptics like Petrarch for whom the idea of a fair trial was a mere thought experiment. Yet, even with alternation, allocation bias—i.e., "insufficient randomization"—remained pervasive in poor experimental designs (e.g., via foreknowledge of assignment) and the intellectual journey was only completed in the 1940s with the landmark British Medical Research Council trials of patulin and streptomycin. Ever since the randomized controlled trial has been the foundation of evidence-based medicine.[1]

Is a similar evidence-based macroeconomics possible and what can it learn from this noble scientific tradition? Ideas from the experimental approach bridge medicine, epidemiology, and statistics, and they have slowly infected empirical economics, although mostly on the micro side.[2] In this paper we delve into the experimental toolkit so as to re-examine a key issue for macroeconomics, the need to ensure treatments are somehow re-randomized in non-experimental data. We do this in the context of the foremost academic and policy dispute of the day—the effects of fiscal policy shocks on output (see, in particular, Alesina and Ardagna 2010; Guajardo, Leigh, and Pescatori 2011).[3]

[1] See Chalmers (2005, 2011), who discusses Petrarch, bloodletting, and the MRC clinical trials.

[2] Angrist and Pischke (2010) judge that "progress has been slower in empirical macro." The fear that standard empirical practices would not work, especially for aggregate economic questions, goes back a long way, at least to J. S. Mill (1836), who favored a priori reasoning alone, arguing that: "There is a property common to almost all the moral sciences, and by which they are distinguished from many of the physical; this is, that it is seldom in our power to make experiments in them."

[3] Ironically enough, fiscal policy debates are now littered with medical metaphors. In 2011 German Finance Minister Wolfgang Schäuble wrote in The Financial Times, that "austerity is the only cure for the Eurozone"; while Paul Krugman, at The New York Times, likened it to "economic bloodletting'. In the FT, Martin Wolf, cautioned that "the idea that treatment is right irrespective of what happens to the patient falls into the realm of witch-doctoring, not science." Martin Taylor, former head of Barclays, put it bluntly: "Countries are being enrolled, like it or not, in the economic equivalent of clinical trials."

Identification of the effects of fiscal consolidation in empirical studies has broadly taken one of two forms. In the context of a vector autoregression, one option is to achieve identification based on exclusion restrictions. In practice, these restrictions are roughly equivalent to a regression-control strategy based on a limited number of observable controls, and (usually) a linear conditional mean assumption. Examples of this strand of the literature include Alesina and Perotti (1995); Perotti (1999); and Mountford and Uhlig (2009). The other strand of the literature has approached the identification problem through instrumental variables. Auerbach and Gorodnichenko (2012, 2013) and Owyang, Ramey and Zubairy (2013) use local projections paired with IV estimation and a more flexible functional form.

Following a new and arguably more promising direction, we take a third fork on the road to identification based on the Rubin Causal Model. This approach has the attractive features of being semiparametric (and hence flexible with respect to the functional form), providing better observables control, and offering a more reliable alternative when the putative instrumental variables for policy action are themselves suspected of being endogenous—a serious problem which we find to be the case here and which is potentially present in many applications in the wider empirical macro literature.

We find that on average, fiscal consolidations are a drag on GDP growth. The effect is also state dependent: if a 1 percent of GDP fiscal consolidation is imposed in a slump rather than in a boom this results in real GDP being around 4 percent lower after five years. We arrive at this conclusion by painstakingly constructing an encompassing framework that allows us to evaluate the type of approach followed by several recent papers in the literature (to be discussed in detail shortly). In addition to accommodating existing methods, the framework allows us to address the identification concerns we uncover via the application of an estimator from the family of "doubly robust" augmented inverse-propensity-score weighted regression adjustment methods (Robins, Rotnizky, and Zhao 1994; Robins 1999; Scharfstein, Rotnitzky, and Robins 1999; Hirano, Imbens, and Ridder 2003; Lunceford and Davidian 2004; Imbens 2004; Glynn and Quinn 2010).

To provide more texture to our results, we evaluate the U.K. austerity program implemented by the Coalition government after the 2010 election. The Global Financial Crisis struck the U.S. and the U.K. in a similar way and these economies ran on parallel trajectories until 2010. Thereafter the U.K. experienced a second slowdown while the U.S. continued to grow. Using our estimates we compute how much of the slowdown could be attributed to the austerity program; we find it to be a very significant contribution (rising to 3.4% of GDP in 2013) and larger than official estimates. Better models with state-dependent features could improve official fiscal policy analyses going forward.

1. The Austerity Debate: A Road Map

Are fiscal consolidations expansionary, neutral, or contractionary? In order to answer this question, and understand the different answers the literature has arrived at so far, we proceed in a series of incremental stages.

First, we use the OECD annual panel dataset adopted in two recent high-profile yet seemingly irreconcilable studies. The "expansionary austerity" idea has come to be associated with the paper by Alesina and Ardagna (2010, henceforth AA) an idea perhaps dating back to at least Giavazzi and Pagano (1990). On the opposite side, the IMF team of Guajardo, Leigh, and Pescatori (2011, henceforth GLP) reached the opposite conclusion of "contractionary austerity." By juxtaposing these two papers we are not implying that the literature falls evenly or comprehensively within these two camps. We use the contraposition as a rhetorical device much like Perotti (2013), who presents a lucid discussion of the empirical pitfalls in this research area.

Second, we use (Jordà 2005) local projections (LPs) to estimate output impacts of fiscal policy dynamically up to 5 years out. LPs are a flexible semi-parametric regression control strategy to estimate dynamic multipliers and include, as a special case, impulse responses calculated by commonly used vector autoregressions (VARs). LPs accommodate possibly nonlinear, or state-dependent responses easily, and indeed we find that the effects of fiscal policy can be very different in the boom and the slump, as emphasized by Keynes in the 1930s. State-dependent multipliers based on LPs have been taken up in some very recent papers (Auerbach and Gorodnichenko 2012, 2013, for the US and OECD; Owyang, Ramey, and Zubairy 2013, for U.S and Canada). Other recent papers on state-dependent multipliers, using various measures of slack, include Barro and Redlick (2011) and Nakamura and Steinsson (2014). For a critical survey see Parker (2011). Long ago, Perotti (1999) explored the idea of "expansionary austerity" with state-dependent multipliers.

We calculate the impact of fiscal policy shocks based on LPs using the AA measure of policy, the change in the cyclically-adjusted primary balance (d.CAPB).[4] When we restrict attention to "large" shocks (changes in CAPB larger in magnitude than 1.5% of GDP, which is the benchmark cutoff value used by AA and proposed earlier by Alesina and Perotti 1995), we replicate the "expansionary austerity" result. However, when we condition on the state of the economy, we find that this result is driven entirely by what happens during a boom. When the economy is in a slump the expansionary effects of fiscal consolidation evaporate.

[4]The d.CAPB measure used by AA is based on Blanchard (1993). The construction of this variable consists of adjusting for cyclical fluctuations using the unemployment rate.

Third, we then use instrumental variable (IV) estimation of the LPs to account for unobserved confounders. Specifically, we instrument the cyclically-adjusted primary balance using the IMF's narrative measure of an exogenous fiscal consolidation in GLP. This type of "narrative-based identification" has been applied by, e.g., Ramey and Shapiro (1998) and Romer and Romer (1989, 1997). Our IV estimation then turns out to replicate the flavor of the GLP results: austerity is contractionary, and strongly so in slumps.

Fourth, we show that the proposed IMF narrative instrumental variable has a significant forecastable element driven by plausible state variables, such as the debt-to-GDP level, the cyclical level or rate of growth of real GDP, and the lagged treatment indicator itself (since austerity programs are typically persistent, multi-year affairs).[5] This calls into question the validity of the narrative instrumental variable. As noted above in the history of medicine, and as with any efforts to construct a narrative policy variable that is exogenous, one has to worry about the possibility that treatment is still contaminated by endogeneity, which would impart allocation bias to any estimates.[6]

Fifth, in order to purge remaining allocation bias we use inverse probability weighting (IPW) estimation to estimate the LP responses. We consider the IMF narrative instrumental variable as a "fiscal treatment"—i.e., a binary indicator rather than a continuous variable—and we are interested in characterizing a dynamic *average treatment effect* (ATE). In new work, Angrist, Jordà, and Kuersteiner (2013) introduce IPW estimators in a time series context to calculate the dynamic ATE responses to policy interventions. We follow a slightly different approach using augmented regression-adjusted estimation instead, denoted AIPW, which combines inverse probability weighting with regression control and adjusts the estimator to achieve semi-parametric efficiency (see, e.g., Lunceford and Davidian 2004). Our AIPW estimator falls into the broad class of "doubly robust" estimators of which Robins, Rotnitzky, and Zhao (1994) is perhaps the earliest reference (see also Robins 1999; Scharfstein, Rotnitzky, and Robins 1999; Hirano, Imbens, and Ridder 2003; Lunceford and Davidian 2004; Imbens 2004; Glynn and Quinn 2010). The "doubly robust" property means that consistency of the estimated ATE is achieved when either the propensity score model and/or the conditional mean is correctly specified.

What do we find? Our results contrast with the expansionary austerity view of AA, and rather amplify the opposing view of GLP: we find that austerity is very contractionary.

[5]The potential endogeneity of fiscal consolidation episodes has been noted by other authors. For example, Ardagna (2004) uses political variables as an exogenous driver for consolidation in a GLS simultaneous equation model of growth and consolidation for the period 1975–2002. Hernández De Cos and Moral-Benito (2013) use economic variables as instruments.

[6]For example, in the debate over the use of narrative methods to assess monetary policy, see the exchange between Leeper (1997) and Romer and Romer (1997).

The effect in slumps is much stronger and of even higher statistical significance; and even in booms there are signs of drag on growth over the five-year horizon.

Allocation bias is therefore a serious empirical issue for the fiscal policy debate. In the historical sample under dispute, policymakers have tended to impose austerity in bad times. Thus, what we have been seeing in the U.K. and Eurozone austerity experiments in the aftermath of the Global Financial Crisis are not unusual in timing, even if the policy shocks are large in scale. These events are out-of-sample for our study, but past austerity has generally been applied in weak economic conditions: *plus ça change*. Yet when it is in a bad current state the economy is more likely to grow faster than trend going forward, simply by construction. By failing to allow for this treatment selection we can end up with far too rosy and optimistic estimates of the effects of fiscal consolidation: a dead cat bounces, regardless of whether it jumped or was pushed.

2. Identification and OLS/IV Estimation of Fiscal Multipliers

The key starting point for our analysis is the idea that fiscal policy is rarely the result of random experimentation. Automatic stabilizers swell the public deficit when economies are in recession. In financial crises, banking sector debts gone bad may be absorbed by the sovereign. And when debt-to-GDP ratios challenge the comfort levels of bond markets or governments, sovereigns will be more likely to consolidate their fiscal balances. In calculating what the counterfactual path of the economy would have been under an alternative fiscal policy intervention, historical data are likely to be a poor control. Much of the variation in fiscal policy is the result of endogenous factors.

Teasing causal effects from observational data is difficult. It depends crucially on the interplay between the modeling approach and identification assumptions. This section introduces the basic LP framework used in the remainder of the paper. In order to facilitate cleaner notation, we drop the country index of the cross section in the panel of data that we later investigate. When appropriate, we discuss the idiosyncrasies of panel data implementation, such as the inclusion of fixed effects in the conditional mean, or calculation of cluster-robust standard errors.

Denote by y_t an outcome variable of interest, say the log of real GDP. More generally, y_t could be a k_y-dimensional vector. Let D_t denote the fiscal policy variable. In the analysis of this section, D_t is a continuous random variable although later in the paper, we will treat D_t as a discrete random variable that can only take two values, $D_t = 0, 1$. In addition, we consider the possibility that there is a k_w-dimensional vector of variables, w_t that are not included in the vector y_t, but which could be relevant predictors of the policy variable

D_t. Instrumental variables, when available, will be collected in the k_z-dimensional vector z_t. Finally, denote X_t the rich conditioning set given by $\Delta y_{t-1}, \Delta y_{t-2}, ...; D_{t-1}, .D_{t-2}, ...;$ and w_t. We assume that policy is determined by $D_t = D(X_t, \psi, \varepsilon_t)$ where ψ refers to the parameters of the implied policy function and ε_t is an idiosyncratic source of random variation. Therefore, $D(X_t, \psi, .)$ refers to the systematic component of policy determination.

To make further progress at this point we will borrow from definition 1 in Angrist, Jordà, and Kuersteiner (2013, henceforth AJK). This defines *potential outcomes* given by $y^{\psi}_{t,h}(d) - y_t$ as the value that the observed outcome variable $y_{t+h} - y_t$ would have taken if $D_t = d$ for all $\psi \in \Psi$ and $d \in \mathcal{D}$. In our application, the difference $y_{t+h} - y_t$ refers to the cumulative change in the outcome from t to $t + h$. The horizon h can be any positive integer. When the policy intervention is continuous, we use $D_t = d_1$ to indicate an intervention of size d_1 that we want to compare to a benchmark given by $D_t = d_0$, where usually $d_0 = 0$, but in general it need not be. Later on, when the policy intervention variable is binary, we use $D_t = 1$ to indicate policy intervention and $D_t = 0$ for no intervention.

The causal effect of a policy intervention is defined as the unobservable random variable given by the difference $(y_{t,h}(d_1) - y_t) - (y_{t,h}(d_0) - y_t)$. Notice that y_t is only used to benchmark the cumulative change and it is observed at time t. We assume that the parameters of the policy function do not change.

Following AJK, we can state the selection-on-observables assumption (or the *conditional ignorability* or *conditional independence* assumption as it is sometimes called) as

$$(y^{\psi}_{t,h}(d) - y_t) \perp D_t | X_t; \psi \qquad \text{for all } h \geq 0, \text{and for } d \in \mathcal{D}, \text{and } \psi \in \Psi. \tag{1}$$

That is, the treatment-control allocation is independent of potential outcomes given controls X_t. This condition does not imply that there is no effect of policy on the outcome given controls. We are simply stating that conditional on controls, policy allocation is independent of the *potential* outcome, whatever that might be. To further understand the role that the conditional independence assumption plays, consider the ideal randomized experiment first.

Suppose that policy intervention can only take two values, $D_t = 0, 1$, and that the treatment allocation to either bin is completely random. The average causal effect of policy intervention on the outcome at time $t + h$ given by

$$E\left[(y_{t,h}(1) - y_t) - (y_{t,h}(0) - y_t))\right]$$

could be simply calculated using *group means* as

$$\hat{\Lambda}^h_{GroupMean} = \frac{1}{n_1}\sum_t D_t(y_{t+h} - y_t) - \frac{1}{n_0}\sum_t (1 - D_t)(y_{t+h} - y_t) \qquad \text{for all } h \geq 0, \tag{2}$$

where $n_1 = \sum_t D_t$ and $n_0 = \sum_t (1 - D_t)$ are the number of observations in treatment and control groups, respectively.

Alternatively, the average treatment effect, Λ^h, could be calculated from the auxiliary regression

$$(y_{t+h} - y_t) = D_t\alpha_1^h + (1 - D_t)\alpha_0^h + v_{t+h} \qquad \text{for all } h \geq 0. \tag{3}$$

The difference in the OLS estimates of the intercepts $\hat{\alpha}_1^h - \hat{\alpha}_0^h = \hat{\Lambda}^h$ in expression (3) is equivalent to that in expression (2).

Even when data are randomly allocated across the treatment and control subpopulations, it would be natural to condition on the X_t to adjust for small sample differences in characteristics between subpopulations and to gain in efficiency, even though the estimator is consistent for the average treatment effect (ATE hereafter) whether or not regressors are included. Notice that the model for the outcomes is unspecified. The estimate of the ATE does not depend on specific assumptions about this model if the conditional ignobility assumption is met.

Allocation to treatment and control groups is not usually random in observational data. To appreciate the role of the selection-on-observables assumption in (1), consider elaborating on the example. First, by the law of iterated expectations, we can write

$$\begin{aligned} & E\left[(y_{t,h}(1) - y_t) - (y_{t,h}(0) - y_t)\right] \\ &= E\left[E\left[y_{t+h} - y_t | D_t = 1; X_t\right] - E[y_{t+h} - y_t | D_t = 0; X_t]\right] \\ &= \Lambda^h \qquad \text{for all } h \geq 0. \end{aligned} \tag{4}$$

Assume that a linear regression control strategy suffices to do the appropriate conditioning for the X_t and hence obtain a consistent estimate of $E[y_{t+h} - y_t | D_t, X_t]$. This is a big assumption that we relax later on in the paper. Then the average causal effect of a policy intervention on the outcome variable at time $t + h$ in the maintained example, can be calculated by expanding expression (3) with

$$y_{t+h} - y_t = D_t\alpha_1^h + (1 - D_t)\alpha_0^h + D_t X_t\beta_1^h + (1 - D_t)X_t\beta_0^h + v_{t+h} \quad \text{for all } h \geq 0. \tag{5}$$

If one imposes the constraint $\beta_1^h = \beta_0^h$, then expression (5) is nothing more than a standard LP and $\Lambda^h = \alpha_1^h - \alpha_0^h$ is the policy response at horizon h. The standard linear

LP is a direct estimate of the typical impulse response derived from a traditional VAR, as Jordà (2005) shows. This naïve constrained specification imposes two implicit assumptions seldom appreciated in the VAR literature. First, the effect of the controls X_t is assumed to be stable across the treated and control groups. Second, the expected value of X_t in each subpopulation is assumed to be the same. The first assumption is potentially defensible. The mechanism describing the effect of interest rates on real GDP may well be the same whether or not there is a fiscal consolidation, for example. The second assumption is another matter. It is unlikely that, say, government debt levels are the same in the treated and control groups. Fiscal consolidations are often driven by high levels of debt.

Using expressions (4) and (5), notice that

$$\begin{gathered} E[E[(y_{t,h}(1)-y_t)-(y_{t,h}(0)-y_t)|X_t]] = \\ E\left\{E\left[(D_t\{\hat{\alpha}_1+X_t\hat{\beta}_1^h\})-((1-D_t)\{\hat{\alpha}_0+X_t\hat{\beta}_0^h\})|X_t\right]\right\} = \\ E[\hat{\alpha}_1^h-\hat{\alpha}_0^h]=E[\hat{\Lambda}^h]=\Lambda^h, \end{gathered}$$

under the maintained assumptions of the example that $E(X_t|D_t=1)=E(X_t|D_t=0)$ and $\beta_1=\beta_0$ and noticing that $E(D_t|D_t=1)=E((1-D_t)|D_t=0)=1$.

More generally, if we do not impose the implicit assumptions of the naïve LP specification, the analogous representation to the group means expression (2) is

$$\hat{\Lambda}_{RA}^h=\frac{1}{n_1}\sum_t D_t(m_1^h(X_t,\hat{\theta}_1^h))-\frac{1}{n_0}\sum_t(1-D_t)(m_0^h(X_t,\hat{\theta}_0^h)) \qquad \text{for all } h\geq 0, \tag{6}$$

where $m_j^h(.)$ is a generic specification of the conditional mean of $(y_{t+h}-y_t)$ in each subpopulation $j=1,0$ and $\theta_j^h=(\alpha_j^h \quad \beta_j^h)'$ for the regression example in (5). The n_1 and n_0 have been defined earlier. Note that this more general form of *regression adjustment* allows the conditional means to be different for the treated and control subpopulations *and* allows their effect to differ as well.

The assumption of selection-on-observables implies that, conditional on a possibly large set of controls, variation in policy interventions is largely random. However, if policy interventions conditional on controls are systematically determined by an unobserved variable that is correlated with the outcome, we will fail to measure the true causal effect of fiscal policy once again.

A solution to this conundrum can be found if instrumental variables (IVs) are available. Rather than relying on a richly saturated specification of the conditional mean to achieve exogenous variation in the policy intervention, IV methods rely on controls thought to

vary exogenously with respect to the selection mechanism driven by the unobservable covariates. If there is correlation between the instruments and the policy variable, then one has a source of exogenous variation in policy interventions with which to estimate the causal effect. In the relevant literature, Auerbach and Gorodnichecko (2013) and Owyang, Ramey, and Zubairy (2013) have used the restricted version of this approach.

If instrumental variables z_t (as defined earlier) are independent of the unobserved selection mechanism and relevant for D_t, then estimation of the response to policy interventions in expression (4) using local projections in expression (5) but estimated with IV based on z_t will deliver a consistent estimate of Λ_h. Specifically, consider the Group Means estimator in expression (2) based on the following two-stage least squares strategy. In the first stage, estimate a binary dependent variable model (such as a probit or logit) of D_t on $Z_t = (z_t \quad X_t)$ from which one obtains an estimate of $P(D_t = 1|Z_t)$ and which we denote $\hat{p}_t$.

Using this estimate, the Group Means estimator in expression (2) simply becomes:

$$\hat{\Lambda}^h_{GMIV} = \frac{\sum_t \hat{p}_t D_t (y_{t+h} - y_t)}{\sum_t D_t \hat{p}_t} - \frac{\sum_t (1-\hat{p}_t)(1-D_t)(y_{t+h} - y_t)}{\sum_t (1-D_t)(1-\hat{p}_t)} \quad \text{for all } h \geq 0. \qquad (7)$$

The easiest way to see this is to define $\zeta_t = \hat{p}_t D_t + (1-\hat{p}_t)(1-D_t)$. If the instruments are valid, then $E(\zeta_t' v_{t+h}) = 0$. Expression (7) is the result of the estimator based on this moment condition. The equivalent to expression (6) requires that the θ^h_j in $m^h_j(X_t, \hat{\theta}^h_j)$ for $j = 1, 0$ be estimated by instrumental variable techniques (see Wooldridge 2010 and references therein). As a way to draw a closer parallel to the estimator in expression (13) below, notice that expression (7) can be rewritten as

$$\hat{\Lambda}^h_{GMIV} = \frac{1}{n_1} \frac{\sum_t \hat{p}_t D_t (y_{t+h} - y_t)}{\bar{p}_1} - \frac{1}{n_0} \frac{\sum_t (1-\hat{p}_t)(1-D_t)(y_{t+h} - y_t)}{(1-\bar{p}_0)} \quad \text{for all } h \geq 0,$$

where $\bar{p}_j = \sum_t \hat{p}_t \mathbf{1}(D_t = j)/n_j$.

In summary, local projection methods afford a very straightforward way to contrast the effect of estimating fiscal multipliers under implicit selection-on-observables assumptions (OLS) relative to estimates where that assumption fails, due to selection-on-unobservables, but instruments are available (IV). We carry the estimation by imposing the assumption $\beta_1 = \beta_0$ to preserve the setup commonly used in this literature where the focus is on matching the impulse response analysis typical in a VAR. In later sections we relax this assumption. The differences between the two estimators is revealing, and forms the basis of the next two sections. And yet, we then show that the exogeneity of the instruments is violated, and discuss how to also deal with that problem in a compatible framework.

3. Replicating Expansionary Austerity: OLS Results

Our first estimates use OLS estimation with the LP method, based on what is the traditional variable in the literature, the change in the cyclically adjusted primary balance (denoted d.CAPB), the same variable used by Alesina and Perotti (1995) and by AA, and used as a reference point by GLP in the IMF study. The local projection is done from year 0, when a policy change is assumed to be announced, with the fiscal impacts first felt in year 1, consistent with the timing in GLP. The LP output forecast path is constructed out to year 5, and deviations from year 0 levels are shown, and also the sum of these deviations, or "lost output" across all of those five years.

To create a benchmark estimating equation that mimics the standard setup in the literature, the typical LP equation that we estimate follows from equation (5) and has the form

$$y_{i,t+h} - y_{i,t} = \alpha_i^h + \Lambda^h D_{i,t+1} + \beta_{L0}^h \Delta y_{i,t} + \beta_{L1}^h \Delta y_{i,t-1} + \beta_C^h y_{i,t}^C + v_{i,t+h}, \tag{8}$$

for $h = 1, ..., 5$, and where $y_{i,t+h} - y_{i,t}$ denotes the cumulative change from time t to $t+h$ in 100 times the log of real GDP, the α_i^h are country-fixed effects, and $D_{i,t}$ denotes the d.CAPB policy variable (measured from time t to time $t+1$ given the assumed timing of the announcement and implementation of fiscal plans). Finally, the term $y_{i,t}^C$ denotes the cyclical component of GDP measured as deviations from an HP trend estimated with a smoothing parameter of $\lambda = 100$. We use the subscripts $L0$ and $L1$ for the β parameters associated to $\Delta y_{i,t-l}$ for $l = 0, 1$ so as not to confuse them with the $j = 1, 0$ treatment-control index.

The specification (8) nests the main elements in AA and GLP to facilitate comparisons of our results with theirs. The coefficient Λ^h is the parameter governing the impact of the continuous policy treatment measured by d.CAPB and corresponds to the constrained version of expression (5) where we have rearranged that expression to get a direct estimate of Λ^h from the regression output, but it is otherwise specified the same way.

Table 1 reports estimates based on expression (8). Estimated log real GDP impacts ($\times 100$) for each year are reported in columns 1 to 5, and for the 5-year sum of the log deviations in the final column 6. In parallel with the main result in AA, although the effects are seen to be economically modest, the data appear to support the notion that fiscal consolidation can be expansionary (especially in the first two years), although the cumulative effect over a five year period is small and negative. If we focus on multiplier estimates based on large consolidations (i.e., changes in CAPB larger than 1.5 percent of GDP using the Alesina and Perotti (1995) and AA cutoff value), then the results are almost identical. Small consolidation packages have a small effect, but the estimates are

Table 1: *Fiscal multiplier, effect of d.CAPB, OLS estimates*

Deviation in log real GDP (relative to Year 0, $\times 100$)

	(1) Year 1	(2) Year 2	(3) Year 3	(4) Year 4	(5) Year 5	(6) Sum
Fiscal multiplier, full sample	0.11**	0.12**	-0.04	-0.21***	-0.32**	-0.42**
	(0.04)	(0.05)	(0.04)	(0.07)	(0.12)	(0.16)
Observations	457	440	423	406	389	389
Fiscal multiplier, large change in CAPB ($> 1.5\%$)	0.12**	0.13**	-0.04	-0.23***	-0.33**	-0.41*
	(0.04)	(0.05)	(0.04)	(0.07)	(0.12)	(0.19)
Observations	457	440	423	406	389	389
Fiscal multiplier, small change in CAPB ($\leq 1.5\%$)	0.06	0.11	0.03	-0.07	-0.23	-0.53
	(0.07)	(0.15)	(0.14)	(0.19)	(0.28)	(0.50)
Observations	457	440	423	406	389	389

Standard errors (clustered by country) in parentheses. ***/**/* indicate $p < 0.01/0.05/0.10$.
Additional controls: cyclical component of y, 2 lags of change in y, country fixed effects.

imprecise.

Would the picture change much if we broke down the analysis of the impact of consolidation as a function of whether the economy is experiencing a boom or a slump? Estimation was next carried out on two bins of the data to allow responses to be state dependent. We sort on the sign of y^C, the time-0 cyclical component of log output (HP filtered) into "boom" and "slump" bins, to capture conditions at time 0 varying across the cycle. This partition places just over 200 observations in each the "boom" bin and the "slump" bin, given the AA-GLP combined dataset with about 450 observations in total, after allowing for observations lost due to lags.

Table 2 shows OLS estimated responses using expression (8) by sorting the data into these two bins. Panel (a) shows the estimated response coefficient at year h based on values of d.CAPB common to the AA and GLP datasets. Panel (b) shows results when we estimate separate response coefficients for "large" and "small" changes in d.CAPB, following the 1.5% of GDP cutoff value employed by Alesina and Perotti (1995) and by AA. These distinctions prove to be relatively unimportant since, as can be seen, all of the action is driven by "large" changes, with similar coefficients on the "large" changes in panel (b) and all changes in panel (a). In panel (b), the coefficients for "small" changes are small and not statistically significant at conventional levels. This is similar to what we found in Table 1.

The results are reasonable and consistent with the literature, and particularly the GLP replication of the AA-type results. The OLS estimates suggest that fiscal austerity

Table 2: *Fiscal multiplier, effect of d.CAPB, OLS estimates, booms v. slumps*

Deviation in log real GDP (relative to Year 0, $\times 100$)

(a) Uniform effect of d.CAPB changes						
	(1) Year 1	(2) Year 2	(3) Year 3	(4) Year 4	(5) Year 5	(6) Sum
Fiscal multiplier, $y^C > 0$, boom	0.21***	0.24***	0.05	-0.17	-0.22	-0.02
	(0.07)	(0.07)	(0.05)	(0.11)	(0.15)	(0.24)
Observations	222	205	192	180	175	175
Fiscal multiplier, $y^C \leq 0$, slump	-0.03	-0.07	-0.17	-0.23*	-0.41**	-0.98**
	(0.04)	(0.07)	(0.11)	(0.12)	(0.18)	(0.40)
Observations	235	235	231	226	214	214
(b) Separate effects of d.CAPB for large ($> 1.5\%$) and small ($\leq 1.5\%$) changes in CAPB						
	(1) Year 1	(2) Year 2	(3) Year 3	(4) Year 4	(5) Year 5	(6) Sum
Fiscal multiplier, large change in CAPB, $y^C > 0$, boom	0.23**	0.24***	0.06	-0.15	-0.18	0.13
	(0.08)	(0.08)	(0.06)	(0.11)	(0.15)	(0.28)
Fiscal multiplier, small change in CAPB, $y^C > 0$, boom	0.06	0.21	-0.04	-0.32	-0.57	-1.55
	(0.11)	(0.35)	(0.40)	(0.37)	(0.41)	(1.14)
Observations	222	205	192	180	175	175
Fiscal multiplier, large change in CAPB, $y^C \leq 0$, slump	-0.02	-0.05	-0.18	-0.30*	-0.52**	-1.16*
	(0.05)	(0.08)	(0.13)	(0.16)	(0.23)	(0.56)
Fiscal multiplier, small change in CAPB, $y^C \leq 0$, slump	-0.05	-0.16	-0.10	0.13	0.17	0.03
	(0.12)	(0.21)	(0.23)	(0.32)	(0.49)	(1.10)
Observations	235	235	231	226	214	214

Standard errors (clustered by country) in parentheses. ***/**/* indicate $p < 0.01/0.05/0.10$.
y^C is the cyclical component of $\log y$ (log real GDP), from HP filter with $\lambda = 100$.
Additional controls: cyclical component of y, 2 lags of change in y, country fixed effects.
The boom bin is for observations where the cyclical component y^C is greater than zero, the slump bin is for observations where the cyclical component is less than or equal to zero.
Large consolidations means larger than 1.5% of GDP; small means less than or equal to 1.5% of GDP.

is expansionary, since the only statistically significant coefficients are ones that have a positive sign. However, our stratification of the results by the state of the cycle at time 0 brings out a new insight, and shows that this result is entirely driven by what happens in booms. It is only in the boom bin that we find a significant positive response of real GDP to fiscal tightening, with a coefficient or "multiplier" (the more general usage of the term, which we follow in the remainder of the paper) of nearly 0.25 in years 1 and 2. Over 5 years the sum of these effects is small, also nearly 0.15. In the slump bin, the estimate of the policy response is not statistically different from zero and in many cases it is negative.

4. Replicating Contractionary Austerity: IV Results

One widely shared concern with the OLS estimates just discussed is that the policy measure d.CAPB may be highly imperfect for the job. It likely suffers from both measurement error and endogeneity. A recent frank discussion of the measurement problems with this concept is presented by Perotti (2013). Moreover, to disentangle the true cyclical component of this variable from the observed actual level outcome has to rely on modeling assumptions about the sensitivity of taxes and revenues to the cycle—effects which may be only imprecisely estimated, and which may not be stable over time or across countries. If that attempt at purging the cyclical part of the variable still leaves some endogenous variation in d.CAPB, then the implicit assumption of exogeneity needed for a causal estimate and policy analysis would be violated.

One potential solution therefore is to seek a different and more direct measure of underlying fiscal policy change, using the so-called "narrative approach" (Romer and Romer 1989). This was the arduous strategy adopted by the IMF's GLP study, which went back over 17 OECD countries and estimated the timing and magnitude of fiscal policy shocks on a year-by-year basis, based on documentary evidence from each country concerning the policies enacted since the 1970s. GLP focused exclusively on fiscal consolidation episodes, where authorities sought to reduce their budget deficit, and they sought events that were not reactions to the contemporaneous or prospective economic conditions, so that they could claim plausible exogeneity. We employ the IMF narrative measures in two ways: much of time we use an indicator of a fiscal treatment (denoted Treatment) which is simply a country-year event binary 0-1 dummy that shows when a consolidation is taking place; the other variable of interest is the IMF's estimate of the magnitude of the consolidation measures in that year as a percent of GDP (denoted Total), and which provides a scaled measure of that year's austerity package.

To bring this IMF approach into our framework, and consistent with our OLS replication of the AA results above, we present in Tables 3, and 4 our IV estimates which make use of the IMF variables. We reestimate expression (8) using the IMF dates of fiscal consolidations as both binary and continuous instruments. If the IMF approach is correct and has found truly exogenous shocks to fiscal policy, then it would be a valid instrument for d.CAPB. It would also be a potentially strong instrument: the raw correlation between d.CAPB (year 1 versus year 0) and Treatment (in year 1) is 0.31, and a bivariate regression has an F-statistic of 51; the same applies when Treatment is replaced by Total (in year 1).

We begin by reestimating the full sample specification reported in the top panel of Table 1 using instrumental variables in two ways. First we use the IMF narrative variables

Table 3: *Fiscal multiplier, effect of d.CAPB, IV estimates*

Deviation in log real GDP (relative to Year 0, $\times 100$)

	(1) Year 1	(2) Year 2	(3) Year 3	(4) Year 4	(5) Year 5	(6) Sum
Fiscal multiplier,	-0.34**	-0.72***	-0.76***	-0.78***	-0.88***	-2.94***
binary Treatment IV	(0.12)	(0.23)	(0.25)	(0.23)	(0.28)	(0.84)
Observations	457	440	423	406	389	389
Fiscal multiplier,	-0.46***	-0.81***	-0.69**	-0.58*	-0.68**	-2.77**
continuous Total IV	(0.13)	(0.23)	(0.31)	(0.28)	(0.30)	(0.97)
Observations	457	440	423	406	389	389

Standard errors (clustered by country) in parentheses. ***/**/* indicate $p < 0.01/0.05/0.10$.
Additional controls: cyclical component of y, 2 lags of change in y, country fixed effects.
d.CAPB instrumented by IMF fiscal action variable in binary 0-1 form (Treatment) in the top panel, and as a continuous (Total) variable in the bottom panel.

on dates of fiscal consolidation as a binary instrument (first row). Second, for a continuous IV we use the size of the consolidation identified by the IMF (second row). The results are reported in Table 3. Strikingly, the message here completely overturns the findings in Table 1. This is of course a well known problem, consistent with the pronounced divergence between the AA and GLP results. Fiscal consolidation is unambiguously contractionary. Using the sum of coefficients reported in column (6) of Table 3, for every 1% in fiscal consolidation, the path of real GDP is pushed down by over 0.57 percent each year on average over the five subsequent years. This result is not sensitive to whether we use the binary or continuous instrument.

The previous section broke down the analysis as a function of whether the economy is in a boom or a slump. For completeness and as a check that the IV results in Table 3 are robust, we reproduced much of the analysis in Table 2 using instrumental variables based on the binary version of the IMF narrative variable. These results are reported in Table 4. Almost identical results (not shown) arise when the continuous IV is used, so the precise choice of IV makes very little difference to the overall message.

The IV-based responses suggest that austerity is contractionary since the only statistically significant coefficients here have a negative sign. However, stratification by the state of the cycle shows that this result is now driven by what happens in slumps. It is only in the slump bin that we find a significant negative response of real GDP to fiscal tightening. In Table 4 we find a coefficient or "multiplier" of between -0.25 and -0.95 in years 1 to 5. Over five years the sum of these effects is -3.35**, so the average loss for a 1% of GDP fiscal consolidation is to depress the output level by about -0.67% per year over this horizon.

Table 4: *Fiscal multiplier, effect of d.CAPB, IV estimates (binary IV), booms v. slumps*

Deviation in log real GDP (relative to Year 0, $\times 100$)

	(1) Year 1	(2) Year 2	(3) Year 3	(4) Year 4	(5) Year 5	(6) Sum
Fiscal multiplier, $y^C > 0$, boom	-0.34	-0.32	-0.13	-0.59	-0.81	-1.36
	(0.33)	(0.50)	(0.51)	(0.52)	(0.59)	(1.78)
Observations	222	205	192	180	175	175
Fiscal multiplier, $y^C \leq 0$, slump	-0.25	-0.76***	-0.95***	-0.79**	-0.93*	-3.35**
	(0.15)	(0.25)	(0.31)	(0.33)	(0.45)	(1.19)
Observations	235	235	231	226	214	214

Standard errors (clustered by country) in parentheses. ***/**/* indicate $p < 0.01/0.05/0.10$.
The boom bin is for observations where the cyclical component y^C is greater than zero, the slump bin is for observations where the cyclical component is less than or equal to zero.
y^C is the cyclical component of $\log y$ (log real GDP), from HP filter with $\lambda = 100$.
Additional controls: cyclical component of y, 2 lags of change in y, country fixed effects.
d.CAPB instrumented by IMF fiscal action variable in binary 0-1 form (treatment).

5. Endogenous Austerity: Is the Narrative Instrument Valid?

So far we have briefly replicated the current state of the literature, but this is not entirely pointless. It serves to show that the LP framework can capture different sides of the debate in a uniform empirical design, on a consistent data sample, allowing us to focus on how differences in estimation and identification assumptions lead to different results. Our work also shows that the LP estimation method makes it very easy to allow for nonlinearity and do a stratification of the results; here we found significant variations in responses across the two bins, a setup designed to capture variations in the state of the economy from boom to slump. We found that indeed fiscal impacts vary considerably across these states in a manner that is intuitive and not unexpected: the output response to fiscal austerity is less favorable the weaker is the economy. Does this mean that Keynes was right?

Before drawing any conclusions we evaluate whether the IMF narrative variable is a legitimate instrument. Have we identified the causal effect of fiscal consolidations on output? If the IMF's narrative variable can be predicted by excluded controls and those controls are correlated with the outcome, we will have failed to resolve the allocation bias in our estimates. The IMF narrative variable will not truly be the exogenous variable on which to make solid causal inferences about policy impacts. This possible shortcoming of the "narrative identification" strategy has been noted before in the context of monetary policy (Leeper 1997) and we have the same concern here. To address this issue we report

Table 5: *Checking for balance in treatment and control sub-populations*

	Difference (Treated minus Control)	
Public debt to GDP ratio	0.13*	(0.03)
Deviation of log output from trend	-0.72*	(0.20)
Output growth rate	-0.63*	(0.18)
Treatment (lagged)	0.56*	(0.04)
Observations	491	

Standard errors in parentheses. ***/**/* indicate $p < 0.01/0.05/0.10$.

three diagnostic tests in this section in Tables 5, 6, and 7.

In the ideal randomized controlled trial, with treatment and control units allocated randomly, the probability density function of each of the controls would be the same for each subpopulation—there would be perfect overlap between the two subpopulations. A simple way to check for this *balance* condition, as it is often referred to in the literature, is to do a test of the equality of the means across subpopulations. Notice that the balance condition also lies behind the implicit assumption that one can estimate the LP by restricting the coefficient of the controls to be the same for the treatment and control groups. The balance condition is evaluated in Table 5 for several potentially important macroeconomic control variables. The null hypothesis of balance is rejected for all of them, strongly suggesting that the IMF narrative dates are not truly exogenous events.

We go beyond this simple check and next evaluate two additional identification conditions. First, we can check if the outcome is predictable by a set of available controls not yet included in the analysis. To be clear, the original AA and GLP papers do include in their analysis a robustness check that includes other controls. However, the controls they consider are typically related fiscal variables rather than the set of macroeconomic controls we consider here.

In Table 6 we report the results of such tests by reexamining whether our candidate model in expression (8) admits as additional explanation the following variables: real GDP growth; real private loan growth; CPI inflation; the change in the investment to GDP ratio; the short-term interest rate on government securities (usually 3-months in maturity); the long-term rate on government securities (usually 5–10 year bonds); and the current account to GDP ratio. The first 3 variables are expressed as 100 times the log difference. In all cases, we consider the value of the variable and one lag. The tests are conducted with the 1-period ahead local projection (the equivalent of the corresponding equation in a VAR) using the full sample according to expression (8).

The objective is to set a higher bar for the possibly omitted regressors to be significant. Partitioning the sample into the growth bins we used earlier could generate spurious

Table 6: *Omitted variables explain output fluctuations*

Model	OLS	IV (binary)	IV (continuous)
Real GDP growth	0.00	0.00	0.00
Real private loan growth	0.24	0.56	0.54
CPI Inflation	0.00	0.00	0.00
Change in investment to GDP ratio	0.11	0.00	0.00
Short-term interest rate	0.00	0.00	0.00
Long-term interest rate	0.00	0.01	0.02
Current account to GDP ratio	0.00	0.00	0.00

Note: See text. Entries are the p-value of a test of the null hypothesis that the given variable and its lag are irrelevant in determining output given the fiscal treatment. The test is applied to three models. "OLS" refers to the LP responses calculated in Table 2; "IV" refers to the LP responses calculated using the binary instrument in Table 4; and "IV-Total" refers to the LP responses calculated using the continuous instrument.

findings since the tests would rely on a smaller sample. Table 6 reports the p-value associated with the joint null that the candidate variable and its lag are not significant. A rejection means that fluctuations in output could be due to reasons other than the fiscal treatment variable. The basic message from the table is clear: most of the excluded controls are highly significant. For now, a cautious interpretation is to view these findings as a source of concern rather than conclusive evidence that the multipliers reported earlier are incorrect.

Next we check for another condition: Do excluded controls predict fiscal consolidations? Table 7 asks whether variation in the IMF binary treatment variable identified by GLP can be predicted. The results indicate that we have a reasonable basis for this concern. This is a set of estimated treatment equations, where we use a pooled probit estimator to predict the IMF fiscal consolidation variable in year 1, presumptively announced at year 0, based on state variables at time 0. As shown in the appendix, our later results are robust to alternative binary classification models such as pooled logit, and fixed-effects probit and logit with controls for global time-varying trends.

Table 7 shows in column (1) that treatment is more likely, as expected, when public debt to GDP ratios are high: the coefficient is positive, meaning that governments tend to pursue austerity when they have a debt problem. In column (2) we add y^C (the cyclical component of y) and the growth rate of y to further condition on the state of the economy: when the economy is growing below potential, there is an increase in the likelihood of consolidation. Moreover, austerity is more likely to be pursued when output is growing slower in stark contrast to what common sense might suggest. But this finding is in line with contemporary experience in Europe and the U.K., although all of the sample data we use here are pre-crisis. Thus, the act of engaging in pro-cyclical fiscal policy is not a new-fangled craze but more of a chronic tendency in advanced countries. Finally,

Table 7: *Fiscal treatment regression, pooled probit estimators (average marginal effects)*

Probit model of treatment at time t+1 (fiscal consolidation event)

Model	(1)	(2)	(3)	(4)
Public debt/GDP (t)	0.33***	0.28***	0.12*	0.11*
	(0.073)	(0.073)	(0.064)	(0.064)
Cyclical component of log y (t) (y^C)		-0.026**	-0.012	
		(0.011)	(0.009)	
Growth rate of output (t)		-0.030**		-0.024**
		(0.012)		(0.010)
Treatment (t)			0.41***	0.41***
			(0.020)	(0.019)
Observations	457	457	457	457
Classification test: AUC	0.61	0.66	0.81	0.82
	(0.03)	(0.03)	(0.02)	(0.02)

Standard errors in parentheses. ***/**/* indicate $p < 0.01/0.05/0.10$.
y^C is the cyclical component of $\log y$ (log real GDP), from HP filter with $\lambda = 100$.
AUC is the area under CCF curve. $AUC \in [0.5, 1]; H_0 : AUC = 1/2$. See Jordà and Taylor 2011 for explanation of CCF curve. AUC reported here equivalent to area under the ROC curve. See text.

columns (3) and (4) add the lag of the dependent variable Treatment and this has a highly significant coefficient: as we know from the raw data series generated by the IMF study, the fiscal consolidation episodes are typically long, drawn-out affairs, so once such a program is started it tends to run for several years. Being in treatment today is thus a good predictor of being in treatment tomorrow. In these last two columns the lagged growth rate rather than the cyclical level of output emerges as the slightly better predictor of treatment.

Further confirmation of the predictive ability of these treatment regressions is provided by the AUC statistic.[7] The AUC is commonly used in biostatistics and machine learning to evaluate classification ability (see, e.g. Jordà and Taylor 2011). Under the null that the covariates have no classification ability, the AUC = 0.5. Perfect classification ability translates into AUC = 1. The AUC has an approximate Gaussian distribution in large samples. Table 7 measures the classification ability of each specification. The AUC statistics show that the probits have very good predictive ability, with AUC at best around 0.65 when lagged treatment is omitted (Column 2), and rising to around 0.8 with lagged treatment (Columns 3 and 4). The AUCs are all significantly different from 0.5.

The key lesson from Table 7 is simply that the IMF treatment variable has a significant forecastable component. Since the same controls also affect the outcome (see Table 6),

[7] AUC stands for *area under the curve*. The *curve* usually refers to the *Receiver Operating Characteristic* curve or ROC curve. It also refers to the *Correct Classification Frontier*. See Jordà and Taylor (2011).

together these two findings indicate that there could be substantial bias in estimated responses of the type shown so far in this paper, and in the wider literature.[8] The question, then, is how to deal with the problem of potentially endogenously determined instruments. The remainder of this paper provides one answer.

6. Estimators of Average Treatment Effects

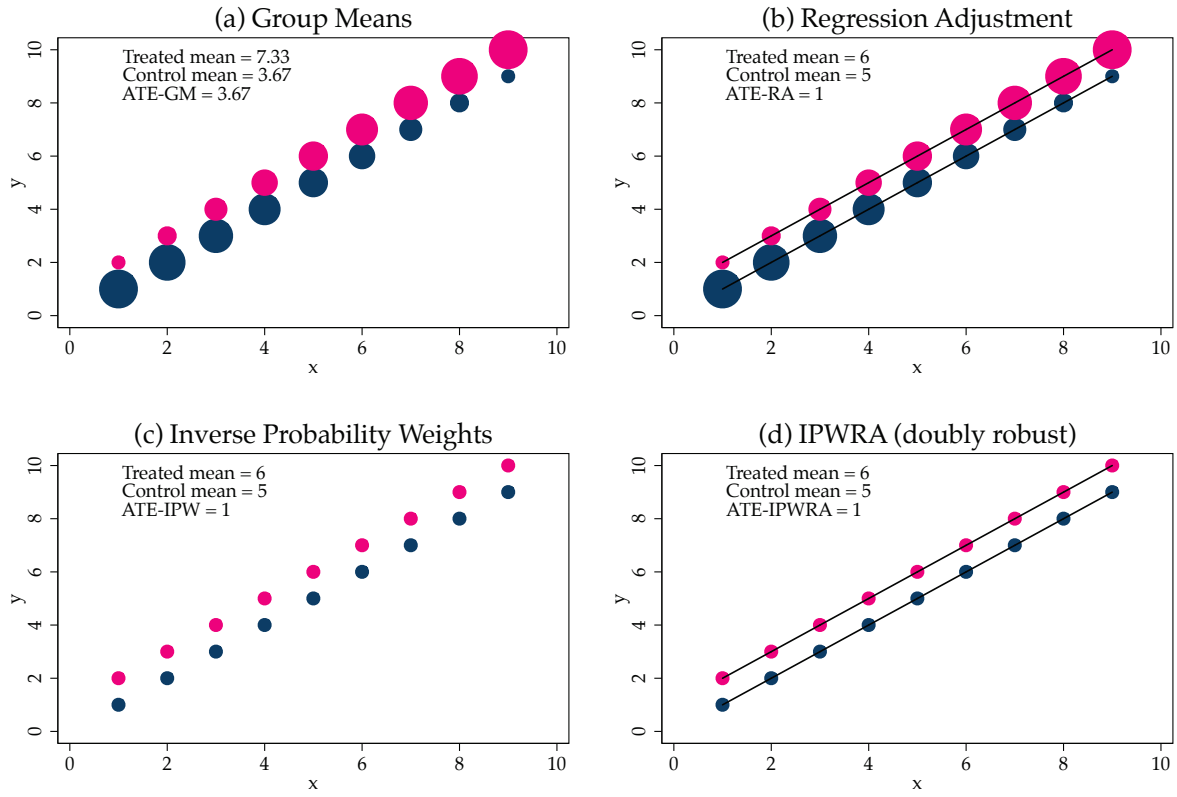
Absent credible instruments, what is the best empirical way forward? In this section we turn our attention to the principles behind medical experimental designs to try to make some headway. Although the technique is relatively new to macroeconomics, matching estimators using inverse propensity score weighting have been frequently applied in cross-sectional data in applied microeconomics. Matching methods more generally constitute a benchmark within the medical research literature when trials are suspected of being contaminated by allocation bias. The provenance of the particular inverse propensity-score weighting method we employ is thus well established.

Figure 1 serves to motivate the methods that we will discuss shortly and exemplifies the perils of allocation bias with a simple bivariate manufactured example based one observable confounder. Panel (a) displays the hypothetical frequencies of observing the control variable x in the treatment and control subpopulations using circles of varying diameter. Think of it as a display of the raw data. In the control subpopulation, we are more likely to observe low values of x. This is indicated with the bigger circles in green that are located near the vertical axis. The opposite is the case for the treated subpopulation, indicated with the orange circles that grow the further they are from the vertical axis. The example is set up so that the true ATE = 1.

The naïve *Group Means* estimator based on expression (2) consists of the difference in means between the two subpopulations delivering an estimate of the ATE = 7.33 (treated mean)- 4.67 (control mean) = 3.67 that is almost four times as large as the true ATE. Panel (b) implements the *regression adjustment* estimator described in (6). Now the ATE is estimated using the conditional mean average implied by the regression estimates of

[8]Hernández De Cos and Moral-Benito (2013) have arrived at a similar conclusion. Their proposed solution to the lack of exogeneity problem is to use an instrumental variable approach. Instruments rely on data for pre-determined controls and on past consolidations. Since data on pre-determined controls already appear in the specification of previous studies (AA, GLP, etc.), the key question is whether past consolidation data predict current consolidation episodes. Fixed-effect panel estimation already takes into account heterogeneity in the unconditional probability of consolidation across countries. Take Australia as an example. It is unlikely that the consolidation observed in 1985 helps determine the likelihood of consolidation in year 1994 beyond the observation that Australia may consolidate more or less often than the typical country (already captured by the fixed effect). There may be little gained from the point of view of strengthening the identification.

Figure 1: *An example of allocation bias and the IPWRA estimator*



Outcome model: y = x + Treated; Treatment model: p(Treated) = x/10; True ATE = 1.

Notes: See text. True ATE = 1. Panel (a) displays the raw hypothetical data using circles of increasing diameter to denote where the data are more frequently observed. Treated units are in orange, control units in green. The *Group Means* estimate of the ATE = 3.67. Panel (b) is the same as panel (a) and adds regression lines to each subpopulation. The *RA* ATE = 1. Panel (c) displays the data in panel (a) once it has been inversely weighted by the frequency with which they are observed. The *IPW* ATE = 1. Panel (d) adds regression lines to panel (c) to show the *IPWRA* estimate of the ATE = 1.

each subpopulation. In this case ATE = 6 (treated regression mean) - 5 (control regression mean) = 1, which is the correct ATE. In the simple example, the effect of the control x is linear so a regression control strategy suffices to obtain the correct ATE.

Suppose that you do not want to make assumptions about the functional form of the regression needed to adjust for the covariate x in the ATE estimator. Panel (c) implements the *IPW* estimator in (13). Using weights based on the inverse frequency with which the data are observed for each value of x in each subpopulation generates a "pseudo-randomized" sample from which the simple difference in mean estimator delivers the correct answer. In this case ATE = 6 (treatment mean using inverse weights) - 5 (control mean using inverse weights) = 1, again the correct value.

In practice one may be unsure about the correct specification of either the regression or the propensity score describing the appropriate reweighing scheme. Panel (d) combines the two approaches (IPW and regression adjustment) based on expression (17). This estimator is "doubly robust" meaning that either the propensity score or the regression may be incorrectly specified and yet still deliver the correct estimate of the ATE. In the example there is no gain from using this procedure, but one can still verify that ATE = 6 (conditional regression mean for treated using inverse weighting) - 5 (conditional regression mean for control using inverse weighting) = 1. Again, the correct ATE.

In what follows we rely again on selection-on-observables arguments to calculate causal effects using IPW-based estimators applied to the LP framework, where the probability of fiscal consolidation is the key policy intervention we want to investigate as a source of allocation bias. The methods build on the intuition contained in Figure 1.

When policy interventions are mostly driven by the endogenous response to controls, we can think of the observable treatment/control subpopulations as being oversampled from the region of the distribution in which the propensity score attains its highest values. Moments calculated with this raw empirical distribution will therefore be biased: not enough probability mass is given to observations with low propensity scores. Weighting by the inverse of the propensity score shifts weight away from the oversampled toward the undersampled region of the distribution. This shift of probability mass reconstructs the appropriate frequency weights of the underlying true distribution of outcomes under treatment and control so that the means estimated from each subpopulation are no longer biased and their difference is an unbiased estimate of the ATE.

Our approach builds on the AJK estimator. The principles behind this estimator are similar to those in the Hirano, Imbens, and Ridder (2003) estimator. That estimator itself relies on Rosenbaum and Rubin (1983) and the earlier Horvitz and Thompson (1952) estimator for stratified samples. In addition, we move further and build on the work by Robins, Rotnitzky, and Zhao (1994); Robins (1999); Lunceford and Davidian (2004); Glynn and Quinn (2010); and Kreif, Grieve, Radice, and Sekhon (2013). These authors discuss IPW in the context of regression estimators that deliver greater robustness and efficiency.

Inverse propensity-score weights (IPW)

We use the same notation as in section 2, referring to outcomes as y_t, the policy variable as D_t, which now is allowed to take only two discrete values $D_t = 1, 0$, and the vector X_t, which collects all information on predetermined outcomes and controls relevant in explaining the policy variable $D_t = D(X_t, \psi, \varepsilon_t)$. We continue to keep the discussion

simple by setting aside notation that refers to the panel dimension of the analysis.

Recall that the critical assumption is the conditional ignorability or selection-on-observables condition (1), repeated here for convenience:

$$(y_{t,h}^{\psi}(d) - y_t) \perp D_t | X_t; \psi \quad \text{for all } h \geq 0, \text{for } d = 1, 0 \text{ and for all } \psi \in \Psi.$$

Rosenbaum and Rubin (1983) show the convenient property that

$$X_t \perp D_t | p(D_t = 1 | X_t, \psi),$$

that is, the propensity score $p(D_t = 1|X_t, \psi)$ is all that is needed to capture the effect of the X_t in the selection-on-observables condition.[9] This result provides further support for the IPW estimator. Recall the average treatment effect (ATE) is, by definition,

$$\Lambda^h = E\left[(y_{t,h}(1) - y_t) - (y_{t,h}(0) - y_t)\right] = E[E\left[(y_{t,h}(1) - y_t) - (y_{t,h}(0) - y_t)|X_t\right]], \quad (9)$$

using the law of iterated expectations. Looking inside the expectations in the final term above, the average policy response conditional on X_t, in terms of observable data, is

$$\begin{aligned} &E\left[(y_{t,h}(1) - y_t) - (y_{t,h}(0) - y_t)|X_t\right] = \\ &E\left[y_{t,h} - y_t | D_t = 1; X_t\right] - E\left[y_{t,h} - y_t | D_t = 0; X_t\right], \qquad \text{for all } h \geq 0, \end{aligned} \quad (10)$$

where it is assumed that the policy environment characterized by $\psi \in \Psi$ remains constant. Estimation of these conditional expectations can be simplified considerably when a model for the policy variable D_t is available.

Angrist and Kuersteiner (2004, 2011) refer to the predicted value from such a policy model the *policy propensity score*. The policy propensity score is meant to ensure the estimation of the policy response (the average treatment effect in the microeconomics parlance) is consistent under the main assumption. In addition, it acts as a dimension-reduction device. Ideally, any predictor of policy should be included, regardless of whether that predictor is a fundamental variable in a macroeconomic model. The probit results reported in Table 7 can be seen as candidate estimates of this policy propensity score. We will instead construct the policy propensity score using a richer specification

[9]Correction of incidental truncation with inverse probability weighting has a long history in statistics (Horvitz and Thompson 1952) and is generally viewed as more general than Heckman's (1976) selection model. Heckman's (1976) approach corrects for incidental truncation using the inverse Mills ratio, requires specific distributional assumptions, and at least one selection variable not affecting the structural equation. Heckman's approach is only known to work for special nonlinear models, such as an exponential regression model (see Wooldridge 1997). See Wooldridge (2002) for a more general discussion.

that includes all the controls used in Table 6 as well.

Denote the policy propensity score $P(D_t = j|X_t) = p_j(X_t, \psi)$ for $j = 1, 0$. Clearly $p_1(X_t, \psi) = 1 - p_0(X_t, \psi)$. Using the selection-on-observables condition in expression (1) shown earlier, then

$$E\left[(y_{t,h} - y_t)\mathbf{1}\{D_t = j\}|X_t\right] = E[(y_{t,h}(j) - y_t)|X_t]p_j(X_t, \psi) \qquad \text{for } j = 1, 0. \tag{11}$$

Solving for $E[(y_{t,h}(j) - y_t)|X_t]$ and taking unconditional expectations, by integrating over X_t, the ATE in (9) can be calculated as

$$\begin{aligned} \Lambda^h &= E\left[(y_{t,h}(1) - y_t) - (y_{t,h}(0) - y_t)\right] \\ &= E\left[(y_{t,h} - y_t)\left(\frac{\mathbf{1}\{D_t = 1\}}{p_1(X_t, \psi)} - \frac{\mathbf{1}\{D_t = 0\}}{p_0(X_t, \psi)}\right)\right] \text{ for all } h \geq 0. \end{aligned} \tag{12}$$

Under standard regularity conditions (detailed in AJK) an estimate of expression (12) can be obtained using sample moments which generalize the sample moments presented earlier in expression (2) for the OLS case.

Suppose that the first-stage treatment model takes the form of a probability of treatment at time t given by the estimated model $\hat{p}_t = p_1(X_t, \hat{\psi})$, where $\hat{\psi}$ is the estimated parameter vector, and $1 - \hat{p}_t = p_0(X_t, \hat{\psi})$. The inverse propensity score weighted (IPW) "ratio estimator" of the average treatment effect is

$$\hat{\Lambda}_{IPW} = \frac{1}{n}\sum_t \left[\frac{D_t(y_{t+h} - y_t)}{\hat{p}_t}\right] - \frac{1}{n}\sum_t \left[\frac{(1 - D_t)(y_{t+h} - y_t)}{1 - \hat{p}_t}\right]. \tag{13}$$

Some improvements can be made to this expression. Imbens (2004) and Lunceford and Davidian (2004) suggest renormalizing the weights so that they sum up to one in small samples. Hence expression (13) becomes

$$\hat{\Lambda}_{IPW} = \frac{1}{n_1^*}\sum_t \left[\frac{D_t(y_{t+h} - y_t)}{\hat{p}_t}\right] - \frac{1}{n_0^*}\sum_t \left[\frac{(1 - D_t)(y_{t+h} - y_t)}{1 - \hat{p}_t}\right], \tag{14}$$

where

$$n_1^* = \left(\sum_t \frac{D_t}{\hat{p}_t}\right) \qquad n_0^* = \left(\sum_t \frac{(1 - D_t)}{(1 - \hat{p}_t)}\right), \tag{15}$$

and the notation n_j^* parallels the notation n_j for $j = 1, 0$ in (2). Note that $E[D_t/p_t] = E[E(D_t|X_t)]/p_t = 1$; similarly $E[(1 - D_t)/(1 - p_t)] = E[E((1 - D_t)|X_t)]/(1 - p_t) = 1$; and hence it follows that in large samples expressions (13) and (14) apply the same weighting, since $E(n_1^*) = E(n_0^*) = n$. These expressions are natural analogs of the Group

Mean estimator in (2), with inverse propensity-score weighting to correct for allocation bias and to achieve a quasi-random distribution of treatment and control observations via reweighting.

Regression adjustment (IPWRA) and Augmented IPW (AIPW)

As a way to enhance robustness, researchers have derived estimators with a regression adjustment component added to the standard IPW estimator presented above. This estimator parallels that in expression (6) but using inverse probability weighting. To further enhance efficiency, the augmented IPW or AIPW estimator combines the IPW and IPWRA estimators in a manner to be discussed shortly.

It is natural to consider extending the estimator in expression (6) using the propensity score. Formally, the basis for such an estimator would be to transition from expression (11) to (12) in the following manner

$$\Lambda^h = E\left[(y_{t+h} - y_t|X_t)\left(\frac{\mathbf{1}\{D_t = 1\}}{p^1(X_t,\psi)} - \frac{\mathbf{1}\{D_t = 0\}}{p^0(X_t,\psi)}\right)\right] \text{ for all } h \geq 0, \tag{16}$$

which can be implemented by first projecting the outcome variable on the set of control variables (see, e.g. Robins and Rotnitzky 1995; Robins, Rotnitzky, and Zhao 1995; and Wooldridge 2007). The inverse propensity-score weighted estimator with regression adjustment (IPWRA) is then given by

$$\hat{\Lambda}^h_{IPWRA} = \frac{1}{n_1^*}\sum\left[\frac{D_t m_1^h(X_t,\hat{\theta_1}^h)}{\hat{p}_t}\right] - \frac{1}{n_0^*}\sum\left[\frac{(1-D_t)m_0^h(X_t,\hat{\theta_0}^h)}{1-\hat{p}_t}\right], \tag{17}$$

where again $m_j^h(X_t,\hat{\theta_j}^h)$ for $j = 1,0$ is the conditional mean from the first-step regression of $(y_{t+h} - y_t)$ on X_t as in expression (6) in Section 2. The n_j^* for $j = 1,0$ are the same as in expression (15). It is clear that equation (17) nests all the previous estimators, the Group Mean (2), the RA (6), and the IPW (14) as special cases.

The estimator in expression (17) falls into the class of *doubly robust* estimators (see, e.g., Imbens 2004; Wooldridge 2007; Lunceford and Davidian 2004; and Kreif et al. 2011). The intuition behind the estimator is to use the regression model as a way to "predict" the unobserved potential outcomes. Consistency of the estimated ATE only requires either the conditional mean model or the propensity score model to be correctly specified.

However, although (17) is one of a large class of unbiased IPWRA estimators of ATE, it is not the most efficient in this class. Starting with Robins, Rotnitzky, and Zhao (1994) and more recently, Lunceford and Davidian (2004), the estimator within the doubly-robust

class having the smallest asymptotic variance, is the (locally) semi-parametric efficient estimator

$$\hat{\Lambda}^h_{AIPW} = \frac{1}{n}\sum_t \left\{ \left[\frac{D_t(y_{t+h} - y_t)}{\hat{p}_t} - \frac{(1-D_t)(y_{t+h}-y_t)}{(1-\hat{p}_t)} \right] - \frac{(D_t - \hat{p}_t)}{\hat{p}_t(1-\hat{p}_t)} \left[(1-\hat{p}_t) m^h_1(X_t, \hat{\theta}^h_1) + \hat{p}_t m^h_0(X_t, \hat{\theta}^h_0) \right] \right\} \tag{18}$$

Thus, the estimator in (18) can be seen as the basic IPW estimator plus an adjustment consisting of the weighted average of the two regression estimators. The adjustment term has expectation zero when the estimated propensity scores and regression models are replaced by their population counterparts. Moreover, the adjustment term stabilizes the estimator when the propensity scores get close to zero or one (Glynn and Quinn 2010), and this alleviates with the need to truncate the propensity score weights as suggested in Imbens (2004). Another way to interpret the AIPW estimator is to realize that

$$\hat{\Lambda}^h_{AIPW} = \hat{\Lambda}^h_{IPW} + (\hat{\Lambda}^h_{RA} - \hat{\Lambda}^h_{IPWRA}). \tag{19}$$

Readers familiar with the bootstrap will notice the similarities between the bootstrap bias correction formula and expression (19).

The AIPW has a number of attractive theoretical properties. Using the theory of M-estimation, Lunceford and Davidian (2004) show that the estimator is asymptotically normally distributed. In addition, they show that the variance can be calculated using the empirical sandwich estimator $V(\hat{\Lambda}^h_{AIPW}) = \frac{1}{n^2}\sum_t (\hat{I}^h_t)^2$, where

$$\hat{I}^h_t = \left\{ \left[\frac{D_t(y_{t+h} - y_t)}{\hat{p}_t} - \frac{(1-D_t)(y_{t+h}-y_t)}{(1-\hat{p}_t)} \right] - \frac{(D_t - \hat{p}_t)}{\hat{p}_t(1-\hat{p}_t)} \left[(1-\hat{p}_t) m^h_1(X_t, , \hat{\theta}^h_1) + \hat{p}_t m^h_0(X_t, \hat{\theta}^h_0) \right] \right\} - \hat{\Lambda}^h_{AIPW}.$$

Later we allow for the possibility that the $\hat{I}^h_t$ are not a martingale difference sequence and calculate standard errors using cluster robust methods. When the propensity score and the regression function are modeled correctly, the AIPW achieves the semi-parametric efficiency bound. Alternatively, Imbens (2004) shows that standard errors for $\hat{\Lambda}^h_{AIPW}$ can be calculated with the bootstrap.

What We Do

The next section reports the results of applying the AIPW estimator (18) to measure the average treatment effect of fiscal consolidations as a counterpoint to the conventional OLS and IV results reported earlier. As a way to understand where the differences come from, we first implement the AIPW estimator by restricting the parameters of the regression (based on LPs) to be the same in the treated and control subpopulations, as is implicit in the OLS and IV approaches. Under that constraint, the results from the AIPW estimator are close to the IV results seen earlier. Next we allow for the parameters to vary across subpopulations, adhering to the way expression (18) is typically applied in the policy evaluation literature. These results deliver the same qualitative implication of contractionary austerity, but show that the effects of consolidations are quantitatively even more painful.

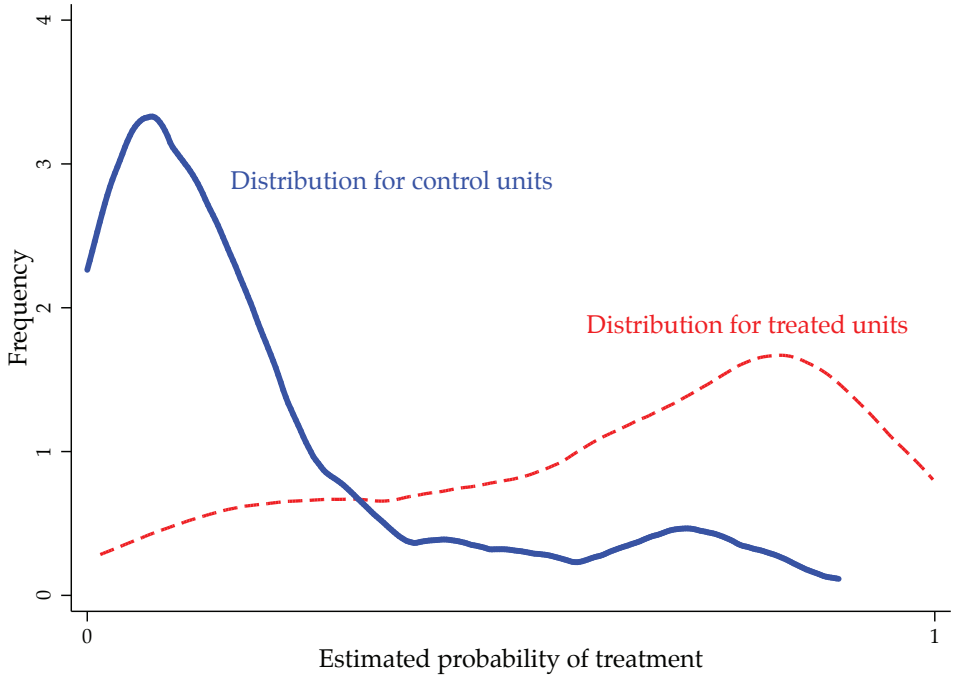
7. Contractionary Austerity Revisited: Estimates of the Average Effect of Fiscal Consolidations

This section presents AIPW estimates of the ATE of fiscal consolidations. Following standard procedures, the propensity score used here is based on a saturated probit model that extends the set of controls used in Table 7 with the current and lagged values of the controls in Table 6. The saturated probit also includes country-fixed effects. Although we do not report the coefficient estimates of this more saturated model, it is worth mentioning that the AUC is now 0.86.

Figure 2 provides smooth kernel density estimates of the distribution of the propensity score for the treated and control units to check for *overlap*. One way to think of overlap is to consider what overlap would be in the ideal RCT. The empirical distributions of the propensity score for treated and control units would be uniform and identical to each other. At the other extreme, suppose that treatment is allocated mechanically on the basis of controls. Then the distribution of treated units would spike at one and be zero elsewhere, and the distribution of control units would spike at zero and be zero elsewhere. Despite the high AUC, the figure indicates considerable overlap between the distributions, which indicates we have a satisfactory first-stage model with which to properly identify the ATE using IPW methods.

However, the figure also indicates that there are some observations likely to get very high weights. Specifically, there are control (treated) units whose propensity score is near zero (one) and hence who get weights in the IPW in excess of 10. In general, it is often

Figure 2: *Overlap check: empirical distributions of the treatment propensity score*



Notes: See text. The propensity score is estimated using the saturated probit specification discussed in the text, which includes country fixed effects. The figure displays the predicted probabilities of treatment with a dashed line for the treatment observations and with a solid line for the control observations.

recommended to truncate the maximum weights in the IPW to 10 (see e.g. Cole and Hernán, 2008 and Imbens 2004). However, the AIPW has the property that high weights in the IPW are compensated at the same rate by the augmentation term. Experiments not reported here indicate that this is indeed what happens in practice and that truncation is unnecessary in our application (see, e.g., Appendix A.3).

Using the more saturated probit, we then estimate cumulated responses and their sum to the 5-year horizon as before. Our indicator of a fiscal consolidation is the narrative IMF indicator, the Treatment variable. Since Treatment is binary, we are estimating average effects only. However, coincidentally, the average treatment size (or dose) is close to 1 percent of GDP in these data (the exact value is 0.97, with a standard deviation of 0.07 in the full sample), so the interpretation of these responses is directly comparable to a conventional multiplier, with only a small upscaling (of 1/0.97) for strict accuracy. We can return to this rescaling issue in a moment when we make a formal comparison with the previous OLS and IV results.

We begin by discussing Table 8, which is the direct counterpart to the OLS and IV result presentations in Tables 1 and 3. Here we show the ATE of fiscal consolidation

Table 8: *Average treatment effect of fiscal consolidation, AIPW estimates, full sample*

Deviations of log real GDP (relative to Year 0, $\times 100$)

	(1) Year 1	(2) Year 2	(3) Year 3	(4) Year 4	(5) Year 5	(6) Sum
Fiscal ATE, restricted ($\theta_1^h = \theta_0^h$)	-0.17	-0.55**	-0.61***	-0.88**	-1.14**	-3.22***
	(0.17)	(0.23)	(0.20)	(0.32)	(0.42)	(0.89)
Fiscal ATE, unrestricted ($\theta_1^h \neq \theta_0^h$)	-0.24	-0.70**	-0.75***	-0.93**	-1.23**	-3.61***
	(0.16)	(0.26)	(0.25)	(0.33)	(0.47)	(1.06)
Observations	456	439	423	406	389	389

Standard errors (clustered by country) in parentheses. ***/**/* indicate $p < 0.01/0.05/0.10$.
Conditional mean controls: cyclical component of y, 2 lags of change in y, country fixed effects.
y^C is the cyclical component of $\log y$ (log real GDP), from HP filter with $\lambda = 100$.
Specification includes country fixed effects in the propensity score model and in the AIPW model.
Propensity score based on the saturated probit model as described in the text. AIPW estimates do not impose restrictions on the weights of the propensity score. Truncated results not reported here but available upon request. See text.

using the AIPW estimator (18), for the full sample (i.e., no use of boom and slump bins, yet) and using the propensity score estimates based on the saturated probit. Both the treatment-equation probit model and the outcome-equation AIPW model include country-fixed effects.

Table 8 is organized into two rows. The first row reports the results based on imposing the restriction $\theta_1^h = \theta_0^h$, the usual implicit restriction used without hesitation in the macro-VAR empirical literature and the same restriction we imposed in reporting the results of Tables 1 and 3. The second row reports the results that do not impose the $\theta_1^h = \theta_0^h$ restriction. The results are qualitatively similar to those reported in Table 3 in that we still find that austerity is contractionary. However, the estimated impacts of fiscal consolidations on output are now even bigger.

Recall that according to the IV estimates, the sum effect was -2.94*** over 5 years. This would imply an average annual real GDP loss of about 0.59% of GDP per 1% of fiscal consolidation over each of the 5 years. Here our AIPW estimate with unrestricted coefficients has a sum effect of -3.61*** over 5 years. This would imply an average annual real GDP loss of about 0.74% of GDP per 1% of fiscal consolidation over each of the 5 years (using a 1/0.97 rescaling factor). Thus the implied output losses due to austerity are about 20% larger under our AIPW estimation than with IV estimation.

Next we once again explore the same partition of the data into booms and slumps, allocating to the bins according to whether output is above or below trend as in earlier sections to provide a more granular view of these results, and Table 9 presents these AIPW

Table 9: *Average treatment effect of fiscal consolidation, AIPW estimates, booms versus slumps*

Deviations of log real GDP (relative to Year 0, $\times 100$)

	(1) Year 1	(2) Year 2	(3) Year 3	(4) Year 4	(5) Year 5	(6) Sum
Fiscal ATE, $y^C > 0$, boom	-0.31	-0.42	-0.53	-0.44	-0.58	-1.25
	(0.22)	(0.35)	(0.42)	(0.60)	(0.84)	(1.97)
Fiscal ATE, $y^C < 0$, slump	-0.19	-0.76***	-0.96***	-0.49	-1.07	-3.83**
	(0.19)	(0.25)	(0.33)	(0.46)	(0.63)	(1.54)
Observations	456	439	423	406	389	389

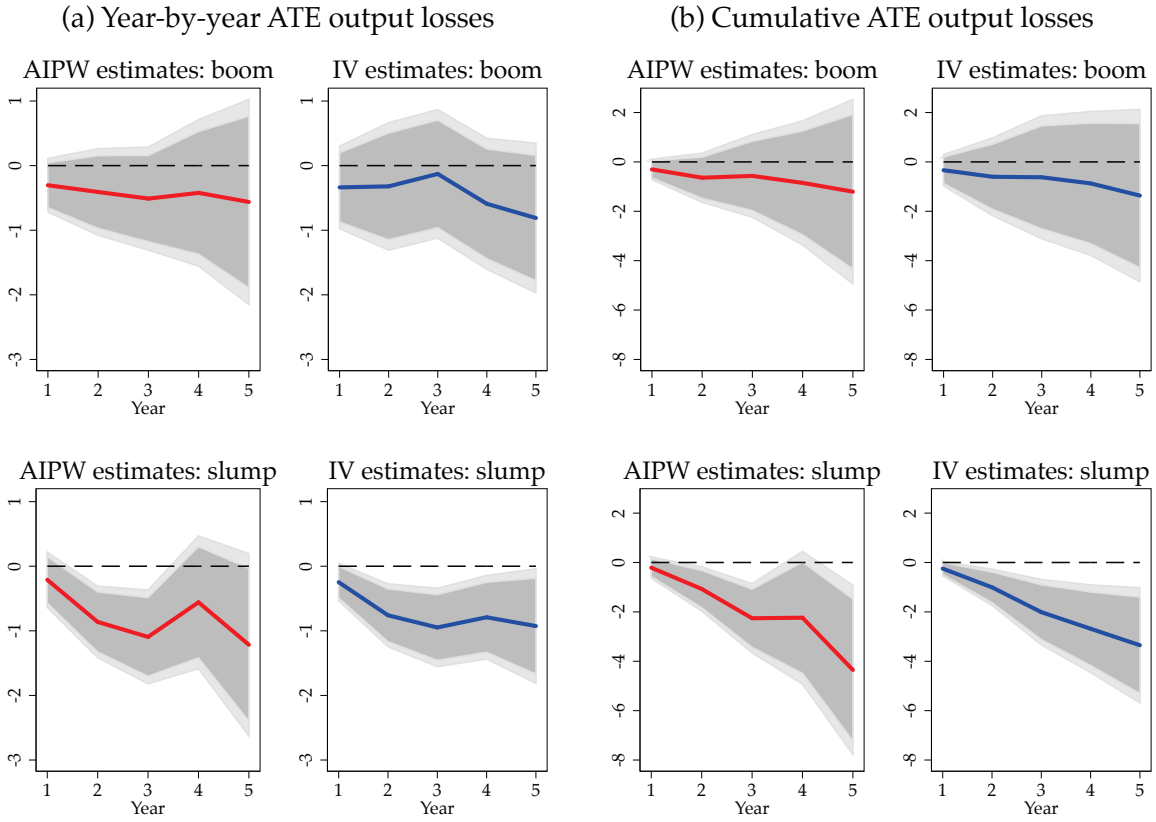
Standard errors (clustered by country) in parentheses. ***/**/* indicate $p < 0.01/0.05/0.10$.
Conditional mean controls: cyclical component of y, 2 lags of change in y, country fixed effects.
y^C is the cyclical component of $\log y$ (log real GDP), from HP filter with $\lambda = 100$.
Specification includes country fixed effects in the propensity score model and in the AIPW model.
Propensity score based on the saturated probit model as described in the text. AIPW estimates do not impose restrictions on the weights of the propensity score. The boom bin is for observations where the cyclical component y^C is greater than zero, the slump bin is for observations where the cyclical component is less than or equal to zero.

estimates based on the same saturated policy propensity score probit model described earlier. These results show that in a boom a fiscal consolidation has on average a small, negative, but imprecisely estimated effect. The first row of the table indicates that the average accumulated loss after five years is -1.25 percent of GDP. In a slump, the results are about three times as strong and highly statistically significant: after five years, the accumulated average loss is -3.83** percent of GDP, as shown in the second row of the table. Scaling these effects for the average treatment size in each bin (0.89 in slumps, 1.03 in booms) the average loss per 1% fiscal consolidation is 0.24% of GDP per year over the five year window in booms, and 0.86% of GDP per year in slumps.

Summing up our LP results, we always find more adverse paths when austerity is imposed in booms rather than in slumps, but there are big differences. OLS suggests that austerity might have a small and imprecisely estimated expansionary effect, although a more granular view indicates that even then, this result holds only in booms. Using the "narrative" instrument we would walk away believing more firmly that austerity is contractionary. The estimated effect with IV is relatively small and imprecisely estimated for the boom, but stronger and significant in the slump, adding up to -3.3% over 5 years. Finally, using the AIPW estimator we find even larger contractionary effects of austerity, about 20% larger, still not statistically significant in booms, and amounting to -3.8% over 5 years in slumps.

Figure 3 displays the coefficients reported in Table 9, with appropriate rescaling in the case of AIPW to allow for the average treatment size in each bin, boom and slump, to

Figure 3: *Comparing AIPW and IV estimates of the response of the output path to a fiscal consolidation, deviations of log real GDP (relative to Year 0, ×100)*



Notes: Panel (a) reports the cumulative ATE responses based on $y_{t+h} - y_t$, where as panel (b) presents the accumulated ATE output loss, which is the running sum of the coefficients displayed in panel (a). 95/90% error bands displayed. The top row shows the results for the subpopulation of observations in the boom measured in deviations above HP trend. The bottom row shows the results for the subpopulation of observations in the slump, measured in deviations below HP trend. AIPW refers to the responses calculated using the AIPW estimator of Section 6; IV refers to the IV estimator discussed in Section 2. AIPW impacts are rescaled to allow for the average size of fiscal consolidation in each bin. See text.

show the dynamic ATE impacts of fiscal consolidations in graphical form and compares them with the responses obtained using the IV coefficient estimates which were reported earlier in Table 4.

Our results underscore that austerity tends to be painful, but that timing matters: the least painful fiscal consolidations, from a growth and hence budgetary perspective, will tend to be those launched from a position of strength, that is, in the boom not the slump. This would seem to require moderately wise policymaking and/or fiscal regimes (councils, rules, etc.), not to mention an ability to stay below any debt limit so as to maintain capital market access to permit smoothing.

The next section puts our new results to work in the context of the austerity program launched in U.K. by the Coalition administration, to show how our analysis can be used in practice. Moreover, by putting our results in a realistic situation outside the sample used for estimation, we obtain a feel for how well calibrated our findings are to the recent macroeconomic experience of a representative economy from our sample.

8. Counterfactual: Coalition Austerity and the U.K. Recession

This section makes a counterfactual forecast of the post-2007 path of the U.K. economy with and without the fiscal austerity policies imposed by the Coalition government after the 2010 election. These estimates are based on a sample that excludes the global financial crisis. Therefore, the exercise has the flavor of an out-of-sample evaluation.

The U.K. experienced a much weaker recovery than in the U.S., where nothing close to a double dip took place. The divergence between the two recovery paths began in 2010 (Schularick and Taylor 2012). Since both countries' central banks acted with aggressive ease, by going to the zero bound and pursuing quantitative easing policies thereafter, explanations for the differences have focused elsewhere. Various explanations have been offered, ranging from tighter U.K. fiscal policy, to spillovers from the Eurozone and weak trade links with fast-growth emerging markets. Other stories have invoked contractions in oversized U.K. sectors such as finance and North Sea oil and gas, the extent of non-bank finance, and differential energy costs (Posen 2012; Davies 2012).

To gain quantitative traction on the share of responsibility that should be borne by fiscal policy we use our AIPW estimates. We scale, and assign the impacts of fiscal shocks as follows. As a measure of the change in fiscal stance we use the change in the U.K. Office of Budget Responsibility's (OBR) cyclically-adjusted primary balance. The changes turn out to be +2.3% of GDP in year 1 (2009–10 to 2010–11), followed by +1.5% in year 2, and +0.1% in year 3, showing a slowing of the pace of tightening in year 3, but with further austerity planned in future years.[10] This gives us a sequence of three fiscal policy shocks. Note that the average treatment in the low bin is 0.89, so for this counterfactual

[10]Considerable controversy attends the question as to whether austerity policy was eased in year 3, with the Chancellor and HM Treasury insisting that consolidation continued, but many critics suggesting the data showed otherwise. This is often referred to as the Plan A versus Plan B debate. See the discussion by Jonathan Portes, `http://www.niesr.ac.uk/blog/fiscal-policy-plan-and-recovery-explaining-economics`. For consistency with official sources we use the official OBR figures, excluding certain accounting credits in year 3 due to Bank of England and Royal Mail transactions which are not related to fiscal plans. See Appendix A.5 and Appendix Table A5. Two alternative measures of fiscal shocks are discussed in the appendix, one from the OBR and one from the IMF. The measure we have chosen is the official UK measure and is more modest than these alternative measures.

exercise we scale treatment effects due to each shock by a factor of 1/0.89.

We then have to compute the impact of each shock at each horizon and make sure we assign it appropriately. Our AIPW estimation already allows for the fact that if at time 0 a treatment occurs, then its measured impact at time $h \geq 1$ includes not just the direct impact of the policy on output, but also its indirect impact arising from the fact that treatment at time 0 also predicts some positive probability of treatment at time $h \geq 1$. To prevent double counting we therefore need to carefully subtract these "expected austerity" measures from any forecast of fiscal impacts in year 1 and beyond.

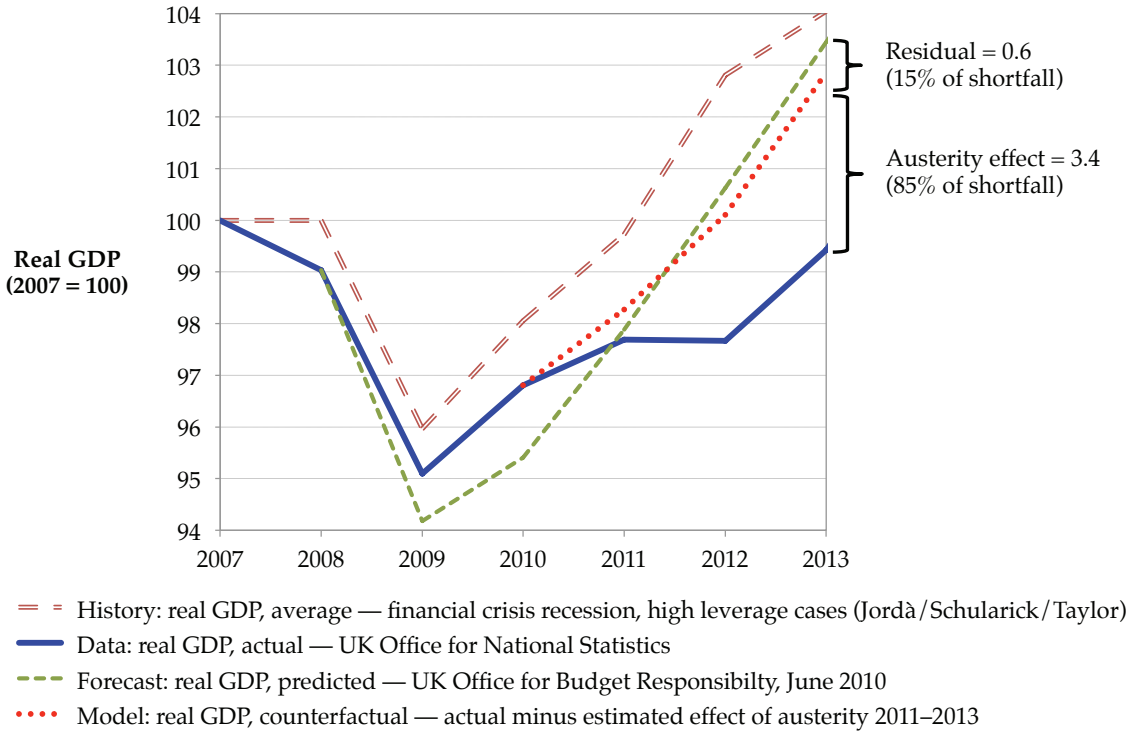
The effects of the first round of austerity in 2010–11 can be computed directly from the AIPW estimates above (for the slump bin, since the U.K. was already in a deep recession then). For example, the effect of the 2010–11 austerity shock in 2011 itself would be computed as the shock magnitude of +2.3 (OBR data, as above) multiplied by the scaling factor of 1/0.89 (noted above), and then multiplied by the AIPW coefficient of -0.19 (from the slump bin in year 1).

However, in other subsequent years an adjustment must be done. For example, the effect of the 2010–11 plus 2011–12 austerity shock in 2012 itself would be computed in two parts. First, there is a similar direct effect of the first year shock on second year output: the first year shock magnitude of 2.3 (again) multiplied by the scaling factor of 1/0.89 (again), and then multiplied by the AIPW coefficient of -0.76 (from the slump bin, but now in year 2). Second, there is the additional effect from *unexpected* treatment in year 2 conditional on treatment in year 1. To get at this problem we estimate a simple LP regression for the forward path of treatment at time h, conditional on treatment today, and use these to weight austerity impacts in Years 2 and 3.[11]

The results of this counterfactual exercise are presented in Figure 4, and for reference we also show various actual and forecast paths for U.K. real GDP from 2007 (the business cycle peak) through 2013. As a starting point, absent knowledge of what was to happen after the 2010 Coalition austerity program, what might have been the ex ante expected

[11]The LP estimations for the forward path of treatment for the necessary 3-year horizon are reported in the appendix in Table A4. We find that the ATE estimate of a change in probability, in the slump bin, of a treatment in year 1 given a treatment in year 0 is 0.51; the model also gives a 28% chance in year 2 and 17% in year 3. For our counterfactual this means that 51% of the Coalition austerity in 2011–12 (and 28% in 2012–13) was "baked in"—in probabilistic terms—by the decision to do austerity in 2010–11. So this component is already accounted for in the AIPW output path estimates. The net effects can be computed mechanically as we illustrate in the following example. First, we can compute the first year shock magnitude of 2.3 multiplied by the scaling factor of 1/0.89, and then multiplied by the AIPW coefficient of -0.76 (from the slump bin in year 2). Second, we can add to this the second year shock magnitude of 1.5 (OBR) multiplied by the scaling factor of 1/0.89, multiplied by the AIPW coefficient of -0.24 (from the slump bin in year 1), and multiplied by the probability of no treatment in year 2 which is 0.49=1-0.51. In a similar way we can assign unexpected and expected effects of contemporaneous treatment to prior treatment in all years along the path.

Figure 4: *U.K. austerity: forecast, actual, and counterfactual paths for real GDP, 2007–13*



Notes: Units are percent of 2007 real GDP, the last peak. OBR forecast is from `http://budgetresponsibility.independent.gov.uk/wordpress/docs/pre_budget_forecast_140610.pdf`. The Jordà, Schularick, and Taylor (2013) path is for real GDP per capita, extended to a 6-year horizon, adjusted by +0.65% per year given the U.K. rate of population growth. Actual data from ONS in March 2014. Model counterfactuals subtract estimated AIPW responses in the slump bin, suitably scaled. See text.

path of the U.K. economy? This question is answered by the two dashed lines. The double-long-dashed line shows the unconditional historical path in a financial crisis recession based on a large sample of all advanced-economy recessions from 1870 to 2007 in Jordà, Schularick, and Taylor (2013), extended to the 6-year horizon. We restrict attention to their average path for highly-leveraged economies after a financial crisis, a category which includes the U.K. case in 2007. Clearly, a seriously painful recession was to be expected anyway: if output is scaled to 100 in 2007, this path shows a 4% drop over two years, to a level of 96 by 2009, followed by recovery thereafter to about 104 in 2013.

What did the authorities expect? According to the June 2010 Pre-Budget report of the OBR they expected something similar but slightly worsened to unfold after 2010, as shown by the short-dashed path in the figure. The bottom in output here is 94.2 and the recovery was predicted to be initially slower, although by 2012 the OBR thought the output level would be 100.6 and by 2013 it would be at 103.4, in the same units. (Thus the difference between the two displayed forecast paths is only 0.6% by 2013.)

Alas, this did not come to pass, as shown by the solid line in the chart using actual UK (ONS) data to depict the outturn of events. Everything was going more or less in line with the forecast path until 2010. After that, a double-dip recession was avoided only by a decimal rounding and the U.K. real economy virtually flatlined for a couple of years before a small uptick in 2013. (In per capita terms, the UK economy actually shrank.)

How much of the U.K.'s dismal performance can be attributed to the fiscal policy choice of instigating austerity during a slump? The answer based on our counterfactual model is about ⅝. This is shown by the dotted line in the chart, which cumulates the effects of each of the three years of austerity on growth from 2010 to 2013. By 2013, the last year in the window, the cumulative effects of these choices amounted to about 3.4% of GDP (in 2007 units) where the total gap relative to the actual path was 4.0%, thus leaving an unexplained residual of 0.6%. Our model also suggests that additional drag from the 2010–13 austerity policies will also continue to be felt into 2014–16, even if there is no further austerity imposed.[12]

In 2013, at the end of the period analyzed here, OBR published an estimate that austerity caused a roughly -1.5% change in output in the year 2013. Our -3.4% estimate of the impact of fiscal austerity on economic activity is more than twice as large. We think this difference is largely due to the fact that, unlike us, OBR does not allow for state-dependence and OBR forces the effects to decay to zero after four years. Both of these modeling choices would appear to be strongly rejected by the data, however.[13]

Even so, our 3.4% estimate could *still* be biased down because we are unable to adjust for monetary policy at the zero lower bound (ZLB). The U.K. out-of-sample counterfactual took place in a liquidity trap environment, but the in-sample data we used for estimation overwhelmingly do not. Our estimates are based on a sample from the 1970s to 2007. Out of 173 consolidation episodes, there are only 7 country-year observations at the ZLB, all relating to Japan in the 1990–2007. Economic theory (Christiano et al. 2011; Eggertsson and Krugman 2012; Rendahl 2012) and also historical evidence from the 1930s (Almunia et al. 2010) indicate that fiscal multipliers are much larger under ZLB conditions than in normal times when monetary policy is away from this constraint. But we cannot hope to convincingly capture the ZLB effect in our sample with just a handful of observations from Japan, so this must remain a goal for future research where we hope to apply our new estimation methods to a large set of contemporary and historical data.

[12]The residual in Figure 4 could be accounted for by factors outside the framework: export patterns, the Eurozone crisis, or idiosyncratic U.K. sector shocks. There may have also been overoptimism in the 2010 forecast (e.g., OBR underestimating either the size or economic impacts of upcoming austerity shocks).

[13]See the impacts for 2010–11, 2011–12, and 2012–13 cumulated to 2012–13 in Chart 2.26 of the OBR's Forecast Evaluation Report, `http://cdn.budgetresponsibility.independent.gov.uk/FER2013.pdf`.

9. Conclusion

Few macroeconomic policy debates generate as much controversy as the current austerity argument, and as Europe stagnates the furore appears to be far from over. Amidst the cacophony of competing estimates of fiscal multipliers, the goal of this paper is not to add another source of noise.

Rather, the main contribution is to harmonize dissonant views into a unified framework where the merits of each approach can be properly evaluated. The effect of fiscal consolidation on macroeconomic outcomes is ultimately an empirical question. In the absence of randomized controlled trials, we have to rely on observational data. And to measure the causal effect of fiscal consolidations on growth, it is critical that identification assumptions be properly evaluated and that empirical methods be suitably adjusted to the demands of the data.

Whenever outcomes are correlated with observables that determine the likelihood of treatment, the effect of the treatment cannot be causally measured without bias. Yet, this allocation bias prevents us from being able to tell whether or not the low or even inverted values of the fiscal multiplier often found in this strand of the literature are indeed close enough to the truth.

If episodes of fiscal consolidation could be separated by whether or not they are explained by circumstances, identification could be, once again, restored. The narrative approach relies on a careful reading of the records to achieve just such a separation. Moreover, results from this approach indicate that the fiscal multiplier is larger in magnitude, especially in depressed economies. However, it is critical that those consolidations believed to be exogenous not be predictable by observable controls. The data indicate this not to be the case and it may appear that we are no better off than before.

Extant results in the literature can be somewhat reconciled by interpreting exogenous consolidations as instrumental variables. After all, if the narrative approach were not very informative about the exogeneity of these episodes, there should not be any difference in the value of the multiplier estimated using simple least squares and IV methods. But this turns out not to be the case. So, while imperfect, the narrative approach (through these IV estimates) seems to be isolating fiscal consolidations that differ from those in the overall population in some important respects. Whether the fiscal multiplier estimated with instrumental variables can be interpreted causally required further analysis.

Dissatisfaction with the violation of exogeneity conditions required for identification could lead one, like Mill (1836), to the nihilistic conclusion that without an *experimentum crucis* mere observational data are hopelessly unsuitable for testing a macroeconomic

hypothesis, but we believe the battle is not lost. Propensity score methods, common in biostatistics, medical research, and in applied microeconomics when ideal randomized trials are unavailable, offer a last line of defense. Recent work by Angrist, Jordà and Kuersteiner (2013) introduced inverse probability weighted estimators of average treatment effects for time series data.

Our appeal to this approach begins by recognizing that fiscal consolidations are not exogenous events, even those identified by the narrative approach. Next we construct a predictive model for the likelihood of fiscal consolidation using various specifications including some with a rich set of available observable controls. The predictive model serves to reallocate probability mass from the regions of the distributions in the treatment/control subpopulations that are oversampled to those regions that are undersampled, enabling identification in the framework of the Rubin Causal Model.

Our estimates are quantitatively close to those from the instrumental variables specification than to those from the least squares specification, although that such would be the outcome was unknowable without doing the analysis. Our analysis suggests even larger impacts than the IMF study when the economy is growing below its long-run trend, however. Generally, in the slump, austerity prolongs the pain, much more so than in the boom. It appears that Keynes was right after all.

References

Alesina, Alberto, and Roberti Perotti. 1995. Fiscal Expansions and Adjustments in OECD Economies. *Economic Policy* 10(21): 207–47.

Alesina, Alberto, and Silvia Ardagna. 2010. Large Changes in Fiscal Policy: Taxes versus Spending. In *Tax Policy and the Economy*, vol. 24, edited by Jeffrey R. Brown. Chicago: University of Chicago Press, pp. 35–68.

Almunia, Miguel, Agustn Bénétrix, Barry Eichengreen, Kevin H. O'Rourke, and Gisela Rua. 2010. From Great Depression to Great Credit Crisis: Similarities, Differences and Lessons. *Economic Policy* 25(62): 219–65.

Angrist, Joshua D., Òscar Jordà, and Guido M. Kuersteiner. 2013. Semiparametric Estimates of Monetary Policy Effects: String Theory Revisited. NBER Working Paper 19355.

Angrist, Joshua D., and Guido M. Kuersteiner. 2004. Semiparametric Causality Tests Using the Policy Propensity Score. NBER Working Paper 10975.

Angrist, Joshua D., and Guido M. Kuersteiner. 2011. Causal Effects of Monetary Shocks: Semiparametric Conditional Independence Tests with a Multinomial Propensity Score. *Review of Economics and Statistics* 93(3): 725–47.

Angrist, Joshua D., and Jörn-Steffen Pischke. 2010. The Credibility Revolution in Empirical Economics: How Better Research Design Is Taking the Con out of Econometrics. *Journal of Economic Perspectives* 24(2): 3–30.

Ardagna, Silvia. 2004. Fiscal Stabilizations: When Do They Work and Why. *European Economic Review* 48(5): 1047–74.

Auerbach, Alan J., and Yuriy Gorodnichenko. 2012. Measuring the Output Responses to Fiscal Policy. *American Economic Journal: Economic Policy* 4(2): 1–27.

Auerbach, Alan J., and Yuriy Gorodnichenko. 2013. Fiscal Multipliers in Recession and Expansion. In *Fiscal Policy after the Financial Crisis* edited by Alberto Alesina and Francesco Giavazzi. Chicago: University of Chicago Press, pp. 98–102.

Barro, Robert J., and Charles J. Redlick. 2011. Macroeconomic Effects from Government Macroeconomic Effects from Government Purchases and Taxes. *Quarterly Journal of Economics* 126(1): 51–102.

Blanchard, Olivier J. 1993. Suggestion for a New Set of Fiscal Indicators. OECD Economics Department Working Papers 79.

Chalmers, Iain. 2005. Statistical Theory was not the Reason that Randomisation was Used in the British Medical Research Council's Clinical Trial of Streptomycin for Pulmonary Tuberculosis. In *Body Counts: Medical Quantification in Historical and Sociological Perspectives* edited by G. Jorland, A. Opinel, and G. Weisz. Montreal: McGill-Queens University Press, pp. 309–34.

Chalmers, Iain. 2011. Why the 1948 MRC trial of Streptomycin Used Treatment Allocation Based on Random Numbers. *Journal of the Royal Society of Medicine* 104(9): 383–86.

Christiano, Lawrence, Martin Eichenbaum, and Sergio Rebelo. 2011. When Is the Government Spending Multiplier Large? *Journal of Political Economy* 119(1): 78–121.

Cole, Stephen R., and Miguel A. Hernán. 2008. Constructing Inverse Probability Weights for Marginal Structural Models. *American Journal of Epidemiology* 186(6): 656–664.

Davies, Gavyn. 2012. Why is the UK Recovery Weaker than the US? Financial Times, November 14. http://blogs.ft.com/gavyndavies/2012/11/14/why-is-the-uk-recovery-weaker-than-the-us/.

Eggertsson, Gauti B., and Paul Krugman. 2012. Debt, Deleveraging, and the Liquidity Trap: A Fisher-Minsky-Koo Approach. *Quarterly Journal of Economics* 127 (3): 1469–1513.

Glynn, Adam N., and Kevin M. Quinn. 2010. An Introduction to the Augmented Inverse Propensity Weighted Estimator. *Political Analysis* 18(1): 36–56.

Guajardo, Jaime, Daniel Leigh, and Andrea Pescatori. 2011. Expansionary Austerity: New International Evidence. IMF Working Paper 11/158. Forthcoming in *Journal of the European Economic Association*.

Giavazzi, Francesco, and Marco Pagano. 1990. Can Severe Fiscal Contractions Be Expansionary? Tales of Two Small European Countries. In *NBER Macroeconomics Annual 1990* edited by Oliver J. Blanchard and Stanley Fischer. Cambridge, Mass.: MIT Press, pp. 75–122.

Heckman, James J. 1976. The Common Structure of Statistical Models of Truncation, Sample Selection, and Limited Dependent Variables and a Simple Estimator for Such Models. *Annals of Economic and Social Measurement* 5(4): 475–492.

Hernández de Cos, Pablo, and Enrique Moral-Benito. 2013. Fiscal Consolidations and Economic Growth. *Fiscal Studies* 34(4): 491–515.

Hirano, Keisuke, Guido W. Imbens, and Geert Ridder. 2003. Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score. *Econometrica* 71(4): 1161–89.

Horvitz, Daniel G., and Donovan J. Thompson. 1952. A Generalization of Sampling without Replacement from a Finite Population. *Journal of the American Statistical Association* 47(260): 663–85.

Imbens, Guido W. 2004. Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review. *Review of Economics and Statistics* 86(1): 4–29.

Jordà, Òscar. 2005. Estimation and Inference of Impulse Responses by Local Projections. *American Economic Review* 95(1): 161–182.

Jordà, Òscar, Moritz Schularick, and Alan M. Taylor. 2013. When Credit Bites Back. *Journal of Money, Credit and Banking* 45(s2): 3–28.

Jordà, Òscar, and Alan M. Taylor. 2011. Performance Evaluation of Zero Net-Investment Strategies. NBER Working Paper 17150.

Kreif, Noémi, Richard Grieve, Rosalba Radice, and Jasjeet S. Sekhon. 2013. Regression-Adjusted Matching and Double-Robust Methods for Estimating Average Treatment Effects in Health Economic Evaluation. *Health Services and Outcomes Research Methodology* 13(2-4): 174–202.

Leeper, Eric M. 1997. Narrative and VAR Approaches to Monetary Policy: Common Identification Problems. *Journal of Monetary Economics* 40(3): 641–57.

Lunceford, Jared K., and Marie Davidian. 2004. Stratification and Weighting via the Propensity Score in Estimation of Causal Treatment Effects: A Comparative Study. *Statistics in Medicine* 23(19): 2937–60.

Mill, John Stuart. 1836. On the Definition of Political Economy; and on the Method of Philosophical Investigation in that Science. *London and Westminster Review* 26(1): 1–29.

Mountford, Andrew, and Harald Uhlig. 2009. What are the Effects of Fiscal Policy Shocks? *Journal of Applied Econometrics* 24(6): 960–92.

Nakamura, Emi, and Jón Steinsson. 2014. Fiscal Stimulus in a Monetary Union: Evidence from US Regions. *American Economic Review* 104(3): 753–92.

Owyang, Michael T., Valerie A. Ramey, and Sarah Zubairy. 2013. Are Government Spending Multipliers Greater during Periods of Slack? Evidence from Twentieth-Century Historical Data. *American Economic Review* 103(3): 129–134.

Parker, Jonathan A. 2011. On Measuring the Effects of Fiscal Policy in Recessions. *Journal of Economic Literature* 49(3): 703–18.

Perotti, Roberto. 1999. Fiscal Policy In Good Times And Bad. *Quarterly Journal of Economics* 114(4): 1399–1436.

Perotti, Roberto. 2013. The Austerity Myth: Gain without Pain? In *Fiscal Policy after the Financial Crisis* edited by Alberto Alesina and Francesco Giavazzi. Chicago: University of Chicago Press, pp. 307–54.

Posen, Adam. 2012. Why is their recovery better than ours? (Even though neither is good enough) Speech at the National Institute of Economic and Social Research, London, 27 March. http://www.bankofengland.co.uk/publications/Documents/speeches/2012/speech560.pdf

Ramey, Valerie A., and Matthew D. Shapiro. 1998. Costly Capital Reallocation and the Effects of Government Spending. *Carnegie-Rochester Conference Series on Public Policy* 48(1): 145–194.

Rendahl, Pontus. 2012. Fiscal Policy in an Unemployment Crisis. Cambridge Working Papers in Economics 1211.

Robins, James M., and Andrea Rotnitzky. 1995. Semiparametric Efficiency in Multivariate Regression Models. *Journal of the American Statistical Association*, 90(429): 122–129.

Robins, James M., Andrea Rotnitzky, and Lue Ping Zhao. 1994. Estimation of Regression Coefficients When Some Regressors are not Always Observed. *Journal of the American Statistical Association* 89(427): 846–66.

Robins, James M., Andrea Rotnitzky, and Lue Ping Zhao. 1995. Analysis of Semiparametric Regression Models for Repeated Outcomes in the Presence of Missing Data. *Journal of the American Statistical Association,* 90(429): 106–121.

Robins, James M. 1999. Robust Estimation in Sequentially Ignorable Missing Data and Causal Inference Models. *Proceedings of the American Statistical Association Section on Bayesian Statistical Science*. Alexandria, Va.: American Statistical Association, pp. 6–10.

Romer, Christina D., and David H. Romer. 1989. Does Monetary Policy Matter? A New Test in the Spirit of Friedman and Schwartz. In *NBER Macroeconomics Annual 1989* edited by Oliver J. Blanchard and Stanley Fischer. Cambridge, Mass.: MIT Press, pp. 121–70.

Romer, Christina D., and David H. Romer. 1997. Identification and the Narrative Approach: A Reply to Leeper. *Journal of Monetary Economics* 40(3): 659–65.

Rosenbaum, Paul R., and Donald B. Rubin. 1983. The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika* 70(1): 41–55.

Scharfstein, Daniel O., Andrea Rotnitzky, and James M. Robins. 1999. Adjusting for Nonignorable Drop-Out Using Semiparametric Nonresponse Models: Rejoinder. *Journal of the American Statistical Association* 94(448): 1135–46.

Schularick, Moritz, and Alan M. Taylor. 2012. Fact-checking financial recessions: US-UK update. VoxEU, October 24. http://www.voxeu.org/article/fact-checking-financial-recessions-us-uk-update.

Wooldridge, Jeffrey M. 1997. Quasi-Likelihood Methods for Count Data. In *Handbook of Applied Econometrics*, vol. 2, edited by M. Hashem Pesaran and Peter Schmidt. Oxford: Blackwell, pp. 352–406.

Wooldridge, Jeffrey M. 2007. Inverse Probability Weighted M-Estimation for General Missing Data Problems. *Journal of Econometrics* 141(2): 1281–1301.

Wooldridge, Jeffrey M. 2010. *Econometric Analysis of Cross Section and Panel Data.* 2nd edition. Cambridge, Mass.: MIT Press.

A. Appendix

A.1. OLS with Country-Fixed Effects and Controlling for World Growth

This section reports estimates of the OLS specification (equation (8) when the model is extended to include the World real GDP growth rate (from the World Bank dataset) as a control to capture global time varying trends. The following Table A1 corresponds to Table 2 using this alternative specification.

Table A1. Fiscal multiplier, d.CAPB, OLS estimate, booms v. slumps
Log real GDP (relative to Year 0, $\times 100$)

	(1) Year 1	(2) Year 2	(3) Year 3	(4) Year 4	(5) Year 5	(6) Sum
(a) Uniform effect of d.CAPB changes						
Fiscal multiplier, $y^C > 0$, boom	0.21***	0.25***	0.06	-0.18*	-0.26*	-0.07
	(0.07)	(0.07)	(0.05)	(0.10)	(0.14)	(0.24)
Observations	222	205	192	180	175	175
Fiscal multiplier, $y^C \leq 0$, slump	-0.03	-0.06	-0.17	-0.23*	-0.41**	-0.97**
	(0.03)	(0.06)	(0.10)	(0.12)	(0.17)	(0.37)
Observations	235	235	231	226	214	214
(b) Separate effects of d.CAPB for Large ($> 1.5\%$) and Small ($\leq 1.5\%$) changes						
Fiscal multiplier,	0.23***	0.25***	0.07	-0.17	-0.22	0.08
large change in CAPB, $y^C > 0$, boom	(0.08)	(0.08)	(0.06)	(0.10)	(0.14)	(0.27)
Fiscal multiplier,	0.04	0.19	-0.02	-0.35	-0.68	-1.68
small change in CAPB, $y^C > 0$, boom	(0.12)	(0.33)	(0.40)	(0.37)	(0.39)	(1.11)
Observations	222	205	192	180	175	175
Fiscal multiplier,	-0.03	-0.05	-0.18	-0.30*	-0.52**	-1.16**
large change in CAPB, $y^C \leq 0$, slump	(0.04)	(0.08)	(0.12)	(0.16)	(0.22)	(0.53)
Fiscal multiplier,	-0.05	-0.15	-0.10	0.13	0.16	0.03
small change in CAPB, $y^C \leq 0$, slump	(0.12)	(0.21)	(0.23)	(0.32)	(0.49)	(1.09)
Observations	235	235	231	226	214	214

Standard errors (clustered by country) in parentheses. ***/**/* indicates $p < 0.01/0.05/0.10$ respectively.
Additional controls: cyclical component of y, 2 lags of change in y, country fixed effects; and also growth rate of world real GDP (World Bank).
y^C is the cyclical component of $\log y$ (log real GDP), from HP filter with $\lambda = 100$.

A.2. IV with Country-Fixed Effects and Controlling for World Growth

The following Table A2 corresponds to Tables 4 when we add the World real GDP growth rate (from the World Bank dataset) as a control to capture global time varying trends.

Table A2. Fiscal multiplier, d.CAPB, IV estimate (binary), booms v. slumps
Log real GDP (relative to Year 0, $\times 100$)

	(1) Year 1	(2) Year 2	(3) Year 3	(4) Year 4	(5) Year 5	(6) Sum
Fiscal multiplier, $y^C > 0$, boom	-0.32	-0.33	-0.14	-0.54	-0.67	-1.18
	(0.32)	(0.52)	(0.51)	(0.45)	(0.45)	(1.54)
Observations	222	205	192	180	175	175
Fiscal multiplier, $y^C \leq 0$, slump	-0.24	-0.76***	-0.95***	-0.79**	-0.94**	-3.38***
	(0.15)	(0.24)	(0.31)	(0.32)	(0.42)	(1.10)
Observations	235	235	231	226	214	214

Standard errors (clustered by country) in parentheses. $^{***}/^{**}/^{*}$ indicates $p < 0.01/0.05/0.10$ respectively.
Additional controls: cyclical component of y, 2 lags of change in y, country fixed effects; and also growth rate of world real GDP (World Bank).
y^C is the cyclical component of $\log y$ (log real GDP), from HP filter with $\lambda = 100$.
d.CAPB instrumented by IMF fiscal action variable in binary 0-1 form (treatment).

A.3. Robustness

As discussed in the text, we explored the sensitivity of our results to different model specifications. These findings are shown in Table A3. In each case we show the impacts that these model changes have on the estimated 5-year summed estimate of the response of output to the fiscal treatment in the two output level bins. We also report the predictive ability test for the first stage in each case based on the area under the curve (AUC) statistic and its standard error.

In the main text we adopted a baseline specification of a pooled probit with country-fixed effects in the first-stage binary treatment regression. In column 1 we add the year-0 World real GDP growth rate (from the World Bank dataset) as a control to capture global time varying trends in both stages. In column 2 we show the first-stage using a pooled logit estimator with country-fixed effects. In column 3 we extend the estimator in column 3 and add the year-0 World real GDP growth rate (from the World Bank dataset) as a control to capture global time varying trends in both stages. Columns 4 and 5 report the results for the baseline probit model in the main text when probability weights are truncated to [0.1,0.9] and [0.2,0.8].

The message from these checks is that our results are not sensitive to the particular choice of first-stage model used to generate the propensity score. In the boom bin, effects are always small and statistically insignificant. In the slump bin the effects are negative and significant.

Table A3. ATE of fiscal consolidation, AIPW estimates, booms v. slumps, various propensity score models and truncations
Sum of log real GDP impacts, years 1 to 5 (all relative to Year 0, $\times 100$)

Estimator	(1) probit CFE + world GDP	(2) logit CFE	(3) logit CFE + world GDP	(4) probit CFE $p \in [0.1, 0.9]$	(5) probit CFE $p \in [0.2, 0.8]$
Fiscal ATE, $y^C > 0$, boom	-1.21	-1.04	-2.74	-1.26	-2.97
	(1.98)	(1.99)	(1.77)	(2.02)	(1.82)
Fiscal ATE, $y^C \leq 0$, slump	-3.76**	-4.31***	-3.19*	-3.87**	-3.87**
	(1.52)	(1.41)	(1.64)	(1.53)	(1.50)
First-stage, AUC	0.88	0.85	0.85	0.87	0.86
	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
Observations	389	389	389	389	389

Standard errors (clustered by country) in parentheses. ***/**/* indicates $p < 0.01/0.05/0.10$ respectively.
Additional controls: cyclical component of y, 2 lags of change in y, country fixed effects.
y^C is the cyclical component of $\log y$ (log real GDP), from HP filter with $\lambda = 100$.
AUC is the area under the Correct Classification Frontier (null = $\frac{1}{2}$); see text.
First-stage p-score models for the fiscal treatment are:
Column 1: As in Table 10, but including the year-0 World real GDP growth rate (from the World Bank dataset) as a control to capture global time varying trends.
Column 2: As in Table 10, but pooled logit estimator.
Column 3: As 2, but including the year-0 World real GDP growth rate (from the World Bank dataset) as a control to capture global time varying trends.
Column 4: As in Table 10, pooled probit, but probability weights truncated to [0.1,0.9].
Column 5: As in Table 10, pooled probit, but probability weights truncated to [0.2,0.8].

A.4. Estimated LP Equation for Future Treatment

For our U.K. counterfactuals we use LP-OLS estimates of future treatment as a response to treatment today. This allows us to compute expected and unexpected components of fiscal shocks in multi-year austerity programs, e.g. U.K. 2010–13. The estimates are shown in Table A4.

Table A4. LP estimate of impact treatment on future treatment, OLS estimates, booms v. slumps
Dependent variable: Treatment in year h (consolidation from year h to $h+1$)

	(1) Treatment $(t+1)$	(2) Treatment $(t+2)$	(3) Treatment $(t+4)$
Treatment (t)	0.509^{***}	0.281^{***}	0.171^{***}
	(0.054)	(0.055)	(0.042)
Observations	439	421	404

Standard errors (clustered by country) in parentheses. $^{***}/^{**}/^{*}$ indicates $p < 0.01/0.05/0.10$ respectively.
Additional controls: cyclical component of y, 2 lags of change in y, country fixed effects.
y^C is the cyclical component of $\log y$ (log real GDP), from HP filter with $\lambda = 100$.

A.5. Measures of U.K. Fiscal Consolidation 2010–13

Measures of the Size of U.K. Fiscal Treatments are shown in shown in Table A5. As discussed in the text, in our U.K. counterfactuals we use the change in the U.K. Office of Budget Responsibility (OBR) cyclically-adjusted primary balance as a measure of the scale of the fiscal treatment in each period (panel a). Alternative measures exist such as the OBR's cyclically-adjusted Treaty balance (panel b) or the IMF government structural balance (panel c) . ("Treaty" refers to Maastricht Treaty definitions.) All three paths are broadly similar; our preferred OBR cyclically-adjusted primary balance series (a) shows smaller changes than the other two series.

Table A5. OBR and IMF measures of the size of U.K. fiscal consolidations, 2010–2013
Levels and changes in percent of GDP

	(1)	(2)	(3)	(4)
Budget year	2009/10	2010/11	2011/12	2012/13
(a) OBR cyc.-adjust. primary bal. (used in text)	-6.8	-4.4	-2.9	-2.8 (-1.0)*
change	—	+2.3	+1.5	+0.1 (+1.9)*
cumulative change	—	+2.3	+3.8	+3.9 (+5.7)*
(b) OBR, cyc.-adjust. Treaty def., sign reversed	-9.5	-7.4	-5.9	-3.6
change	—	+2.1	+1.5	+2.3
cumulative change	—	+2.1	+3.6	+5.9
IMF calendar year	2010	2011	2012	2013
(c) IMF, government structural balance	-8.5	-6.6	-5.4	-4.0
change	—	+1.9	+1.2	+1.4
cumulative change	—	+1.9	+3.1	+4.5

Data from IMF WEO October 2012 database, HM Treasury Autumn Statements 2011 and 2012, and HM Treasury and OBR Budget 2013 and 2014 documents online. The data in panel (a) are updated based on March 2014 OBR updates and are consistent with the estimates computed by Simon Wren-Lewis (`http://mainlymacro.blogspot.com/2014/03/i-got-to-third-sentence-of-osbornes.html`). The figures in parentheses (*) indicate headline figures which include "distortions" due to credits taken for accounting adjustments involving the Bank of England's asset purchase program and the Royal Mail.