FEDERAL RESERVE BANK OF SAN FRANCISCO

WORKING PAPER SERIES

The Time for Austerity:
Estimating the Average Treatment Effect of Fiscal Policy

Òscar Jordà,
Federal Reserve Bank of San Francisco and University of California, Davis

Alan M. Taylor,
University of California, Davis, NBER, and CEPR

September 2013

Working Paper 2013-25
http://www.frbsf.org/publications/economics/papers/2013/wp2013-25.pdf

The views in this paper are solely the responsibility of the authors and should not be interpreted as reflecting the views of the Federal Reserve Bank of San Francisco or the Board of Governors of the Federal Reserve System.

September 2013

The Time for Austerity:

Estimating the Average Treatment Effect of Fiscal Policy [⋆]

Abstract

Elevated government debt levels in advanced economies have risen rapidly as sovereigns absorbed private-sector losses and cyclical deficits blew up in the Global Financial Crisis and subsequent slump. A rush to fiscal austerity followed but its justifications and impacts have been heavily debated. Research on the effects of austerity on macroeconomic aggregates remains unsettled, mired by the difficulty of identifying multipliers from observational data. This paper reconciles seemingly disparate estimates of multipliers within a unified framework. We do this by first evaluating the validity of common identification assumptions used by the literature and find that they are largely violated in the data. Next, we use new propensity score methods for time-series data with local projections to quantify how contractionary austerity really is, especially in economies operating below potential. We find that the adverse effects of austerity may have been understated.

Keywords: austerity, multipliers, booms, slumps, fiscal policy, output fluctuations, allocation bias, matching, average treatment effect, inverse probability weighting, local projection, identification.

JEL Codes: C54, C99, E32, E62, H20, H5, N10.

Òscar Jordà (Federal Reserve Bank of San Francisco and University of California, Davis)
e-mail: oscar.jorda@sf.frb.org; ojorda@ucdavis.edu

Alan M. Taylor (University of California, Davis, NBER, and CEPR)
e-mail: amtaylor@ucdavis.edu

[⋆] 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 or the Board of Governors of the Federal Reserve System. We thank seminar participants at the Federal Reserve Bank of San Francisco and the Swiss National Bank 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.

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

1 Introduction

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 be 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 tradition? In this paper we explore the key lesson, the need to ensure treatments are random or exogenous, in the context of the foremost academic and policy dispute of the day — the effects of fiscal policy shocks on output. For the optimistically minded, this might seem like a natural and helpful line of debate. Natural, since experimental techniques drawn from medicine have been fruitfully incorporated in other fields of economics, notably in applied microeconomic work in areas such as labor economics or development. Helpful, since clarity on fiscal impacts might be welcome, especially given the uproar over ongoing European and U.K. austerity programs.

Ironically enough, policy debates are now littered with medical metaphors. In 2011 German Finance Minister Wolfgang Schäuble wrote that "austerity is the only cure for the Eurozone"; while Paul Krugman 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."[2]

Bridging medicine and economics, and drawing from epidemiology and statistics, ideas from the experimental approach have slowly infected economics. Yet, in a recent survey of work in this tradition, Angrist and Pischke (2010) judge that "progress has been slower in empirical macro."[3]

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

[2] Quotations from ft.com, nytimes.com, ft.com, and ft.com, respectively.

[3] Even on the micro side the uptake wasn't immediate; Angrist and Pischke (2010) also note (following Meyer 1995) earlier social science research methods seen in texts from the 1960s and 1970s that were "well known in some disciplines, but distinctly outside the econometric canon."

The fear that standard empirical practices from the natural sciences would not work, especially for aggregate economic questions, is powerful and it took root after John Stuart Mill (1836), who pushed for reliance on *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." Yet even as he elevated deduction, Mill was capable of seeing (albeit without modern statistical methods at his disposal) how induction could work, even in macroeconomic settings; in the same essay he notably sketched out the essence of our approach here, a treatment-control framework using matching methods.[4]

2 The Austerity Debate

In this paper we lean on these empirical traditions as we speak to the macroeconomic debate over fiscal policy, and propose a simple empirical setup for use with a variety of estimation methods. We think that this approach provides a transparent and consistent framework which encompasses the key conflicting views (and the empirical designs behind them) in the current debate over expansionary versus contractionary austerity. This helps to clearly isolate the main driver of these different findings: the appropriate identification of the response to fiscal policy interventions. Into this setup we then introduce a treatment-control design which, we argue, offers an even more promising way to identify the effects of fiscal policy, since we can use statistical techniques designed for situations, experimental or otherwise, where underlying allocation bias may still persist. This problem of nonrandom allocation, as in many clinical situations, turns out to be very serious here — as it is likely to be in many other non-experimental macroeconomic contexts where policy choices are afflicted by an underlying "insufficient randomization" problem.

Our argument proceeds in three stages and can be summed up as follows. For consistency we utilize only the OECD annual panel dataset that is common to the two recent high-profile yet seemingly irreconcilable studies: the "expansionary austerity" idea has come to be defined by the paper of Alesina and Ardagna (2010, henceforth AA), but an IMF study reached the opposite conclusion of "contractionary austerity" (Guajardo, Leigh, and Pescatori 2011; henceforth GLP).[5]

We use the local projection (LP) method (Jordà 2005) to estimate output impacts of fiscal policy dynamically up to 4 years out. These LP methods make it simple to allow for possibly nonlinear,

[4] Mill: "The consequence of this unavoidable defect in the materials of the induction is, that we can rarely obtain what Bacon has quaintly, but not unaptly, termed an *experimentum crucis*.... In any science which admits of an unlimited range of arbitrary experiments, an experimentum crucis may always be obtained... But this can seldom be done in the moral sciences, owing to the immense multitude of the influencing circumstances, and our very scanty means of varying the experiment.... How, for example, can we obtain a crucial experiment on the effect of a restrictive commercial policy upon national wealth? We must find two nations alike in every other respect, or at least possessed, in a degree exactly equal, of everything which conduces to national opulence, and adopting exactly the same policy in all their other affairs, but differing in this only, that one of them adopts a system of commercial restrictions, and the other adopts free trade. This would be a decisive experiment, similar to those which we can almost always obtain in experimental physics. Doubtless this would be the most conclusive evidence of all if we could get it. But let any one consider how infinitely numerous and various are the circumstances which either directly or indirectly do or may influence the amount of the national wealth, and then ask himself what are the probabilities that in the longest revolution of ages two nations will be found, which agree, and can be shown to agree, in all those circumstances except one?"

[5] By setting up these two ideas in this oppositional way we follow Perotti (2013), who presents a lucid discussion of the empirical pitfalls in this research area. The idea of expansionary austerity dates to Giavazzi and Pagano (1990).

or state-dependent dynamic responses, and indeed we will find that the effects of fiscal policy can be very different in the boom and the slump, as emphasized by Keynes in the 1930s.[6] But the LP method per se does not guarantee that the measured response to a fiscal intervention is appropriately identified. It has to be accompanied by a proper identification strategy and a great deal of our analysis dwells on how to confront this issue.

As a first step we estimate LP-OLS impacts of fiscal policy shocks using the AA measure of policy, the change in the cyclically-adjusted primary balance (d.CAPB).[7] We can consider all such shocks, or we can 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 (1997). Either way, we replicate previous results by these authors and in the IMF study: austerity is expansionary using this estimator. 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 these expansionary effects of fiscal consolidation evaporate (but there are no significant contractionary effects). It remains to be determined whether the focus on large fiscal consolidations is sufficient to achieve identification. We show below this is unlikely.

As a next step we replace the LP-OLS estimator with an LP-IV estimator, where the cyclically-adjusted primary balance is instrumented by the IMFs narrative measure of an exogenous fiscal consolidation. This type of "narrative-based identification" has been applied by, e.g., Ramey and Shapiro (1998) and Romer and Romer (2010). Here we replicate the GLP results: austerity is contractionary.

We then offer a new take in the third and final step. Here we consider the IMF indicator as a "fiscal treatment" — i.e., a binary indicator rather than a continuous variable — and we are interested in characterizing the *average treatment effect* (ATE). However, 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.[8]

Our worry is well founded. The IMF treatment 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).[9] Ultimately, we can incorporate other omitted controls to get as complete a picture as possible of the determinants of a fiscal consolidation event.

[6] This issue of state-dependent multipliers has 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) but we use a new identification approach and take the question directly to the data at the heart of the debate on the AA/GLP findings. Other recent papers on state-dependent fiscal multipliers, using various measures of slack, include Barro and Redlick (2011) and Nakamura and Steinsson (2011). For a critical survey see Parker (2011). Longer ago, Perotti (1999) explored the idea of "expansionary austerity" with state-dependent fiscal multipliers.

[7] The CAPB used by AA is based on Blanchard (1993). It consists of adjusting for cyclical fluctuations using the unemployment rate.

[8] 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).

[9] 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 (2012) use economic variables as instruments. We also use economic variables, but with a different estimating approach for reasons as discussed below.

In order to purge allocation bias we use inverse probability weighting (IPW) estimation. In new work, Angrist, Jordà and Kuersteiner (2013, or AJK) introduce simple average IPW estimators to calculate impulse responses in time series data and, drawing on their ideas, we apply similar methods to our problem. IPW requires a first stage model that characterizes the likelihood of treatment. In AJK this model is called the *policy propensity score*. Intuitively, the goal of reweighting in this framework is to focus the estimator on a rebalanced sample in each part of the treatment and control groups that closely resemble each other, and so obtain an ATE estimate as if the allocation had been randomly assigned.

However, because we are interested in keeping a similar framework across experiments to facilitate comparability, we will apply IPW regression adjusted estimation. This estimator falls into the broad class of "doubly robust" estimators of which Robins, Rotnitzky and Zhao (1994) is perhaps the earliest reference. The "doubly robust" or DR moniker comes from controlling for observables through the conditional mean (the local projection part) and through (inverse propensity score) reweighing. If either the propensity score model or the conditional mean model are misspecified, but not both, the estimator will still be consistent for the ATE. We call such an estimator LP-IPWRA when applied to measure dynamic responses, as here.

The IPW and IPWRA estimators have been widely used in medical statistics. In economics, a seminal work is Hirano, Imbens, and Ridder (2003), and since that work the relevance of the IPW estimator for the social sciences has become clear, in areas ranging from political science to applied microeconomics. In macroeconomics, the novelty of the Angrist, Jordà, and Kuersteiner (2013) IPW estimator is to apply similar ideas to evaluating the effects of monetary policy in a time-series setting and providing the proper way to do so. Imbens (2004) reviews IPWRA estimators in the context of a wider literature on this approach and Graham, Campos De Xavier Pinto and Egel (2012) extend the IPWRA framework to problems with missing data in the context of a study of the effects of Black-White differences in cognitive achievement on earnings. The LP-IPWRA fits naturally between these two strands of the literature.

What do we find? Our results contrast with the expansionary austerity view of AA, and tend to echo or even amplify the opposing view of the IMF. Indeed, we find even stronger results than the IMF: austerity is contractionary when using the LP-IPWRA estimator; the effect in slumps is stronger and of even higher statistical significance; and even in booms there are signs of drag.

Clearly, allocation bias is 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 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*. But even when in a bad current state the economy is still 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.

3 Identification

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 grow beyond the comfort of bond markets, sovereigns have little choice but 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 validity of the implicit or explicit identification assumptions particular to the empirical approach used. One of the recurring themes in the analysis that follows is this: the divergence of results in the fiscal policy debate can be reconciled by determining when identification conditions are more or less likely to have been met.

The interplay between the modeling approach and the identification assumptions made in the two poles on the fiscal policy debate can be summarized as follows. Studies based on vector autoregressions implicitly characterize the system of difference equations that determine the dynamics for outcomes and policy interventions jointly. Identification of the causal effect of a fiscal policy intervention relies on what we could call the weak form of the selection-on-observables assumption. First, these papers assume that the variables included in the VAR (raw, transformed, filtered, or otherwise) are sufficient to identify exogenous sources of variation in fiscal policy interventions. Second, it is implicitly assumed that a regression control strategy is sufficient to remove allocation bias. The advantage of this approach is that if the untested identification assumptions are correct and the model explaining how outcomes and policy interventions is well characterized by linearity, counterfactual experimentation is simple to implement. Experiments involving fiscal policy interventions, or for that matter other shocks, can be traced over time and across covariates. The preferred tool of communication for these experiments is a plot of structural impulse response functions.

There are several potential weaknesses to this approach. The set of observables used to achieve identification is limited to the dimension of the VAR. Identification is often an article of faith even though it need not be (and the literature is gradually adapting accordingly). One could test whether the structural shocks identified can be explained by pre-determined values of covariates included as well as excluded in the VAR, although this is not usually done. Identification also relies on the joint model of outcomes and policy interventions being linear. Nonlinear specifications of a VAR are in principle available, but they are rarely used because of their complexity. As we shall see, the response of the economy to a fiscal policy intervention, conditioned on the state of the economy at the time of intervention, can vary substantially (see Auerbach and Gorodnichenko 2012, 2013). That is, the assumption of linearity is largely rejected by the data.

Auerbach and Gorodnichecko (2013) and Owyang, Ramey, and Zubairy (2013) use the local projection (LP) approach introduced by Jordà (2005) to break free from some of the assumptions in the VAR approach. Local projections have several advantages over VARs. The average response to policy can be modeled more richly at little cost because responses can be calculated on a per variable basis. For example, Auerbach and Gorodnichenko (2013) model the conditional mean expression for outcomes with a smooth two-state nonlinear model that allows the economy's response to vary as a function of the state of the economy.

More importantly, the local projection approach improves on VARs on the identification front. There are two ways in which this can be achieved. First, the conditioning set is not limited to covariates in the system. Other variables and their lags can be brought into the specification of the conditional mean to achieve a, let's call it, stronger-form of selection-on-observables. Stronger and not simply strong because the effect of additional controls is limited by the functional form of the conditional mean (usually linear). But what about selection-on-unobservables?

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 calculate the causal effect. Auerbach and Gorodnichecko (2013) and Owyang, Ramey, and Zubairy (2013) have used this approach.

The IMF study (GLP) has generated a set of potential instruments. The narrative instruments in GLP consist of dates of fiscal consolidations that, through a reading of the historical record, the authors claim can be reasonably considered to be exogenous. These can be used in conjunction with the local projection approach to obtain a more reliable measure of the effects of fiscal policy on macroeconomic outcomes. In what follows, we show that many of the results in AA can be reconciled with the results in GLP when untying the VAR straightjacket in favor of instrumental variable local projection methods (LP-IV).

Naturally, the key question one may ask is how exogenous these instruments really are? And, if they are not, what identification approach is there left? As a preview of results reported later in the paper, we are able to show that the episodes of fiscal consolidation identified by GLP can be predicted using pre-determined macroeconomic controls, perhaps not surprisingly. Hence these are not really exogenous and may not be valid instruments. So what can be done?

Angrist, Jordà, and Kuersteiner (2013) attack this identification conundrum using an inverse probability weighted (IPW) estimator for time-series estimation of an average treatment effect (ATE). In the context of our problem, this approach can be adapted as follows. Starting with the IMF dataset, the episodes of fiscal consolidation identified by GLP can serve to narrow the set of policy fiscal interventions thought to be exogenous. Next, we can think of these interventions in discretized fashion and construct a policy prediction model for an "austerity" intervention — the policy propensity score. The model indicates that many of the consolidations in the GLP database would have been predicted by macroeconomic controls.

Next, we rely again on selection-on-observables arguments to calculate causal effects using a matching estimator applied to the LP framework. The principles behind the AJK 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. Robins, Rotnitzky and Zhao (1994), Robins (1999), Lunceford and Davidian (2004), and Kreif, Grieve, Radice and Sekhon (2011) discuss IPW regression adjusted estimators that deliver greater robustness and efficiency. These two strands of thinking naturally lead to our approach — combining the AJK inverse-weighting methods with an LP regression framework. We will describe the LP-IPWRA estimator and its advantages over existing methods below, but there is a basic intuition for how IPW works. Weighting by the inverse of the probability, or propensity score, puts more weight on observations for treated outcomes ("austerity") for which a policy intervention had a low probability of occurring according to the policy model. Similarly, observations for untreated outcomes ("no austerity") receive more weight when the policy model predicted a policy intervention.

The next section introduces the local projection approach which will serve as a platform to try to reconcile the divergent views offered by the literature. We then go on to present OLS and IV results, before probing the assumed exogeneity of the IMF fiscal consolidation episodes. The paper then switches to the LP-IPWRA approach, leading us to ask how this estimator moves the needle in the great fiscal multiplier debate. Does it tilt our views of austerity toward the AA "expansionary" view, or do we confirm the IMF's "contractionist" conclusion?

4 Local Projection Methods

Our empirical work begins with results that replicate the two main poles of the fiscal policy debate so far. Since empirical methodologies and specifications differ across the many studies in this literature, we try to preserve a controlled, uniform empirical design so as to emphasize the fundamental differences arising from the identification assumptions used.

Throughout the paper we use the local projection framework (LP) due to Jordà (2005) to estimate h-step ahead cumulative forecasts of an economy's output response to a fiscal intervention at time $t = 0$, up to 4 years out. The LP method is flexible enough to still encompass all of our different identifications and refinements, including instrumental variables and nonlinearities, as well as the matching methods just noted. As the LP method is still somewhat unfamiliar in macroeconomics, we briefly describe its key features for this application. The reader is referred to Jordà (2005, 2009) for a detailed exposition. In what follows, we use richer notation from AJK that will be useful when we later describe the IPW and IPWRA estimators.

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_0 and d_1. In addition, we consider the possibility that there is a k_x-dimensional vector of variables, x_t that are not included in the vector y_t, but which could be relevant predictors of the policy variable D_t. To account for the possible availability of instrumental variables, define the k_z-dimensional vector z_t of instruments. Finally, denote w_t the rich conditioning set given by $\Delta y_{t-1}, \Delta y_{t-2}, ...; D_{t-1}, .D_{t-2}, ...$; and x_t. In particular, we assume that policy is determined by $D_t = D(w_t, \psi, \varepsilon_t)$ where ψ refers to the parameters of the implied policy function and ε_t is an idiosyncratic source of random variation. Therefore, $D(w_t, \psi, .)$ refers to the systematic component of policy determination.

Next, we will borrow from definition 1 in Angrist, Jordà and Kuersteiner (2013). This defines *potential outcomes* given by $y^{\psi}_{t,h}(d_j) - y_t$ as the value that the observed outcome variable $y_{t+h} - y_t$ would have taken if $D_t = d_j$ for all $\psi \in \Psi$ and all possible realizations $d_j \in \mathcal{D}_t$. Specifically and in the context of our application, the difference $y_{t+h} - y_t$ refers to the cumulative change in the outcome from t to $t+h$. Using $D_t = d_0$ to denote a baseline policy intervention, the causal effect of a policy change is defined as the unobservable random variable given by the difference $(y_{t,h}(d_j) - y_t) - (y_{t,h}(d_0) - y_t)$. Notice that y_t is only used to benchmark the cumulative change and it is observed at time t. We will drop the superscript ψ from here on for clarity; that is, we assume that the parameters of the policy function do not change.

Given a fixed policy regime, so that ψ is fixed, then observed outcomes are given by the following latent variables model:

$$y_{t+h} - y_t = \sum_{d \in \mathcal{D}} y_{t,h}(d) \mathbf{1}\{D_t = d\} - y_t.$$

Angrist, Jordà and Kuersteiner (2013) state the selection-on-observables assumption (or the *conditional ignorability* or *conditional independence* assumption as it is sometimes called) as

$$(y^{\psi}_{t,h}(d_j) - y_t) \perp D_t | w_t; \psi \quad \text{for all } h \geq 0, \text{and for all } d_j, \text{and } \psi \in \Psi. \tag{1}$$

This conditional independence assumption will play an important role later on when we discuss the AJK estimator. Under the selection-on-observables assumption in (1), and further assuming that a regression control strategy suffices to do the appropriate conditioning, the average causal effect of a policy intervention d_j relative to a baseline d_0 on the outcome variable at time $t+h$, given by

$$E\left[(y_{t,h}(d_j) - y_t) - (y_{t,h}(d_0) - y_t))\right]$$

can be calculated by the local projection

$$y_{t+h} - y_t = \alpha^h + \theta^h D_t + \gamma^{h\prime} w_t + v_{t+h}; \quad \text{for } h = 0, 1, ..., H, \tag{2}$$

assuming that the per-period conditional mean can be linearly approximated. Country fixed effects are momentarily omitted to simplify the exposition. With panel data, these could be incorporated easily as we do in our application. More elaborate specifications of the local projection in (2) are, of course, possible. Auerbach and Gorodnichenko (2013) is one example. We will also relax this assumption later on, but for now, presentation of the main results is facilitated with this simpler environment.

We can then write

$$\begin{aligned} & E\left[(y_{t,h}(d_j) - y_t) - (y_{t,h}(d_0) - y_t)\right] \\ = & E\left[E\left[y_{t+h} - y_t | D_t = d_j; w_t\right] - E[y_{t+h} - y_t | D_t = d_0; w_t]\right] \\ = & \theta^h \left(d_j - d_0\right); \quad \text{for } h = 1, ..., H. \end{aligned} \tag{3}$$

Note that θ^h can be easily estimated using OLS in expression (2). Notice also that the local projection directly conditions on observables (under the assumption of linearity) and facilitates the computation of (3). Expression (3) is equivalent to the impulse response estimated from a VAR when w_t is limited to lags of the outcome variables but does not include x_t under the maintained assumption of linearity and correct specification (see Jordà 2005).

If the conditional independence assumption (1) fails, OLS applied to (2) will deliver a biased and inconsistent estimate of θ^h. Instrumental variables can be brought in to fix this inconsistency, but need to meet two well-known conditions. First, they need to be independent of the unobserved selection mechanism. Second, the instruments z_t need to be predictive for D_t.

Assuming these two conditions are met, estimation of the response to policy interventions in expression (3) using local projections in expression (2) estimated by instrumental variables methods with z_t will deliver a consistent estimate of θ_h.

In summary, local projection methods afford a very straightforward way to contrast the effect of estimating fiscal multipliers under implicit selection-on-observables assumptions (LP-OLS) relative to estimates where that assumption fails due to selection-on-unobservables but instruments are available (LP-IV). The differences between the two estimators will be revealing, and form 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.

5 Replicating Expansionary Austerity: LP-OLS Results v. AA

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 4, and deviations from year 0 levels are shown, and also the sum of these deviations, or lost output across all of those four years.

The typical LP equation that we estimate has the following form:

$$y_{i,t+h} - y_{i,t} = \alpha_i^h + \theta^h D_{i,t+1} + \gamma_0^h \Delta y_{i,t} + \gamma_1^h \Delta y_{i,t-1} + \xi^h y_{i,t}^C + v_{i,t+h} \tag{4}$$

for $h = 1, ..., 4$ and with $y_{i,t+h} - y_{i,t}$ denoting the accumulated change from time t to $t+h$ in 100 times the log of GDP, α_i^h are country-fixed effects, $D_{i,t}$ denotes the d.CAPB variable. This variable is measured from time t to time $t+1$ for consistency with the GLP study and reflects the timing

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

Log real GDP (relative to Year 0, ×100)

	Year 1	Year 2	Year 3	Year 4	Sum
Fiscal multiplier, full sample	0.11*	0.12*	-0.04	-0.21*	-0.08
	(0.04)	(0.05)	(0.04)	(0.07)	(0.09)
Observations	457	440	423	406	406
Fiscal multiplier, large consolidation $> 1.5\%$	0.12*	0.13*	-0.04	-0.23*	-0.07
	(0.04)	(0.05)	(0.04)	(0.07)	(0.12)
Observations	457	440	423	406	406
Fiscal multiplier, small consolidation $\leq 1.5\%$	0.06	0.11	0.03	-0.07	-0.07
	(0.07)	(0.15)	(0.14)	(0.19)	(0.41)
Observations	457	440	423	406	406

Standard errors (clustered by country) in parentheses. $+p < 0.10$, $*p < 0.05$.
Additional controls: cyclical component of y, 2 lags of change in y, country fixed effects.
Large consolidation means change in CAPB larger than 1.5% of GDP; small means less than or equal to 1.5% of GDP.

of the announcement and implementation of fiscal plans. Finally, the term $y^C_{i,t}$ denotes the cyclical component of GDP measured as deviations from an HP trend estimated with a smoothing parameter of 100, which is typical for annual data. This specification nests the main elements in AA and GLP to facilitate comparisons of our results with theirs.

In a parallel with the main result in AA, Table 1 reports estimates based on expression (4). Although the effect is economically modest, the data appear to support the notion that fiscal consolidation can be expansionary (specially in the first two years), although the cumulative effect over a four year period is largely negligible. 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), the results are almost identical. Small consolidation packages have close to a zero effect, but the estimates are 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?

To allow responses to be state dependent, estimation is carried out on two bins of the data, where 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 both the "boom" bin and the "slump" bin, given the AA-GLP combined dataset with just almost 450 observations in total, after allowing for observations lost due to lags.

Table 2 shows OLS estimated responses using expression (4) 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

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

Log real GDP (relative to Year 0, ×100)

(a) Uniform effect of d.CAPB changes					
	(1) Year 1	(2) Year 2	(3) Year 3	(4) Year 4	(5) Sum
Fiscal multiplier, $y^C > 0$, boom	0.21*	0.24*	0.05	-0.17	0.23
	(0.06)	(0.07)	(0.05)	(0.10)	(0.16)
Observations	222	205	192	180	180
Fiscal multiplier, $y^C \leq 0$, slump	-0.03	-0.07	-0.17	-0.23+	-0.55+
	(0.03)	(0.06)	(0.10)	(0.12)	(0.27)
Observations	235	235	231	226	226

(b) Separate effects of d.CAPB for large ($> 1.5\%$) and small ($\leq 1.5\%$) changes					
	(1) Year 1	(2) Year 2	(3) Year 3	(4) Year 4	(5) Sum
Fiscal multiplier, $y^C > 0$, boom, large consolidation $> 1.5\%$	0.23*	0.24*	0.06	-0.15	0.32
	(0.08)	(0.08)	(0.05)	(0.10)	(0.20)
Fiscal multiplier, $y^C > 0$, boom, small consolidation $\leq 1.5\%$	0.06	0.21	-0.04	-0.32	-0.61
	(0.11)	(0.33)	(0.38)	(0.35)	(1.00)
Observations	222	205	192	180	180
Fiscal multiplier, $y^C \leq 0$, slump, large consolidation $> 1.5\%$	-0.02	-0.05	-0.18	-0.30$^+$	-0.62$^+$
	(0.04)	(0.08)	(0.12)	(0.15)	(0.36)
Fiscal multiplier, $y^C \leq 0$, slump, small consolidation $\leq 1.5\%$	-0.05	-0.16	-0.10	0.13	-0.18
	(0.12)	(0.20)	(0.22)	(0.31)	(0.68)
Observations	235	235	231	226	226

Standard errors (clustered by country) in parentheses. $+p < 0.10$, $*p < 0.05$.

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.

(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 LP-OLS forecasts suggest that fiscal austerity 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 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 between a coefficient or multiplier of about 0.2 in years 1 and 2. Over 4 years the sum of these effects is small, also about 0.2. In the slump bin, the estimate of the policy response is not statistically different from zero and in many cases it is negative.

6 Replicating Contractionary Austerity: LP-IV Results v. IMF

One widely shared concern with the LP-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 IMFs 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 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 IMFs 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 in a manner consistent with our LP-OLS replication of the AA results, we present in Tables 3, 4 and 5 some LP-IV estimates which make use of the IMF variables. We reestimate expression (4) using the IMF dates of fiscal consolidations as 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). The IMFs narrative variables, if we can assume that they are valid instruments, cannot be described as weak.

We begin by estimating 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 on dates of fiscal consolidation as a binary instrument (first row). Second, we use the IMF narrative variables as a continuous variable. That is, directly using the size of the consolidation identified by the IMF as an instrument (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 IMF results (see GLP). Fiscal consolidation is unambiguously contractionary. Using the sum of coeffi-

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

Log real GDP (relative to Year 0, $\times 100$)

	(1) Year 1	(2) Year 2	(3) Year 3	(4) Year 4	(5) Sum
Fisc multiplier, full sample, binary IV	-0.34*	-0.72*	-0.76*	-0.78*	-2.29*
	(0.11)	(0.22)	(0.24)	(0.22)	(0.69)
Fisc multiplier, full sample, continuous IV	-0.46*	-0.81*	-0.69*	-0.58*	-2.28*
	(0.12)	(0.22)	(0.29)	(0.27)	(0.81)
Observations	457	440	423	406	406

Standard errors (clustered by country) in parentheses. $+p < 0.10$, $*p < 0.05$.
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 variable in the bottom panel.

Table 4: Fiscal multiplier, effect of d.CAPB, LP-IV estimates (binary IV), 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) Sum
Fiscal multiplier, $y^C > 0$, boom	-0.34	-0.32	-0.13	-0.59	-0.87
	(0.30)	(0.46)	(0.47)	(0.47)	(1.36)
Observations	222	205	192	180	180
Fiscal multiplier, $y^C \leq 0$, slump	-0.25^{+}	-0.76*	-0.95*	-0.79*	-2.68*
	(0.14)	(0.23)	(0.29)	(0.31)	(0.84)
Observations	235	235	231	226	226

Standard errors (clustered by country) in parentheses. $+p < 0.10$, $*p < 0.05$.
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).

cients reported in column (5) 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 four subsequent years. This result is not sensitive to whether the IMF narrative variable is used as a binary or as a 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 reproduce much of the analysis in Table 2 using instrumental variables based on the binary and the continuous versions of the IMF narrative variable. These results are reported in Tables 4 and 5 respectively. Again, the precise choice of IV makes very little difference to the overall message.

The LP-IV 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 4. Over 4 years the sum of these effects is

Table 5: Fiscal multiplier, effect of d.CAPB, LP-IV estimates (continuous IV), booms v. slumps
Log real GDP (relative to Year 0, ×100)

	(1) Year 1	(2) Year 2	(3) Year 3	(4) Year 4	(5) Sum
Fiscal multiplier, $y^C > 0$, boom	-0.51*	-0.57$^+$	-0.13	-0.39	-1.12
	(0.26)	(0.31)	(0.32)	(0.43)	(1.08)
Observations	222	205	192	180	180
Fiscal multiplier, $y^C \leq 0$, slump	-0.35*	-0.68*	-0.71*	-0.55$^+$	-2.24*
	(0.15)	(0.27)	(0.35)	(0.31)	(1.01)
Observations	235	235	231	226	226

Standard errors (clustered by country) in parentheses. $+p < 0.10$, $*p < 0.05$.
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 continuous form (total, in % of GDP).

-2.68, 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. Similar conclusions, with slightly smaller impacts, are seen in Table 5, where the continuous IV is used, but the results are qualitatively similar.

7 Endogenous Austerity: The Fiscal Treatment is not Randomly Allocated

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 the 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 Keynes was right?

Before drawing any conclusions, we must push a little further to deal with a nagging but potentially important problem: is the IMF narrative variable a legitimate instrument? Have we identified the causal effect of fiscal consolidations on output? If the IMF's narrative variable can be predicted by 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.

Table 6: 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. ${}^{+}p < 0.10; {}^{*}p < 0.05$

In the ideal RCT, with treatment and control units allocated randomly, the probability density function of the controls would be the same for each subpopulation — there would be perfect overlap between the two. A simple way to check for this *balance* condition, as it is often referred to in the literature, is to do a simple test of the equality of the means across subpopulations. This is reported in Table 6 for the four control variables we have entertained in the specification of the LP-OLS and LP-IV models. The results indicate that the null hypothesis is rejected for all of them, strongly suggesting that the IMF narrative dates cannot be considered 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 7 we report the results of such tests by reexamining whether our candidate model in expression (4) admits as 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 (4).

Table 7: Omitted Variables Explain Output Fluctuations

		Model	
	OLS	IV (binary)	IV (continuous)
Real GDP growth	0.00	0.00	0.00
Real private loan growth	0.22	0.29	0.39
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	0.00	0.00
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 in Table 5

Table 8: Fiscal treatment regression, pooled probit estimators (average marginal effects)

Model	(1) Treatment (t+1)	(2) Treatment (t+1)	(3) Treatment (t+1)	(4) Treatment (t+1)
Public debt/GDP (t)	0.32*	0.27*	0.11$^+$	0.11$^+$
	(0.07)	(0.07)	(0.06)	(0.06)
Cyclical component of log y (t) (y^C)		-0.02*	-0.01	
		(0.01)	(0.01)	
Growth rate of output (t)		-0.03*		-0.02*
		(0.01)		(0.01)
Treatment (t)			0.42*	0.42*
			(0.02)	(0.02)
Observations	459	459	459	459
Predictive ability test, AUC	0.61	0.65	0.81	0.82
s.e.	0.03	0.03	0.02	0.02

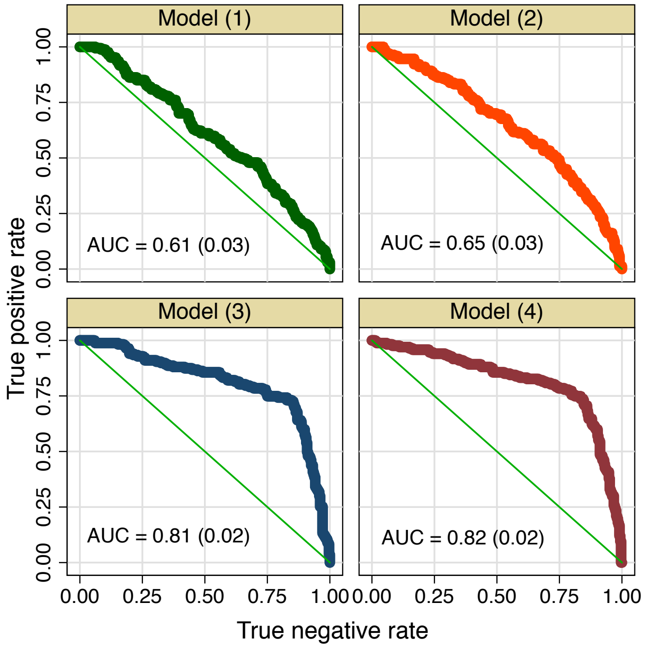
Standard errors in parentheses. $+p < 0.10$, $*p < 0.05$.
y^C is the cyclical component of $\log y$ (log real GDP), from HP filter with $\lambda = 100$.
AUC is the area under the Correct Classification Frontier (null = $\frac{1}{2}$); see text.

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 findings since the tests would rely on a smaller sample. Table 7 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. 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 these omitted controls predict fiscal consolidations? Table 8 shows whether variation in spending levels and outcomes could have been explained by omitted controls, that is, could the IMF binary treatment variable identified by GLP 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. Table 8 reports probit estimates. As shown in an appendix, the 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 8 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

Figure 1: Correct Classification Frontiers for 4 Probit Treatment Models



Notes: See text. The CCF plots all true positive rates $TP(c)$ and true negative rates $TN(c)$ for a treatment classifier based on $I(\hat{p} > c)$, for all values of the threshold c, where $\hat{p}$ is the fitted value from the probit model. AUC denotes area under the curve, with the reference null uninformative classifier having AUC equal to $\frac{1}{2}$. Standard errors in parentheses.

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, 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.

For further confirmation of the predictive ability of these treatment regressions, Figure 1 shows the correct classification frontier (CCF) for each of the 4 models in Table 8 (for details on the CCF, see Jordà and Taylor 2011). Each of these CCF lines plots all true positive rates $TP(c)$ and true negative rates $TN(c)$ for a treatment classifier based on $I(\hat{p} > c)$, for all values of the threshold c, where $\hat{p}$ is the fitted value from the probit model. If c is set to $-\infty$ then we are at one end of the unit simplex, if c is set to $+\infty$ then we are at the other end of the unit simplex. An uninformative classifier would, as we vary c, trace out a CCF along the simplex, and would be no better than a random signal. Informative classifiers generate a CCF that lies above the simplex, and the simplest predictive ability test is whether the area under the curve or AUC exceeds 0.5, a statistic that is asymptotically normal and easy to compute.

The AUC statistics in Table 8 and the CCF curves in Figure 1 show that the probits have very good predictive ability, with AUC at best around 0.65 when lagged treatment is omitted (Model 2), and rising to around 0.8 with lagged treatment (Columns 3 and 4), and these AUCs are all significantly different from 0.5.

The key lesson from Table 8 is simply that the IMF treatment variable has a significant forecastable component. Since the same controls also affect the outcome (see Table 7), 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. The question, then, is how to deal with this problem. The remainder of this paper provides one answer.

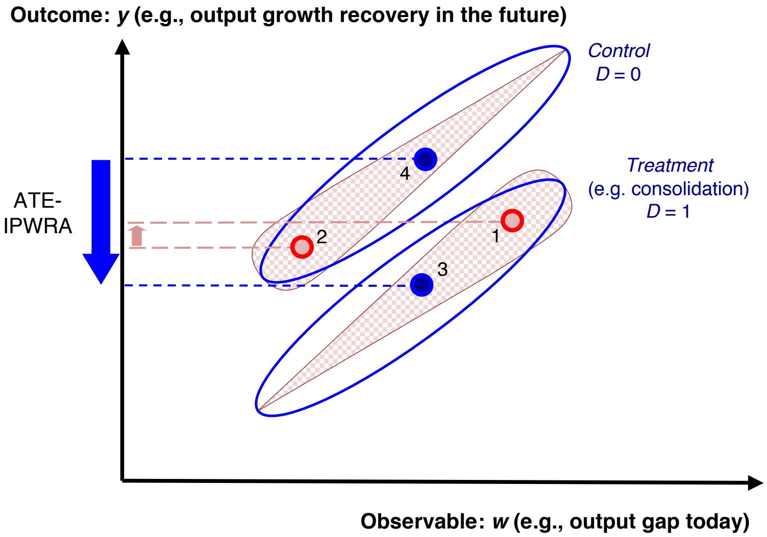
8 Average Treatment Effects for Macroeconomics

Based on the results reported so far, proper identification of the causal effects of fiscal consolidation remains elusive. In this section we turn our attention to the principles behind medical experimental designs to try to make some headway. It is often hard for empirical designs to guarantee that the allocation of treatment, by those actually designated to make that "policy" choice, will remain unaffected by the characteristics of the treated. Alternation, as in the 1809 bloodletting trial, was the first step round this, but by the twentieth century it was known that this design could be polluted by those administering trials. For example, there could be a temptation to ensure that healthier patients, with a better chance of recovery, get a shot at some potential miracle drug. If best practice randomization is properly followed, an average treatment effect (ATE) will be unbiased. But many tests for drugs and therapies cannot guarantee these ideal conditions. The world of policy and macroeconomics is similarly imperfect and it is well understood that policy endogeneity creates a risk of allocation bias.

The hypothetical Figure 2 displays the perils of allocation bias with a simple bivariate example using one covariate for illustration. The vertical axis is an outcome, here future output relative to today; the horizontal axis is an observable, here the output gap today, which is assumed to be predictive for whether treatment (fiscal consolidation) is implemented or not. In general, the observable could be a vector of variables. Here high output gap, or a depressed economy, is associated with austerity treatment, as in our probit. But the output gap may also be correlated with the short-run outcome, e.g., because of convergence to the long run output trend. In this setup, direct comparison of the sample means of the two populations, treated versus control, would suggest that fiscal consolidations are expansionary. Of course, this is a result of the positive association between the outcome and the observable and nothing to do with the effectiveness of the treatment, which in reality is contractionary. The example is similar to the observation that individuals with high blood pressure tend to have cardiovascular disease. We would not therefore conclude that statins (which are usually taken by those with high blood pressure) cause cardiovascular disease.

IPW-based methods like AJK and our LP-IPWRA are based on a propensity score model. Inverse weighting can be seen as a form of matching, and indeed a form of matching applied to the problem of unobserved potential outcomes — that is, a type of missing data problem.

Figure 2: An Example of Allocation Bias and the IPW Estimator


Outcome: *y* (e.g., output growth recovery in the future)
Control
D = 0
Treatment
(e.g. consolidation)
D = 1
ATE-
IPWRA
4
2
1
3
Observable: *w* (e.g., output gap today)

Notes: See text. The shaded teardrop shapes indicate the hypothetical observed distributions under treatment and control, and the relationship to the observable w. The response of the outcome y (future output) depends positively on the observable, as assumed in the example. Treatment also depends positively on the observable, and treatment adversely affects the outcome. Points 1 and 2 depict the unweighted treatment and control means. Points 3 and 4 depict the re-weighted treatment and control means. The inverse probability weights are illustrated by the unshaded ellipse. The naive unweighted ATE estimator, the difference in means, is $y_1 - y_2$. The IPW estimator is $y_3 - y_4$.

This is depicted in Figure 2 by the difference between the teardrop shapes (showing the lack of covariate balance in the observed data) and the ellipses (the reweighted data display a similarly balanced set of covariate distributions). Notice that in that figure, the sample means calculated with the weighting (points 3 and 4) suggest that consolidations tend to have a negative effect on the outcome, not positive as one would infer from the unweighted means (points 1 and 2). In the statins example, some individuals with high blood pressure do not take statins while others with relatively low blood pressure do take them. Out of the two populations, these individuals' intake of statins is similar to that in a randomized trial and provide the best basis to evaluate the effect of statins on cardiovascular disease.

Inverse weighting is clearly not the only possible solution to the problem of endogenous treatment. The previous section suggests that the dates of fiscal consolidations identified by GLP are predictable by observable characteristics of economies prior to consolidation. Hernández De Cos and Moral-Benito (2012) 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.

9 Estimator of Average Treatment Effects

Absent credible instruments, what is the best empirical way forward? 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 new LP-IPWRA method is thus well established. Here we present the mechanics of the approach. The estimated path can be seen as an analog to the well worn notion of the structural impulse response. Or, as here, it can be used as another way of estimating a set of local projections at different horizons. The probability of fiscal consolidation is the key policy intervention we want to investigate.

We use the same notation that we introduced in section 4, thus referring to outcomes as y_t, the policy variable as D_t, which now is allowed to take only a two discrete values d_1 and d_0, and the vector w_t, which collects all information on predetermined outcomes and controls relevant in explaining the policy variable $D_t = D(w_t, \psi, \varepsilon_t)$. We 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:

$$(y^{\psi}_{t,h}(d_j) - y_t) \perp D_t | w_t; \psi \quad \text{for all } h \geq 0, \text{for } d_1, d_0 \text{ and for all} \psi \in \Psi.$$

The average policy response, conditional on w_t in terms of observable data, is:

$$\begin{gathered} E\left[(y_{t,h}(d_1) - y_t) - (y_{t,h}(d_0) - y_t)|w_t\right] = \\ E\left[y_{t,h} - y_t | D_t = d_1; w_t\right] - E\left[y_{t,h} - y_t | D_t = d_0; w_t\right], \qquad \text{for all } h = 0, ..., H, \end{gathered} \tag{5}$$

where it is assumed that the policy environment characterized by $\psi \in \Psi$ remains constant. Recall that d_1 refers to the policy intervention and d_0 refers to a policy benchmark (for example, no consolidation). 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 acts as a dimension-reduction device and is meant to ensure the estimation of the policy response (the average treatment effect in the microeconomics parlance) is consistent under the main assumption. 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 8 can be seen as candidate estimates of this policy propensity score. We will instead construct the policy propensity score using a richer specification that includes all the controls used in Table 7.

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

$$E\left[(y_{t,h} - y_t)\mathbf{1}\{D_t = d_j\}|w_t\right] = E[(y_{t,h}(d_j) - y_t)|w_t]p^j(w_t, \psi) \qquad \text{for } j = 1, 0. \tag{6}$$

Solving for $E[(y_{t,h}(d_j) - y_t)|w_t]$ and integrating over w_t then the policy response parameter of interest can be calculated as

$$\begin{aligned} \theta^h = E\left[(y_{t,h}(d_1) - y_t) - (y_{t,h}(d_0) - y_t)\right] \quad &= \quad E\left[(y_{t,h} - y_t)\left(\frac{\mathbf{1}\{D_t = d_1\}}{p^1(w_t, \psi)} - \frac{\mathbf{1}\{D_t = d_0\}}{p^0(w_t, \psi)}\right)\right] \\ \text{for } d_1, d_0 \text{ and all } h = 0, ..., H. \end{aligned} \tag{7}$$

Under standard regularity conditions detailed in AJK, an estimate of expression (7) can be obtained using the equivalent sample moment:

$$\begin{aligned} \widehat{\theta}^h \quad &= \quad \frac{1}{T}\sum_{t=h+1}^{T} \frac{(y_{t+h} - y_t)\mathbf{1}\{D_t = d_1\}}{\widehat{p}^1(w_t, \psi)} - \frac{1}{T}\sum_{t=l+1}^{T} \frac{(y_{t+h} - y_t)\mathbf{1}\{D_t = d_0\}}{\widehat{p}^0(w_t, \psi)}; \\ &\text{for all } h = 0, ..., H. \end{aligned} \tag{8}$$

That is, using a first-stage estimate of the policy propensity score, $\widehat{p}^1(w_t, \psi)$, then the dynamic response of the outcome variable $y_{t+h} - y_t$ observed h periods after intervention d_1, which took place at time t, can be simply computed as a weighted sample average as shown in expression (8). AJK collect the $\widehat{\theta}_h$ into a large vector for all $h = 0, ..., H$ and estimate all the relevant responses jointly using a minimum distance expression that facilitates derivation of the appropriate standard errors given first stage estimation uncertainty of the propensity score.

The theory of Robins, Rotnitzky and Zhao (1994) suggests that IPW semi-parametric regression adjusted estimation is the most efficient within its class. Hence, consider the natural extension of the ATE estimator in expression (8) and the LP framework in expression (4). Begin by defining the following auxiliary weighting variables:

$$\widehat{\delta}_t = \left(\frac{\mathbf{1}\{D_t = d_1\}}{\widehat{p}^1(w_t, \psi)} - \frac{\mathbf{1}\{D_t = d_0\}}{\widehat{p}^0(w_t, \psi)}\right), \tag{9}$$

$$\widehat{\phi}_t = \left(\frac{\mathbf{1}\{D_t = d_1\} - \widehat{p}^1(w_t, \psi)}{\widehat{p}^1(w_t, \psi)} - \frac{\mathbf{1}\{D_t = d_0\} - \widehat{p}^0(w_t, \psi)}{\widehat{p}^0(w_t, \psi)}\right). \tag{10}$$

Then, expression (8) can be rewritten as:

$$\widehat{\theta}^h = \sum_{t=h+1}^{T}\left[(y_{t+h} - y_t)\widehat{\delta}_t\right].$$

Instead, the LP-IPWRA estimator can be obtained from:

$$\widehat{\theta}^h = \sum_{t=h+1}^{T} \left[(y_{t+h} - y_t)\widehat{\delta}_t - \widehat{\phi}_t m(w_t, \gamma^h) \right], \tag{11}$$

where $m(w_t, \gamma^h)$ in general is a semi-parametric estimator of the conditional mean model but in our analysis is simply a local projection such as that given by expression (4).

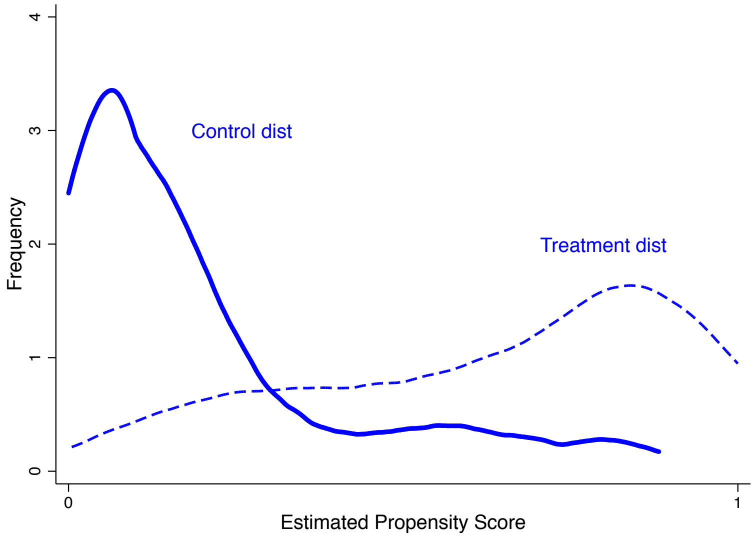
This LP-IPWRA specification is equivalent to the LP-OLS specification based on weights $\widehat{\delta}_t$ and $\widehat{\phi}_t$. As mentioned earlier, the estimator in (11) falls into the class of *doubly robust* estimators (see, e.g. Imbens, 1994; Lunceford and Davidian, 2004; and Kreif et al. 2011). The general intuition behind the estimator is to use the regression model as a way to "predict" the unobserved potential outcomes. The reweighing $\widehat{\phi}_t$ ensures that these predictions are appropriately weighted to match the size of the two subpopulations. When the conditional mean is correctly specified, the regression term in expression (11) improves efficiency. When it is not, the estimator will still be consistent if the propensity score model is correctly specified. Vice versa, if the propensity score model is misspecified but the regression model is correct, the estimate of the ATE will still be consistent. Hence the "doubly robust" label attached to such methods.

10 Contractionary Austerity Revisited: Estimates of the Average Effect of Fiscal Consolidations

This section presents LP-IPWRA estimates of the ATE of fiscal consolidations. We allow for endogenous treatment and allocation bias using inverse probability weighting. The propensity score is based on a probit model that extends the set of controls used in Table 8 with the current and lagged values of the controls in Table 7. The 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 3 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 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 helps properly identify the ATE. However, the figure also indicates that there are some observations likely to get very high weights. Specifically, there are control units whose propensity score is near zero and hence get weights in excess of 10. As a robustness check, we follow the literature (see e.g. Cole and Hernàn, 2008) in truncating the maximum weights in the IPW to 10. We report these results below.

Using the more saturated probit, we then estimate cumulated responses and their sum to the 4 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,

Figure 3: Overlap Check: Empirical Distribution of 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.

coincidentally, the average treatment size (or dose) is equal to 1 percent of GDP in these data (the exact value is 0.97, with a standard deviation of 0.90, with no statistically significant difference between booms and slumps), 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 begin by discussing Table 9, which is the counterpart to Tables 1 and 3. That is, we estimate average treatment effects of fiscal consolidation using the IMF narrative binary indicator as our treatment variable, using the full sample (no binning) and using propensity score estimates based on the saturated probit described earlier. Both the probit and the ATE include country-fixed effects. Table 9 is organized into two rows. The first row reports the raw results whereas the second row reports the results when the inverse probability weights are capped at a maximum value of 10 (a robustness check). The results that are very similar to those reported in Table 3; our LP-IPWRA estimate is a sum effect of -2.08^* over 4 years, that is, a real GDP decline of about 0.52% per 1% fiscal consolidation.

Next we explore the same partition of the data into bins according to whether output is above or below trend that we have used in earlier sections to provide a more granular view of these results. Table 10 presents LP-IPWRA estimates with truncated weights for robustness, based on the same saturated policy propensity score probit model just described above. The results show that Treatment has negative but marginally significant effects in the boom bin, and has significant negative effects in the slump bin. This evidence adds more caution to our earlier results. Summed over 4 years, the LP-IPWRA effects are $-1.13+$ in the boom, and -2.48^* in the slump.

Table 11 sums up based on the cumulative effects over 4 years estimated with our three methods. Using OLS we would walk away believing in expansionary austerity, although with no

Table 9: Average treatment effect of fiscal consolidation, LP-IPWRA estimates

Log real GDP (relative to Year 0, ×100)

	(1) Year 1	(2) Year 2	(3) Year 3	(4) Year 4	(5) Sum
Fiscal ATE, unrestricted weights	-0.21+	-0.62*	-0.63*	-0.74*	-2.08*
	(0.12)	(0.16)	(0.18)	(0.21)	(0.52)
Fiscal ATE, truncated weights	-0.17	-0.55*	-0.61*	-0.88*	-2.08*
	(0.14)	(0.19)	(0.18)	(0.25)	(0.52)
Observations	457	440	423	406	406

Standard errors (robust sandwich) in parentheses. $+p < 0.10$, $*p < 0.05$.
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 $\lambda = 100$.
Specification includes country fixed effects in the propensity score and in the ATE calculation. "Unrestricted weights" refers to LP-IPWRA estimates that impose no restrictions on the weights of the propensity score. "Truncated weights" refers to the LP-IPWRA estimates where the maximum weight is truncated at 10.

Table 10: Average treatment effect of fiscal consolidation, LP-IPWRA estimates, booms v. slumps

Log real GDP (relative to Year 0, ×100). Truncated weights for robustness.

	(1) Year 1	(2) Year 2	(3) Year 3	(4) Year 4	(5) Sum
Fiscal ATE, $y^C > 0$, boom	-0.15	-0.42+	-0.20	-0.63*	-1.13+
	(0.14)	(0.22)	(0.25)	(0.28)	(0.69)
Fiscal ATE, $y^C < 0$, slump	-0.19	-0.69*	-0.89*	-0.60*	-2.48*
	(0.14)	(0.17)	(0.19)	(0.25)	(0.58)
Observations	457	440	423	406	406

Standard errors (robust sandwich) in parentheses. $+p < 0.10$, $*p < 0.05$.
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 $\lambda = 100$.
Specification includes country fixed effects in the propensity score and in the ATE calculation.
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.

effect when the economy is in a slump. Using the "narrative" IV we would walk away believing in contractionary austerity, but only when the economy is in a slump. Using the IPWRA method we find stronger evidence of contractionary austerity in slumps, where our estimates are more precisely estimated, but also to some extent even in booms, where the point estimates are larger and verging on statistical significance.

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. Of course, history suggests that the presence of such wisdom and institutions cannot be assumed; unfortunate as this may be in most respects, we can see how misguided experiments can at least provide empiricists like us with helpful identification.

Table 11: Summary of Results: Expansionary or Contractionary Austerity?

Estimation method	LP-OLS	LP-IV	LP-IPWRA
Fiscal variable	AA Change in CAPB continuous	AA Change in CAPB continuous	IMF GLP consolidation indicator binary (ATE)
Control for endogeneity	none	instrumented by IMF consolidation indicator	inverse propensity score for IMF indicator & regression adjustment (IPWRA)
Booms	mildly expansionary*	no effect	mildly contractionary
Slumps	no effect	contractionary*	contractionary*

11 Counterfactual: Coalition Austerity and The U.K. Recession

To illustrate the usefulness of our modeling approach and provide a quantitative illustration, we apply our LP-IPWRA estimates to make an (out-of-sample) counterfactual forecast of the post-2007 recession/recovery path of the U.K. economy with and without the fiscal austerity policies imposed by the Coalition government after the 2010 election.

That the U.K.'s economic performance in this period was subpar and below ex ante expectation is now well known, and yet is also a matter for contentious debate. It has been a much weaker recovery than in the U.S., where no double dip took place, and 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, explanations for the differences have tended to look in other areas. Various explanations have been offered, including not just tighter U.K. fiscal policy, but also spillovers from the Eurozone, which has a large impact given the U.K.'s strong trade orientation with the Continent and weak trade links with emerging markets, as compared to the U.S. (Davies 2012). Other stories have invoked contractions in oversized U.K. sectors such as finance and North Sea oil and gas.

To gain quantitative traction on the share of responsibility that should be borne by fiscal policy we use our LP-IPWRA 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 +2.5% of GDP from 2010 to 2011, followed by +1.3% from 2011 to 2012 and a forecast +1.7% from 2012 to 2013. This gives us a sequence of three fiscal policy shocks.[10] Note that, as above, the average treatment in the low bin is 1/0.97 (weighted), so for this counterfactual exercise we scale treatment effects due to each shock by a factor of 1/0.97.

We then have to compute the impact of each shock at each horizon and make sure we assign it appropriately. Of course, our LP 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

[10] Two alternative measures of fiscal shocks are discussed in an appendix, and Table A6, one from the OBR and one from the IMF. The measure we have chosen is the official UK measure and is bounded by these alternative measures.

in year 1 and beyond.

The effects of the first round of austerity in 2010–11 can be computed directly from the LP-IPWRA 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 2.5 (OBR data, as above) multiplied by the scaling factor of $1/0.97$ (noted above), and then multiplied by the LP-IPWRA 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.

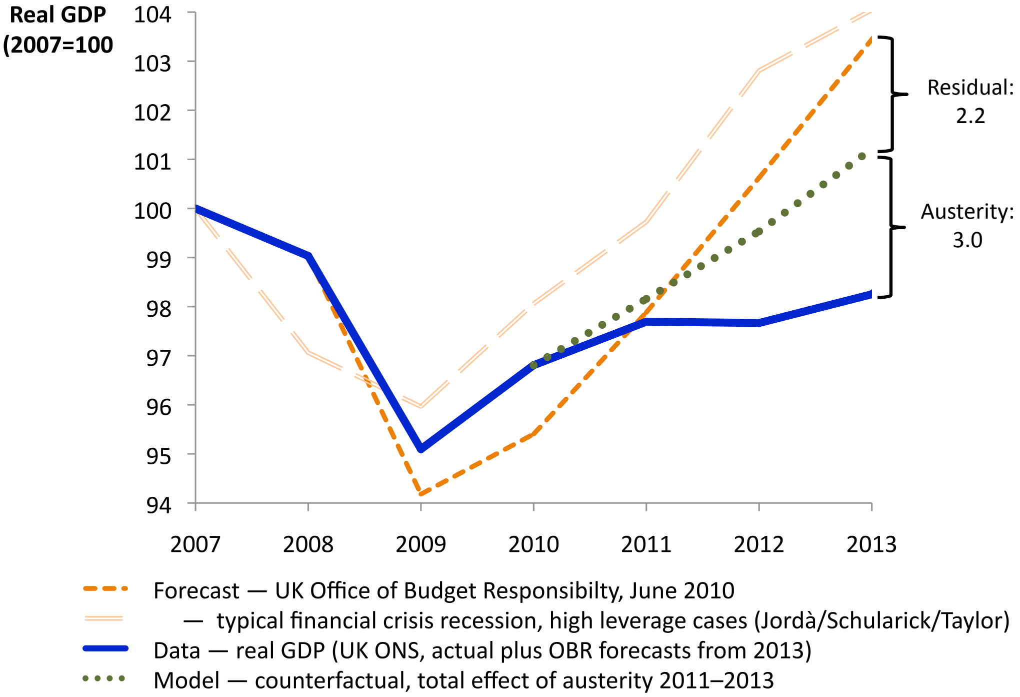
First, there is a similar direct effect of the first year shock on second year output: the first year shock magnitude of 2.5 (again) multiplied by the scaling factor of $1/0.97$ (again), and then multiplied by the LP-IPWRA of -0.69 (from the slump bin, but now in year 2). Second, there is the additional 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. These estimations for the necessary 3-year horizon are reported in the appendix in Table A5. From there 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 0.53. The model thus says that once policymakers did austerity in year 0 there was a 53% chance they would do it again in year 1; the model also gives a 25% chance in year 2 and 11% in year 3. For our model this means that 53% of the coalition austerity in 2011–12 (and 25% in 2012–13) was "baked in" — in probabilistic terms — by the decision to do austerity in 2010–11. So this is already accounted for in the LP-IPWRA output path estimates, and must be subtracted to get the correct estimate of the marginal impact of the actual 2011–12 austerity decision outcome.

The net effects can be computed mechanically as we illustrate in the following example. First, we can compute the first year shock magnitude of 2.5 (again) multiplied by the scaling factor of $1/0.97$ (again), and then multiplied by the LP-IPWRA of -0.69 (from the slump bin in year 2). Second, we can add to this the unexpected second year shock magnitude of 1.3 (OBR) multiplied by the scaling factor of $1/0.97$ (again), multiplied by the LP-IPWRA of -0.19 (from the slump bin in year 1), and multiplied by the probability of no treatment in year 2 which is $0.47 = 1 - 0.53$. In a similar way we can assign unexpected and expected effects of contemporaneous treatment to prior treatment in all years along the path.

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 over the period from 2007 (the business cycle peak) through 2013.

As a starting point, absent any knowledge of what was to happen after the 2010 Coalition austerity program, what might have been the expected path of the U.K. economy? This question is answered by the two dashed lines. The double-long-dashed line shows the unconditional forecast path in a financial crisis recession based on the large sample of all advanced-economy recessions since 1870 in the work of Jordà, Schularick, and Taylor (2011), extended to the 6 year horizon. We restrict attention here to their unconditional forecast path for highly leveraged economies after a financial crisis, a category which includes the U.K. case in 2007. Clearly, a

Figure 4: U.K. Austerity: Forecast, Actual, and Counterfactual Paths for Real GDP (2007=100)



Notes: See text. Units in the chart are expressed in percent of 2007 real GDP, the last peak. The OBR forecast is from `http://budgetresponsibility.independent.gov.uk/wordpress/docs/pre_budget_forecast_140610.pdf`. The Jordà, Schularick, and Taylor (2011) path is for real GDP per capita, extended to a 6-year horizon; it is adjusted up by 0.65% per year to reflect the current average U.K. rate of population growth. Actual data from IMF World Economic Outlook (WEO) database, October 2012, and UK Office of National Statistics (ONS) as of March 2013. The extrapolation of the ONS data for the year 2013 is based on the March 2013 OBR forecast of 0.6% real GDP growth in 2013. Model counterfactuals remove estimated LP-IPWRA responses, suitably scaled, from the actual ONS-OBR path.

seriously painful recession was to be expected anyway: if output is set to 100 in 2007, this path shows a 4% drop over two years, to a level of 96 by 2009, followed by recovery thereafter, with output rising to around 104 in 2013, where the 6 year forecast ends.

What did the authorities expect? According to the June 2010 Pre-Budget report of the OBR they had expected something similar but slightly worse (or slightly lagged) relative to typical historical experience to unfold after 2010, as shown by the short-dashed path in the figure. The bottom in output here is 94.2 and the collapse starts a bit later, reflecting actual data up to 2009, probably because the crisis-recession was only just starting in late 2007, and full blown financial panic did not start until after the Lehman collapse in September 2008. This could explain why actual 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. (i.e., 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/OBR) data to depict the outturn of events (actual plus latest 2013 forecast, as shown). Everything was going more or less according to the above two forecast paths until 2010. After that, a double-dip recession took hold and the U.K. real economy virtually flatlined for three straight years. (In per capita terms, it actually sank.)

How much of the dismal performance can be attributed to the fiscal policy choice of instigating austerity during a slump? The answer, using our model as described above, is: about three fifths. This is shown by the dotted line in the chart, which cumulates the effects of each of the three years of austerity on contemporaneous and future years' growth from 2010 to 2013. By 2013, the last year in the window, the cumulative effects of these choices amounted to about 3.0% of GDP (in 2007 units) where the total gap relative to the OBR's June 2010 forecast was 5.2%, thus leaving an unexplained residual of 2.2% relative to OBR's forecast. Our model also suggests that additional drag from the 2010–12 policies will also continue to be felt into 2014–16, even not allowing for any further austerity.

The residual relative to the forecasts in Figure 4 could be accounted for by various omitted factors, some as noted: export patterns and the Eurozone, idiosyncratic U.K. sector shocks, or overoptimism in the 2010 forecast. For example, Schularick and Taylor (2012) argue that the banking and shadow banking system in the U.K. (as in the U.S.) created unusually large credit overhang and hence a further drag on the recovery, compared to historical norms. However, one major caveat suggests that we also likely have a biased underestimate of the effects of U.K. fiscal policy. This caveat is the zero lower bound (ZLB) of monetary policy: the U.K. out-of-sample counterfactual corresponds to a liquidity trap environment, but the in-sample data overwhelmingly do not. In that case, the true residuals in 2010–13 could be much smaller than above, and the effects of austerity (i.e., the fiscal multipliers) even bigger, big enough to possibly explain most or all of the growth shortfall after 2010. Why?

Our estimates were based on empirical work using a sample from the 1970s to 2007, when the ZLB was virtually absent from any country-year observations during consolidation periods in our IMF dataset (the only exceptions being 7 country-year observations, out of a total of 173 consolidation episodes, all of these relating to Japan in the 1990–2007 period). As is well known in theory (Christiano et al. 2011; Eggertsson and Krugman 2012; Rendahl 2012) and also from historical evidence from the Great Depression (Almunia et al. 2010), fiscal multipliers are much larger in such conditions than in normal times when monetary policy is away from this constraint. In theory we could try to estimate this difference using stratification on the interest rate, but our in-sample estimation cannot hope to convincingly capture the ZLB effect with just a handful of observations from Japan, and so this must remain a goal for future research where we hope to apply our new estimation methods to a larger set of both contemporary and historical data. But in the post-2008 forecast period for the U.K. the ZLB was a binding constraint which would tend to make even our already large estimated fiscal impacts an underestimate of the true impacts. If so, then the vast majority of the difference between the actual U.K. recovery and the ex ante forecast (or the typical historical path) could be attributed to the Coalition's austerity policy choices in 2010–13.

12 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 estimates, 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 values of the fiscal multiplier often found in this strand of the literature are indeed close enough to the truth.

If episodes of fiscal consolidations could be separated by whether or not they can be explained by the 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, the results from this approach seem to 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 fiscal 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 instrumental variable methods. But this turns out not to be the case. So, while imperfect, the narrative approach (through these instrumental variable 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, to the nihilistic conclusion that observational data are hopelessly unsuitable for the problem at hand, but we believe the battle is not lost. Matching methods, common in biostatistics and in medical research, as well as in applied microeconomics when ideal randomized trials are unavailable, offer a last line of defense. Recent work by Angrist, Jordà and Kuersteiner (2013) introduce 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. By narrowing the focus on those observations where allocation to treatment is “as if” at random (on the basis of the predictive first stage model), we have a measure of the multiplier that explicitly accounts for failures of identification due to observable controls.

Our estimates are quantitatively closer to those from the instrumental variables specification than to those from the least squares specification. In fact, they suggest even larger impacts than the IMF study when the state of the economy worsens. 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*, edited by Jeffrey R. Brown, vol. 24. 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 Before and Since the Great Recession: String Theory Revisited. Paper presented at the NBER Summer Institute.

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: 1047–74.

Auerbach, Alan J, and Yuriy Gorodnichenko. 2012. Measuring the Output Responses to Fiscal Policy. *American Economic Journal: Economic Policy* 4: 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. 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 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.

Guajardo, Jaime, Daniel Leigh, and Andrea Pescatori. 2011. Expansionary Austerity: New International Evidence. IMF Working Paper 11/158.

Giavazzi, Francesco, and Marco Pagano. 1990. Can Severe Fiscal Contractions Be Expansionary? Tales of Two Small European Countries. *NBER Macroeconomics Annual*, Cambridge, Mass.: MIT Press, pp. 95–122.

Graham, Bryan S., Crisitine Campos De Xavier Pinto and Daniel Egel. 2012. Inverse Probability Tilting for Moment Condition Models with Missing Data. *Review of Economic Studies* 79(3): 1053–79.

Hernández de Cos, Pablo, and Enrique Moral-Benito. 2012. Endogenous Fiscal Consolidations. Banco de Espana Working Paper 1102.

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: 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, Moritz Schularick, and Alan M. Taylor. 2011. When Credit Bites Back: Leverage, Business Cycles, and Crises. NBER Working Paper 17621.

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. 2011. Regression-Adjusted Matching and Double-Robust Methods for Estimating Average Treatment Effects in Health Economic Evaluation. London School of Hygiene and Tropical Medicine. Preprint. `http://www.lshtm.ac.uk/php/hsrp/reducing-selection-bias/output/publications.html`.

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: 2937–60.

Meyer, Bruce D. 1995. Natural and Quasi-Experiments in Economics. *Journal of Business and Economic Statistics* 13(2): 151–61.

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.

Nakamura, Emi and Jon Steinsson. 2011. Fiscal Stimulus in a Monetary Union: Evidence from U.S. Regions. NBER Working Paper 17391.

Owyang, Michael T., Valerie A. Ramey, and Sarah Zubairy. 2013. Are Government Spending Multipliers Greater During Periods of Slack? Evidence from 20th Century Historical Data. NBER Working Paper 18769.

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.

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

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. 1999. Robust Estimation in Sequentially Ignorable Missing Data and Causal Inference Models. *Proceedings of the American Statistical Association Section on Bayesian Statistical Science*, 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.

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`.

Appendix

LP-OLS with Country-Fixed Effects and Controlling for World Growth

This section reports estimates of the LP-OLS specification (equation (4) 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, ×100)

	(1)	(2)	(3)	(4)	(5)
	Year 1	Year 2	Year 3	Year 4	Sum
(a) Uniform effect of d.CAPB changes					
Fiscal multiplier, $y^C > 0$, boom	0.21*	0.25*	0.06	-0.18^{+}	0.24
	(0.07)	(0.07)	(0.05)	(0.10)	(0.16)
Observations	222	205	192	180	180
Fiscal multiplier, $y^C \leq 0$, slump	-0.03	-0.06	-0.17	-0.23^{+}	-0.55*
	(0.03)	(0.06)	(0.10)	(0.11)	(0.26)
Observations	235	235	231	226	226
(b) Separate effects of d.CAPB for Large (> 1.5%) and Small (≤ 1.5%) changes					
Fiscal multiplier, $y^C > 0$, boom, large consolidation	0.23*	0.25*	0.07	-0.17^{+}	0.33
	(0.08)	(0.08)	(0.05)	(0.09)	(0.20)
Fiscal multiplier, $y^C > 0$, boom, small consolidation	0.04	0.19	-0.02	-0.35	-0.60
	(0.11)	(0.32)	(0.38)	(0.35)	(0.99)
Observations	222	205	192	180	180
Fiscal multiplier, $y^C \leq 0$, slump, large consolidation	-0.03	-0.05	-0.18	-0.30^{+}	-0.62^{+}
	(0.04)	(0.07)	(0.12)	(0.15)	(0.35)
Fiscal multiplier, $y^C \leq 0$, slump, small consolidation	-0.05	-0.15	-0.10	0.13	-0.17
	(0.12)	(0.20)	(0.23)	(0.31)	(0.69)
Observations	235	235	231	226	226

Standard errors (clustered by country) in parentheses. $+p < 0.10$, $*p < 0.05$.
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 $\lambda = 100$.

LP-IV with Country-Fixed Effects and Controlling for World Growth

The following Tables A2 and A3 correspond to Tables 4 and 5 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, ×100)

	(1) Year 1	(2) Year 2	(3) Year 3	(4) Year 4	(5) Sum
Fiscal multiplier, $y^C > 0$, boom	-0.32	-0.33	-0.14	-0.54	-0.85
	(0.29)	(0.48)	(0.47)	(0.41)	(1.25)
Observations	222	205	192	180	180
Fiscal multiplier, $y^C \leq 0$, slump	-0.24$^+$	-0.76*	-0.95*	-0.79*	-2.69*
	(0.14)	(0.23)	(0.28)	(0.30)	(0.81)
Observations	235	235	231	226	226

Standard errors (clustered by country) in parentheses. $+p < 0.10$, $*p < 0.05$.
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).

Table A3. Fiscal multiplier, d.CAPB, IV estimate (continuous), booms v. slumps
Log real GDP (relative to Year 0, ×100)

	(1) Year 1	(2) Year 2	(3) Year 3	(4) Year 4	(5) Sum
Fiscal multiplier, $y^C > 0$, boom	-0.34	-0.39	-0.08	-0.47	-1.18
	(0.25)	(0.33)	(0.36)	(0.47)	(1.22)
Observations	222	205	192	180	180
Fiscal multiplier, $y^C \leq 0$, slump	-0.34*	-0.69*	-0.72*	-0.55$^+$	-2.27*
	(0.16)	(0.27)	(0.34)	(0.29)	(0.96)
Observations	235	235	231	226	226

Standard errors (clustered by country) in parentheses. $+p < 0.10$, $*p < 0.05$.
Additional controls: cyclical component of y, 2 lags of change in y, country fixed effects and change from period 0 to period h in world real GDP (World Bank), or its sum.
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 continuous form (total, in % of GDP).

Robustness

As discussed in the text, we explored the sensitivity of our results to different model specifications. These findings are shown in Table A4. In each case we show in Table A4 the impacts that these model changes have on the estimated 4-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 other robustness checks: when we truncate the inverse probability weights at values of 5, and when we do not truncate at all.

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. Under stricter truncation the adverse impacts get even larger.

Table A4. ATE of fiscal consolidation, LP-IPWRA estimates, booms v. slumps, various p-score models
Sum of log real GDP impacts, years 1 to 4 (all relative to Year 0, ×100)

	(1)	(2)	(3)	(4)	(5)
Estimator	probit	logit	logit	Truncate	Not
	CFE +	CFE	CFE +	$ipw < 5$	Truncated
	world GDP		world GDP		
Fiscal ATE, $y^C > 0$, boom	-1.20	-1.13	-1.20	-1.17	-1.16
	(0.72)	(0.70)	(0.73)	(0.69)	(0.68)
Fiscal ATE, $y^C \leq 0$, slump	-2.49***	-2.48***	-2.49***	-2.53***	-2.44***
	(0.56)	(0.58)	(0.56)	(0.58)	(0.58)
First-stage, AUC	0.87	0.87	0.87	0.86	0.87
s.e.	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
Observations	406	406	406	406	406

Standard errors (clustered by country) in parentheses. $+p < 0.10$, $*p < 0.05$.
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 inverse probability weights truncated to (0,5], 13 observations truncated.
Column 5: As in Table 10, pooled probit, but inverse probability weights not truncated.

Estimated LP Equation for Future Treatment

As discussed in the text, in 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 A5. Treatment is persistent, with but a slight difference between the boom and slump bins.

Table A5. LP estimate of impact treatment on future treatment, LP-OLS estimates, booms v. slumps
Dependent variable: Treatment in year h (consolidation from year h to $h+1$)

	(1)	(2)	(3)
	Treatment $(t+1)$	Treatment $(t+2)$	Treatment $(t+4)$
Treatment (t), boom	0.481	0.334	0.260
	(0.082)	(0.086)	(0.091)
Treatment (t), slump	0.527	0.245	0.113
	(0.057)	(0.054)	(0.038)
Observations	439	421	404

Standard errors (clustered by country) in parentheses. $+p < 0.10$, $*p < 0.05$.
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 $\lambda = 100$.

Measures of U.K. Fiscal Consolidation 2010–13

Measures of the Size of U.K. Fiscal Treatments are shown in shown in Table A6. 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 and the conclusions would be little affected by the use of the alternate measures. In size terms, our preferred OBR cyclically-adjusted primary balance series (b) is larger than (c) but smaller than (a) over the three years.

Table A6. 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)
Estimator	2010	2011	2012	2013
(a) OBR, cyclically-adjusted primary balance (used in text)	−7.0	−4.5	−3.2	−1.5
change	—	+2.5	+1.3	+1.7
cumulative change	—	+2.5	+3.8	+5.5
(b) OBR, cyclically-adjusted Treaty deficit, 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
(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

IMF years as in column heads; OBR years are year shown to following year (e.g., 2010 denotes 2010–11, etc.) Data from IMF WEO October 2012 database, HM Treasury Autumn Statements 2011 and 2012, and HM Treasury and OBR Budget 2013 documents online.