CESifo WORKING PAPERS

10485
2023

June 2023

Misallocation in Firm Production: A Nonparametric Analysis Using Procurement Lotteries

Paul Carrillo, Dave Donaldson, Dina Pomeranz, Monica Singhal

CESifo

Impressum:

CESifo Working Papers
ISSN 2364-1428 (electronic version)
Publisher and distributor: Munich Society for the Promotion of Economic Research - CESifo GmbH
The international platform of Ludwigs-Maximilians University's Center for Economic Studies and the ifo Institute
Poschingerstr. 5, 81679 Munich, Germany
Telephone +49 (0)89 2180-2740, Telefax +49 (0)89 2180-17845, email office@cesifo.de
Editor: Clemens Fuest
https://www.cesifo.org/en/wp
An electronic version of the paper may be downloaded
· from the SSRN website: www.SSRN.com
· from the RePEc website: www.RePEc.org
· from the CESifo website: https://www.cesifo.org/en/wp

CESifo *Working Paper No. 10485*

Misallocation in Firm Production: A Nonparametric Analysis Using Procurement Lotteries

Abstract

How costly is the misallocation of production that we might expect to result from distortions such as market power, incomplete contracts, taxes, regulations, or corruption? This paper develops new tools for the study of misallocation that place minimal assumptions on firms' underlying technologies and behavior. We show how features of the distribution of marginal products can be identified from exogenous variation in firms' input use, and how these features can be used both to test for misallocation and to quantify the welfare losses that it causes. We then consider an application in which thousands of firms experience demand shocks derived from a lottery-based assignment of public procurement contracts for construction services in Ecuador. Using administrative tax data about these firms, we reject the null of efficiency but estimate that the welfare losses resulting from misallocation are only 1.6% relative to the first-best. Standard parametric assumptions applied to the same setting would suggest losses that are at least an order of magnitude larger.

JEL-Codes: D240, D610, H570, L100, O400.

Keywords: allocative efficiency, misallocation, aggregate productivity.

Paul Carrillo
George Washington University / USA
pcarrill@gwu.edu

Dave Donaldson
MIT / Cambridge / MA / USA
ddonald@mit.edu

Dina Pomeranz
University of Zurich / Switzerland
dina.pomeranz@uzh.ch

Monica Singhal
University of California, Davis / USA
msinghal@ucdavis.edu

May 2023

We thank our team of outstanding, dedicated research assistants, as well as David Baqaee, Richard Blundell, Kirill Borusyak, Arnaud Costinot, Pete Klenow, Jeremy Majerovitz, Matt Masten, Virgiliu Midrigan, Emi Nakamura, Whitney Newey, Michael Peters, Richard Rogerson, Cian Ruane, Jesse Shapiro, John Sturm, Alex Torgovitsky, Ivan Werning, and numerous seminar participants for helpful comments. We are grateful to the Centro de Estudios Fiscales and the Departamento de Control of the Ecuadorian Tax Authority for outstanding collaboration and to Innovations for Poverty Action (IPA), International Growth Centre (IGC) and the Stanford Institute for Innovation in Development Economies (SEED) for generous research support. This paper has also benefited from funding from CEPR and UK Department for International Development (DFID) (under the Private Enterprise Development in Low-Income Countries program, reference MRG004 3834), the European Research Council (grant reference 758984), and the Swiss National Science Foundation (grant 100018 192588). The views expressed are not necessarily those of CEPR, DFID, ERC or SNF.

1 Introduction

There is no shortage of reasons to suspect that the economies around us produce goods and services in an inefficient manner. Many firms seem to enjoy market power, many contracts look incomplete, and many policy actions (be they taxes and subsidies, regulations, or even corruption) appear willing to sacrifice production efficiency in pursuit of other goals. But just as market failures may often be qualitatively glaring, the quantification of distortions and their misallocative costs is often challenging. And this is made all the more difficult when, as the theory of the second best reminds us, the arrival of any additional distortion may actually mitigate misallocation rather than exacerbate it.

In this paper we develop and apply new techniques for assessing the extent of misallocation among any given set of firms. Our techniques leverage exogenous shocks to firms' input use in order to quantify features of the distribution of firms' marginal products, for each input, which allows us to test for the presence of misallocation and to estimate its welfare costs. Such exogenous variation can stem from either output demand or input supply. We apply these new procedures to Ecuador's construction industry, in which a lottery component of the country's public procurement system generates random sources of firm-level demand. Using such variation, we find that misallocation appears to be limited in this context.

To arrive at this conclusion, we begin in Section 2 by describing the economic environment that motivates our empirical procedures. An econometrician observes a set of potentially multi-product firms producing in a given initial cross-section. The goal is to assess whether there is misallocation in this cross-section—or equivalently, whether it displays allocative efficiency of production (AEP). To ease comparisons with prior work, we distinguish two notions of AEP. The first, which we term unconstrained AEP (U-AEP), is generically necessary for Pareto efficiency. Such an allocation requires that, for any firm and produced good or service, the ratio of the value marginal product of any input to the input's price—a ratio we refer to as the "wedge" for that firm-product-input—is equal to one no matter where the input is in use. The second notion is analogous but conditions on the aggregate amount of each input type that can be used by the firms in question, as standard definitions of aggregate productivity do. An allocation consistent with this notion of constrained AEP

(C-AEP) would feature wedges that are equal to a common value across firms and products for each input type, but where that common value is not necessarily one.

Testing for either constrained or unconstrained AEP would be simple if marginal products, and hence wedges, were observed. But a firm's marginal products, by definition, depend on its production function, which cannot be identified in the cross-section if firms' technologies differ in unknown respects. Existing methods pool the data across firms, under the assumption that they have common production functions (up to a low-dimensional parameter). But doing so may inherently increase the scope for apparent heterogeneity in wedges, rather than in technology, to explain observed differences in firm behavior. These methods may therefore overestimate the extent of misallocation—especially that due to departures from C-AEP, which rests on wedge heterogeneity.

By contrast, the methods we develop proceed by placing minimal restrictions on the technologies that firms use, the demand or supply relations they face, the extent of their optimizing behavior, or the underlying sources of market failure that cause misallocation.[1] This flexibility is possible for two reasons. First, we do not aim to identify each firm's wedge on each input and product. Instead, we focus on features of misallocation—such as its existence or its welfare consequences—that are functions of the wedge *distribution*, rather than each individual wedge itself. Second, we observe that any wedge is simply the appropriately price-adjusted "treatment effect" of a change in a firm's input on its output. This implies that exogenous variation in firm input use can be used to identify features of the distribution of such treatment effects (and hence wedges), drawing on recent advances in the literature on treatment effects estimation due to Masten and Torgovitsky (2016).

Intuitively, if a given allocation features C-AEP, for example, then, starting from this allocation, the treatment effect of an exogenous change in any input will be the same anywhere in this economy. And if it features U-AEP, then these treatment effects will also equal one. Our procedures simply invert this intuition and ask what amount of misallocation is implied by the distribution of treatment effects that arises due to the exogenous input variation that is available to the econometrician.

[1]The main substantive assumption we do require is that each firm uses a technology that is differentiable at the point at which it is operating.

While the techniques we develop can be applied in many settings, as we discuss more below, our application focuses on Ecuador's construction services sector. As detailed in Section 3, a component of that country's public procurement process allocates contracts (for specific projects at specific prices) by lottery. We study the approximately 9,000 firms that took part in over 18,000 multi-participant procurement lotteries that were held in our sample period (2008-2015), as these firms were subject to a component of demand that was determined randomly. We trace the response of both outputs and inputs to these random demand shocks by merging the lottery contracts data with administrative tax records, which contain information on sales and costs, firm-to-firm transaction values, as well as employer-employee matches. Comparing these responses allows us to use the techniques described above in order to both test for U-AEP and C-AEP, and to estimate the cost of misallocation.

Our first set of results, in Section 4, begins with an event-study specification that reveals the average time-path of such output and input responses to a unit of randomly-determined procurement lottery winnings. Firms rapidly scale up sales as well as both labor and non-labor inputs, with effects peaking about 6 months after the lottery and dissipating by 14 months. Cumulative sales responses are larger than cost responses by a factor of 1.15. Turning to the heterogeneity in these responses, we observe little dispersion in either output or input responses when grouping firms according to a variety of pre-determined firm characteristics, such as baseline sales or number of employees, despite the fact that firms are highly heterogeneous along such dimensions. These descriptive findings are therefore consistent with a setting in which wedges are larger than one (in violation of U-AEP) but display limited dispersion across firms (consistent with C-AEP). However, these conclusions may mask firm-level heterogeneity in unobserved respects.

Section 5 therefore turns to a formal test for both forms of AEP that does not rely on any ability to pre-specify the sources of heterogeneity across firms and inputs. We implement these tests via randomization inference to avoid relying on asymptotic inferential assumptions. Our results are consistent with the aforementioned descriptive analysis. The test for U-AEP—a test for the null that all wedges equal one—resoundingly rejects, with a p-value less than 0.001. On the other hand, the test for C-AEP—which asks whether wedges take a common value for each firm, within each input type—does not reject ($p = 0.35$).

The previous results imply that deviations from U-AEP are large enough to detect at standard levels of statistical significance, but they do not tell us how large the cost of misallocation is in economic terms. Equally, the failure to reject C-AEP may reflect a test with low power even if the gain from equalizing wedges might be substantial. In Section 6 we therefore move beyond such tests to estimates of certain weighted moments of the wedge distribution that are motivated by a second-order approximation to the cost of misallocation, following Baqaee and Farhi (2020). In particular, we estimate that the sales-weighted mean of wedges across firms is 1.126 and the corresponding variance of wedges is 0.014. Even at conservative values for elasticities of output demand and input supply, these point estimates imply that the total cost of misallocation in this context is just 1.6%.[2]

This estimated cost of misallocation arises in roughly equal parts from two components. First, the fact that wedges are dispersed across firms implies that there are gains (of 1.0%) from reallocating inputs to higher value marginal product firms, even while holding aggregate inputs constant at their factual levels (consistent with C-AEP). Second, the fact that average wedges exceed one implies that there would be further gains (of 0.6%) from increasing aggregate inputs until the value marginal product of each input equals its price.

It may seem surprising that the estimated cost of misallocation in this context is so much smaller than typical estimates in the literature (discussed below). However, as argued above, by placing only minimal restrictions on firms' technologies, it is natural that our procedure may arrive at a lower estimated amount of wedge dispersion. To shed light on this, we can ask what our data would imply if we were to impose the assumption that all firms have technologies that feature common rates of returns-to-scale, as is common in the existing literature on misallocation, but are nevertheless free to differ in arbitrary ways across firms. Applying this assumption to the same data would lead to the conclusion that the firms in our setting exhibit substantially greater wedge dispersion and a cost of misallocation that is many times larger. For example, assuming constant returns implies that this cost is 66% rather than 1.6% using our procedure. An interpretation of our findings is, therefore, that the firms in our context truly have heterogeneous technologies—featuring even heterogeneous

[2]The block-bootstrapped distribution of such estimates contains a relatively small mass of substantially higher estimates, though 80% of such values are below 7%.

returns returns-to-scale—and that procedures that restrict such heterogeneity by assumption would reach different conclusions about allocative efficiency.

Prior work has been focused on quantifying the allocative efficiency of inputs across firms for some time. For example, Restuccia and Rogerson (2008) and Hsieh and Klenow (2009) offer seminal contributions and Hopenhayn (2014) and Restuccia and Rogerson (2017) provide reviews. As discussed above, the typical approach in existing work has been to leverage parametric assumptions about firms' production functions in order to estimate marginal products, and hence wedges, in any given cross-section. Such an approach is invaluable if one hopes to identify a separate wedge for every firm and input. However, as discussed above, this procedure may lead to biased wedge estimates if the specification of firms' technologies is incorrect (for example, in terms of assumed returns-to-scale as in Haltiwanger et al. (2018)) or understates cross-firm technological heterogeneity (as emphasized in Gollin and Udry (2021), for example). A distinguishing feature of our approach is that, instead of seeking to identify each wedge, we focus on identifying features of the wedge distribution that are sufficient for our questions of interest: whether all wedges equal one (in order to test for U-AEP), whether the wedge distribution is degenerate for each input type (to test for C-AEP), and how large are various weighted first and second moments of the wedge distribution (to quantify the cost of misallocation).

To do so, we leverage insights from the Hall (1988, 2018) method of estimating markups. Hall (2018), for example, uses a time-series regression of a U.S. sector's output on a weighted bundle of its inputs, with instrumental variables (IVs) based on military purchases and oil price fluctuations, to identify the sector's markup under the assumptions that the markup is constant over time within the panel dataset. We extend this idea in several directions. First, our techniques estimate weighted moments of the wedge distribution within a single cross-section of firms, and hence require no restrictions on markup changes over time.[3] Second, we provide a method for translating such weighted moments into estimates of the cost of

[3]As discussed below, the Masten and Torgovitsky (2016) estimator is suitable for models like ours, in which coefficients are arbitrarily heterogeneous and regressors are endogenous. This estimator has been previously used by Gollin and Udry (2021) when estimating averages of heterogeneous Cobb-Douglas production function parameters in a model of Tanzanian and Ugandan farms. Klette (1999) also uses a random coefficient estimator to estimate unweighted markup dispersion across firms and years in a panel of Norwegian firms, but one assuming that regressors are exogenous.

misallocation. And third, consistent with the misallocation literature, we allow for arbitrary wedges on each input rather than simply an overall markup.

Our procedures for assessing misallocation rely on the availability of variation in firm-level input changes that is exogenous with respect to the firm's production set. Recent work on firms has exploited experimental or quasi-random variation to isolate exactly such variation, stemming from either the output demand side or the input supply side. In terms of the former, the pioneering study of Ferraz et al. (2015) exploits quasi-random variation arising from Brazil's procurement auction system to study effects of demand shocks on firm growth; similarly, Fadic (2020) examines the Ecuadorian lottery system studied here and documents temporary effects on revenues, labor payments and assets of winning firms.[4] Other examples of exogenous sources of demand that have been isolated in prior work derive from market integration (Jensen and Miller, 2018), firm entry (Bergquist and Dinerstein, 2020; Busso and Galiani, 2019), and import competition (Felix, 2021). Turning to shocks from the input supply side, innovative studies have used variation in firms' access to capital (de Mel et al., 2008; Banerjee and Duflo, 2014; Kaboski and Townsend, 2011), labor (de Mel et al., 2016; Beerli et al., 2021), and management consulting services (Bloom et al., 2013; Giorcelli, 2019). Our paper complements and contributes to these literatures by formalizing a method through which firms' responses to such exogenous shocks can be used to test for the existence of allocative inefficiency and to quantify its magnitude.[5]

2 Theoretical Framework

This section describes the theoretical environment in which we aim to test for allocative efficiency and to measure the welfare consequences of departures from such efficiency.

[4]Other examples of studies utilizing procurement-based variation to study firm growth include Lee (2017) and Hvide and Meling (2023). Relatedly, Kroft et al. (2022) use variation from procurement auctions in the U.S. to estimate firm-level labor supply elasticities in order to study implications for market competition.

[5]This also complements recent work (such as McCaig and Pavcnik (2018), Rotemberg (2019), Bau and Matray (2023), and Sraer and Thesmar (2023)) that has used quasi-experimental designs, combined with existing methods for estimating firms' wedges (such as that due to Hsieh and Klenow (2009)), to ask the important but distinct question of whether policy changes appear to *change* the extent of misallocation.

2.1 Setup

We consider an economy consisting of $\mathcal{I}$ potential firms indexed by i. Each has a technology for converting a bundle of inputs (indexed by m in the set $\mathcal{M}$) into a bundle of products (indexed by j in the set $\mathcal{J}$). The output and input bundles of firm i are denoted by the vectors $\mathbf{y}_i \geq 0$ and $\mathbf{x}_i \geq 0$, respectively. An econometrician observes this economy at time "0" and wishes to assess the extent to which the allocation observed is allocatively efficient (in a sense defined below). Letting a "bar" over any variable denote the time-0 allocation, the econometrician observes the quantities $\overline{\mathbf{y}}_i$ and $\overline{\mathbf{x}}_i$ among a set of active firms (i.e., those with $\overline{y}_{ij} > 0$ for at least one product j) that we denote by $\overline{\mathcal{I}} \subseteq \mathcal{I}$. The analogous sets of active products and inputs (those produced and used in strictly positive amounts) for firm i are denoted by $\overline{\mathcal{J}}(i)$ and $\overline{\mathcal{M}}(i)$, respectively.

We describe the technology that each firm $i \in \mathcal{I}$ uses to produce output bundle $\mathbf{y}_i$ from input bundle $\mathbf{x}_i$ via the transformation function

$$F^{(i)}(\mathbf{y}_i, \mathbf{x}_i) \leq 0. \tag{1}$$

Our analysis rests on the following assumption about firms' technologies and behavior.

Assumption 1 (Firms). *(a) For all firms $i \in \mathcal{I}$, $F^{(i)}(\mathbf{y}_i, \mathbf{x}_i)$ is strictly increasing in $(\mathbf{y}_i, -\mathbf{x}_i)$. (b) For all firms $i \in \overline{\mathcal{I}}$, $F^{(i)}(\mathbf{y}_i, \mathbf{x}_i) = 0$. (c) For all firms $i \in \overline{\mathcal{I}}$, $F^{(i)}(\cdot)$ is differentiable at $(\overline{\mathbf{y}}_i, \overline{\mathbf{x}}_i)$ with respect to the outputs in $\overline{\mathcal{J}}(i)$ and the inputs in $\overline{\mathcal{M}}(i)$.*

The first part of this assumption is made essentially for expositional purposes. The second part assumes that firms produce on their technological frontiers. And the third part rules out cases where an active firm is producing at a kink in its frontier. Beyond these relatively mild restrictions, however, technologies are free to vary in arbitrary ways across firms and also across each firm's levels of production. Arbitrary extents of (finite) increasing or decreasing returns-to-scale, non-homotheticity, and input substitution are allowed for. In addition, firm behavior is unrestricted—firms may maximize profits subject to any assumption about market conduct, or they may fail to optimize at all—as long as no firm produces at a point where it could obtain strictly more output from the same bundle of inputs.

2.2 Allocative Efficiency of Production

We define two notions of allocative efficiency to which the time-0 allocation can be compared. These correspond to the choices made by a hypothetical social planner who acts in either an unconstrained manner or subject to a set of aggregate resource constraints, described below. To describe the planner's objective, we specify households' preferences about both the utility from consuming produced goods and the disutility incurred when providing inputs for production. We denote such preferences, for any household $h \in \mathcal{H}$, by

$$U^{(h)}(\{\mathbf{y}_i^h\}, \{\mathbf{x}_i^h\}), \tag{2}$$

where, for example, y_{ij}^h denotes the amount of y_{ij} consumed by household h and feasibility implies that $\sum_h y_{ij}^h \leq y_{ij}$ for all i and j.[6] Given our interest in the allocative efficiency of *production*, we assume that households behave in a simple manner, as follows:

Assumption 2 (Households). *For each household $h \in \mathcal{H}$, the following hold. (a) $U^{(h)}(\cdot)$ is differentiable at the point of consumption, $(\{\overline{\mathbf{y}}_i^h\}, \{\overline{\mathbf{x}}_i^h\})$. (b) The household chooses $(\{\mathbf{y}_i^h\}, \{\mathbf{x}_i^h\})$ to maximize $U^{(h)}(\cdot)$, while taking output prices $\mathbf{p}_i$ and input prices $\mathbf{w}$ as given, subject to the budget constraint of $\sum_{i\in\mathcal{I},j\in\mathcal{J}} p_{ij} y_{ij}^h \leq \sum_{i\in\mathcal{I},m\in\mathcal{M}} w_m x_{im}^h$.*

Below we will study a planner that selects a Pareto-efficient allocation based on the preferences in (2). This means that Assumption 2 implies, via first-order conditions of the household's problem, that for any value of $(\{\mathbf{y}_i\}, \{\mathbf{x}_i\})$ relative prices will equal each household's marginal rate of substitution. As such, the essential role of Assumption 2 is to provide sufficient conditions such that the prices paid by households for consumption goods reflect the marginal social value of such consumption (and similarly for prices received by households for inputs supplied). Assumption 2 therefore effectively normalizes to one any wedges between prices and consumers' marginal utilities (ruling out, for example, taxes on consumption or income), as these distortions are not part of our focus on productive inefficiencies.[7]

[6] We use the braces notation "$\{q_k\}$" to denote a vector whose elements are q_k.

[7] More generally, in the case of inputs that are sourced from other firms rather than households, our analysis can be thought of as normalizing to one any wedges in upstream firms so as to focus on misallocation among the firms under study.

For any firm i, and given any set of prices $(\mathbf{p}_i, \mathbf{w})$ and an allocation $(\mathbf{y}_i, \mathbf{x}_i)$, we define the *wedge* on any input m inside firm i's production of product j as

$$\mu_{ij,m}(\mathbf{y}_i, \mathbf{x}_i, \mathbf{p}_i, \mathbf{w}) \equiv -\frac{p_{ij}}{w_m} \frac{F^{(i)}_{x_{im}}(\mathbf{y}_i, \mathbf{x}_i)}{F^{(i)}_{y_{ij}}(\mathbf{y}_i, \mathbf{x}_i)}, \tag{3}$$

where $F^{(i)}_x(\cdot) \equiv \partial F^{(i)}(\cdot)/\partial x$, etc.[8] That is, a wedge is the ratio of the value marginal product of an input (for producing a given product in a given firm) to the price of that input.[9]

A standard derivation then implies that, under Assumptions 1 and 2, any Pareto-efficient allocation selected by a planner who uses firms' technologies embodied in (1) in order to maximize (2) will display *unconstrained allocative efficiency of production* (U-AEP). This entails a set of outputs $\{\mathbf{y}^{**}_i\}$ and inputs $\{\mathbf{x}^{**}_i\}$ that satisfies the first-order conditions

$$\frac{U^{(h)}_{y_j}(\{\mathbf{y}^{**}_i\}, \{\mathbf{x}^{**}_i\}) F^{(i)}_{x_{im}}(\mathbf{y}^{**}_i, \mathbf{x}^{**}_i)}{U^{(h)}_{x_{im}}(\{\mathbf{y}^{**}_i\}, \{\mathbf{x}^{**}_i\}) F^{(i)}_{y_{ij}}(\mathbf{y}^{**}_i, \mathbf{x}^{**}_i)} \leq 1 \qquad \text{for all} \quad i \in \mathcal{I}, j \in \mathcal{J}, m \in \mathcal{M}, h \in \mathcal{H} \tag{4}$$

with equality for any (i, j, m) that satisfies $y^{**}_{ij} > 0$ and $x^{**}_{im} > 0$. Using (3) and Assumption 2, however, this can be expressed more succinctly by saying that, for all produced products and used inputs, U-AEP requires

$$\mu_{ij,m}(\mathbf{p}^{**}_i, \mathbf{w}^{**}, \mathbf{y}^{**}_i, \mathbf{x}^{**}_i) \equiv \mu^{**}_{ij,m} = 1, \tag{5}$$

where the prices $(\mathbf{p}^{**}_i, \mathbf{w}^{**})$ are those that would prevail in a decentralized equilibrium that corresponds to the planner's allocation. That is, the planner's wedges $\mu^{**}_{ij,m}$ on active products and inputs would all be equal to one in the U-AEP case.

Turning to the case of constrained efficiency, we now imagine that the planner faces an aggregate input constraint on the use of any type of input. In particular, we consider a constraint that no more of the input may be used, in total, than is observed in use at time-0.

[8] Throughout, for expositional simplicity, we refer to derivatives of functions (such as $F^{(i)}(\cdot)$) that are not necessarily differentiable everywhere (since Assumptions 1 and 2 only restrict their differentiability at the time-0 allocation).

[9] In the case of a single-product firm we can write its technology as $y_i = \widetilde{F}^{(i)}(\mathbf{x}_i)$, so that $\mu_{i,m} = \frac{p_i}{w_m} \frac{\partial \widetilde{F}^{(i)}}{\partial x_{im}}$.

Such a constraint can be written as:

$$\sum_{i\in\mathcal{I}} x_{im} \leq \sum_{i\in\mathcal{I}} \overline{x}_{im} \equiv \overline{X}_m \qquad \text{for all} \;\; m \in \mathcal{M}. \tag{6}$$

It is common in the misallocation literature to study allocative efficiency in the presence of such constraints. One benefit of doing so is that the investigation of misallocation can be decomposed into two sources: (a) misallocation of inputs conditional on $\overline{X}_m$; and (b) misallocation in the total input level $\overline{X}_m$, itself. Another motivation involves a connection to aggregate TFP, which holds aggregate inputs constant. In line with the steps used earlier, under Assumptions 1 and 2, and if the constraint (6) holds, any Pareto-efficient allocation $(\mathbf{y}_i^*, \mathbf{x}_i^*)$ displays *constrained allocative efficiency of production* (C-AEP) in which

$$\mu_{ij,m}(\mathbf{y}_i^*, \mathbf{x}_i^*, \mathbf{p}_i^*, \mathbf{w}^*) \equiv \mu_{ij,m}^* \leq \chi_m \qquad \text{for all} \;\; i \in \mathcal{I}, j \in \mathcal{J}, m \in \mathcal{M}, \tag{7}$$

for some constant $\chi_m > 0$, and with equality for any (i, j, m) that features $y_{ij}^* > 0$ and $x_{im}^* > 0$. In this constrained case, the planner's wedges $\mu_{ij,m}^*$ will be equal to each other, i.e., equal to some common value χ_m, within the same input type (among strictly produced outputs and used inputs), but this common value is not necessarily equal to one. This highlights how the essence of C-AEP is a lack of dispersion in wedges rather than their level.

2.3 Testing for Allocative Efficiency in Production

Having seen what an AEP allocation would look like, we turn to a test for whether a given allocation in the data appears consistent with AEP. We consider an econometrician who, at time-0, observes outputs $\{\overline{\mathbf{y}}_i\}$ and inputs $\{\overline{\mathbf{x}}_i\}$ for all firms $i \in \overline{\mathcal{I}}$, as discussed above. In addition, the analyst observes the prices of such outputs and inputs, $\{\overline{\mathbf{p}}_i\}$ and $\overline{\mathbf{w}}$ respectively. Using the definition in equation (3), the actual wedges at time-0 therefore correspond to

$$\overline{\mu}_{ij,m} \equiv \mu_{ij,m}(\overline{\mathbf{y}}_i, \overline{\mathbf{x}}_i, \overline{\mathbf{p}}_i, \overline{\mathbf{w}}) = -\frac{\overline{p}_{ij}}{\overline{w}_m} \frac{F_{x_{im}}^{(i)}(\overline{\mathbf{y}}_i, \overline{\mathbf{x}}_i)}{F_{y_{ij}}^{(i)}(\overline{\mathbf{y}}_i, \overline{\mathbf{x}}_i)}. \tag{8}$$

It is then apparent that the observed allocation is efficient if it corresponds to the planner's allocation in the following sense.

Definition 1. *If the observed allocation* $\{\overline{\mathbf{y}}_i, \overline{\mathbf{x}}_i\}$ *is U-AEP, then it satisfies* $\overline{\mu}_{ij,m} = 1$ *for all* $i \in \overline{\mathcal{I}}, j \in \overline{\mathcal{J}}(i), m \in \overline{\mathcal{M}}(i)$. *If the observed allocation is C-AEP, then it satisfies* $\overline{\mu}_{ij,m} = \chi_m$, *where* $\chi_m > 0$, *for all* $i \in \overline{\mathcal{I}}, j \in \overline{\mathcal{J}}(i), m \in \overline{\mathcal{M}}(i)$.

As is implied by this definition, the null hypotheses to be tested can be stated as:

$$\begin{aligned} &H_0\,(\text{U-AEP}) : \overline{\mu}_{ij,m} = 1 \;\text{ for all }\; i \in \overline{\mathcal{I}}, j \in \overline{\mathcal{J}}(i), m \in \overline{\mathcal{M}}(i), \\ &H_0\,(\text{C-AEP}) : \overline{\mu}_{ij,m} = \chi_m \;\text{ for some } \chi_m > 0 \;\text{ and for all }\; i \in \overline{\mathcal{I}}, j \in \overline{\mathcal{J}}(i), m \in \overline{\mathcal{M}}(i). \end{aligned} \tag{9}$$

This makes it clear that testing for AEP would be straightforward if the set of wedges $\{\overline{\mu}_{ij,m}\}$ were observable. Unfortunately, as the definition of $\mu_{ij,m}$ in (3) makes clear, the wedge $\overline{\mu}_{ij,m}$ depends on the marginal product of input m for product j in firm i—the ratio of two derivatives of firm i's transformation function at the observed allocation, $-F^{(i)}_{x_{im}}(\overline{\mathbf{y}}_i, \overline{\mathbf{x}}_i)/F^{(i)}_{y_{ij}}(\overline{\mathbf{y}}_i, \overline{\mathbf{x}}_i)$. Knowledge of such marginal products can be obtained from knowledge of the transformation function $F^{(i)}(\cdot)$, but obtaining knowledge of $F^{(i)}(\cdot)$ is not possible without parametric restrictions here because, with a separate function $F^{(i)}(\cdot)$ per firm, there is no variation with which the econometrician could estimate these functions nonparametrically.

We overcome this challenge by analyzing changes over time. Let $\Delta y_{ij} \equiv y_{ij,t=1} - \overline{y}_{ij}$ denote the change in any variable, such as y_{ij}, from time-0 to some later date that we refer to as "time-1". Following Hall (1988), we then apply a first-order Taylor expansion to the transformation function $F^{(i)}(\cdot)$ around the point $(\overline{\mathbf{y}}_i, \overline{\mathbf{x}}_i)$, as is valid under Assumption 1(c). Using the definition of $\overline{\mu}_{ij,m}$, changes in the output bundle must therefore relate to changes in the input bundle via

$$\sum_{j \in \overline{\mathcal{J}}(i)} \frac{\overline{\mu}_{ij,m_0}}{\overline{\mu}_{ij_0,m_0}} \overline{p}_{ij} \Delta y_{ij} = \sum_{m \in \overline{\mathcal{M}}(i)} \overline{\mu}_{ij_0,m} \overline{w}_m \Delta x_{im} + \varepsilon_i, \tag{10}$$

for any firm $i \in \overline{\mathcal{I}}$, where m_0 and j_0 denote arbitrarily chosen reference inputs and products, respectively.[10] Here, ε_i captures any possible changes in the production function itself (such

[10] This expression assumes that $m_0 \in \overline{\mathcal{M}}(i)$ for all i, but this is purely for simplicity of notation.

as, but not limited to, a change in TFP), the consequences of any higher-order terms in the Taylor expansion, and the consequences of any new products $j \notin \overline{\mathcal{J}}(i)$ or inputs $m \notin \overline{\mathcal{M}}(i)$. Importantly, this expression is valid for any change in outputs Δy_{ij} and inputs Δx_{im}, and hence does not take a stand on why (or indeed whether) the firm changed certain of its inputs and hence its outputs. We use this equation extensively in what follows.

As with any test, we seek a function of observables that could distinguish H_0 from an alternative hypothesis (in this case, the existence of misallocation). Towards that goal, we define the following change in (fixed-price) revenues minus (fixed-price and $\boldsymbol{\chi}$-adjusted) costs

$$\Delta\Pi_i(\boldsymbol{\chi}) \equiv \sum_{j \in \overline{\mathcal{J}}(i)} \overline{p}_{ij} \Delta y_{ij} - \sum_{m \in \overline{\mathcal{M}}(i)} \chi_m \overline{w}_m \Delta x_{im}. \tag{11}$$

Importantly, $\Delta\Pi_i(\boldsymbol{\chi})$ is a function of data alone, given any candidate value of $\boldsymbol{\chi}$. Combining equations (10) and (11), we see that under the null of (either U- or C-) AEP, we have $\Delta\Pi_i(\boldsymbol{\chi}) = \varepsilon_i$. Finally, we introduce an additional observable, an "instrument," denoted by Z_i, that is assumed to satisfy the following exogeneity restriction.

Assumption 3 (Exogeneity). *The econometrician observes an instrumental variable (IV), denoted Z_i, that satisfies statistical independence with respect to ε_i: $Z_i \perp\!\!\!\perp \varepsilon_i$.*

We discuss practical considerations for assessing the validity of this assumption below. When Assumption 3 holds, and under the null, it is the case that $\Delta\Pi_i(\boldsymbol{\chi}) \perp\!\!\!\perp Z_i$. Since both Z_i and $\Delta\Pi_i(\boldsymbol{\chi})$ are observable under any candidate value of $\boldsymbol{\chi}$, this hypothesis of statistical independence is testable. One method for doing so is to consider the following nonparametric relationship among all firms $i \in \overline{\mathcal{I}}$

$$\Delta\Pi_i(\boldsymbol{\chi}) = f(Z_i) + \nu_i, \tag{12}$$

for a set of flexible functions $f(\cdot)$. Then the AEP hypotheses can be stated as:

$$\begin{aligned} &H_0(\text{U-AEP}) : \text{when } \boldsymbol{\chi} = \mathbf{1},\ f(\cdot) = 0, \\ &H_0(\text{C-AEP}) : \text{for some } \boldsymbol{\chi} > \mathbf{0},\ f(\cdot) = 0. \end{aligned} \tag{13}$$

We summarize the discussion so far in the following proposition.

Proposition 1. *Suppose Assumptions 1-3 hold. Then a nonparametric test for U-AEP can be performed by estimating the function $f(\cdot)$ in equation (12), evaluating $\Delta\Pi_i(\boldsymbol{\chi})$ at $\boldsymbol{\chi} = \mathbf{1}$, and rejecting the null whenever $f(\cdot) \neq 0$. A nonparametric test for C-AEP can similarly be performed by estimating the function $f(\cdot)$ in equation (12), but while evaluating $\Delta\Pi_i(\boldsymbol{\chi})$ at all $\boldsymbol{\chi} > \mathbf{0}$, and rejecting the null whenever there exists no value of $\boldsymbol{\chi} > \mathbf{0}$ at which $f(\cdot) = 0$.*

The intuition behind the test in Proposition 1 is as follows. Suppose that Z_i represents a demand shock that applies to some firms but not others. Further, suppose that Z_i is observed to be positively correlated with $\Delta\Pi_i(\mathbf{1})$, which would correspond to a case in which one estimates an $f(\cdot) \neq 0$. Then this would imply that the demand shock caused some firms' (time-0 price-valued) outputs to grow by more than their (time-0 price-valued) inputs. As long as this demand shock Z_i satisfies Assumption 3, such a scenario is inconsistent with U-AEP because it implies that some wedges are larger than one. Further, suppose we find some value of $\boldsymbol{\chi} > \mathbf{0}$ at which there appears to be no function $f(\cdot)$, within some flexible set of functions, that delivers an estimate with $f(\cdot) \neq 0$. This is consistent with C-AEP because it implies that there exists a set of "prices" χ_m for each input type m such that, when inputs are valued in a way that includes χ_m, input growth does now equal output growth.

In practice, there are a number of ways to choose flexible functional forms for $f(\cdot)$ in order to carry out the test of $f(\cdot) = 0$ in equation (12). In Section 5 we do so by estimating a quantile regression relationship between $\Delta\Pi_i(\boldsymbol{\chi})$ and Z_i, at a range of candidate values of $\boldsymbol{\chi}$, and testing for the presence of non-zero effects at any quantile. We also employ randomization inference, given that the null is sharp (for any $\boldsymbol{\chi}$) and the stochastic distribution of Z_i is known (given the randomization protocol) in our setting.

Returning to the choice of instruments Z_i, there are two considerations. First, Z_i must satisfy Assumption 3. Recall that ε_i consists of three components: (a) changes to the production technology (such as a TFP shock) between time-0 and time-1; (b) non-linear terms in the Taylor expansion of equation (10); and (c) the impact of new types of inputs or outputs. An instrument satisfies the first of these components if it derives from characteristics of the firm's environment that are unrelated to its own technology. As discussed in the

Introduction, potential examples include changes in firm demand, in the firm's competitive environment or its competitors' characteristics, or in subsidies to the firm's inputs. Our application uses a randomized component of government procurement to generate a Z_i that can be interpreted as a demand shock. Concerns due to component (b) can be probed through the use of different values of Z_i that represent relatively small and large external shocks, as we do for our context in Section 4.1. Component (c) will cause bias to the extent that the new inputs or outputs caused by the instrument have large value products but, as equation (7) makes clear, under the null of efficiency these effects are expected to be small.

The second consideration in choosing instruments concerns the power of the test in Proposition 1. Power can be augmented, or directed towards particular alternatives, through the use of instruments Z_i that drive different extents of variation in input use, $\overline{w}_m \Delta x_{im}$. If Z_i had only a weak correlation with input changes, then the econometrician would have no ability to learn whether there is zero correlation between Z_i and $\Delta\Pi_i(\boldsymbol{\chi})$, at any $\boldsymbol{\chi}$. This is akin to a "first-stage" relevance condition in standard IV settings.[11] Ultimately, the strength of the first-stage correlation between Z_i and input changes is an empirical matter that needs to be assessed in specific settings, as we do in Section 4 below.

2.4 Measuring the Distribution of Wedges

Proposition 1 develops nonparametric tests for the existence of misallocation in production at a given observed point in time. These tests amount to evaluating whether the distribution of wedges $\{\overline{\mu}_{ij,m}\}$ satisfies certain features: constancy across firm-products within input types, in the case of C-AEP; and degeneracy at the value of one for all firms, products and inputs, in the case of U-AEP. Our final procedure complements such tests by providing point estimates of features of the wedge distribution such as its weighted moments (with observed weights). These estimates may often be of interest in their own right. But, as we describe further in

[11] One theoretical example that would lack power occurs in the case of a price-taking firm with increasing marginal costs, and where Z_i is a demand shock that offers to purchase more at the market price. Such a firm would be indifferent to this demand shock and hold its scale constant, but alternative instruments (such as an output subsidy) would work in this setting. A second example that lacks power would arise if the instrument derives from a market-level shock that induces equal input responses among all firms. More generally, since the essence of this test is to observe the effects of input changes, all else equal, it will have less power to detect the misallocation of types of inputs that feature large adjustment costs.

Section 2.5, they also play a central role in estimates of the aggregate costs of misallocation.

To proceed with such estimation, we rearrange equation (10) as

$$\bar{p}_{ij_0}\Delta y_{ij_0} = -\sum_{j\in\bar{\mathcal{J}}(i)-j_0} \frac{\bar{\mu}_{ij,m_0}}{\bar{\mu}_{ij_0,m_0}}\bar{p}_{ij}\Delta y_{ij} + \sum_{m\in\overline{\mathcal{M}}(i)} \bar{\mu}_{ij_0,m}\bar{w}_m\Delta x_{im} + \varepsilon_i, \tag{14}$$

which holds for each firm $i \in \bar{\mathcal{I}}$. We begin by considering this equation when applied only to the set of firms that produce a single product at time-0, but we return to the multi-product firm case below. To simplify notation, we therefore drop the product subscript j temporarily. For single-product firms, equation (14) then becomes

$$\bar{p}_i\Delta y_i = \sum_{m\in\mathcal{M}} \bar{\mu}_{i,m}\bar{w}_m\Delta x_{im} + \varepsilon_i, \tag{15}$$

where here (and henceforth) we follow the convention that $\Delta x_{im} = 0$ if $\bar{x}_{im} = 0$.

This equation corresponds to a cross-firm regression model relating $\bar{p}_i\Delta y_i$ to a set of regressors given by $\bar{w}_m\Delta x_{im}$ for each $m \in \mathcal{M}$. In particular, this model takes the form of an instrumental variables correlated random coefficients (IVCRC) model. This is because: (a) each unit of observation i not only has its own unobserved intercept ε_i but also its own unobserved coefficient $\bar{\mu}_{i,m}$ on each regressor; (b) it features endogeneity as we expect any regressor (the change in inputs of type m) to be correlated with the error term since this term captures changes in the firm's production technology, to which the firm's input choices may respond; and (c) we can expect the coefficients $\bar{\mu}_{i,m}$ to be potentially correlated with the error term ε_i, for example, if firms with high markups are more likely to receive productivity shocks. Masten and Torgovitsky (2016) develop tools for the study of such IVCRC models in cases with suitable instrumental variables, as described in the following condition.

Assumption 4 (IV). *The econometrician has access to a vector of instruments Z_{im}, one for each $m \in \mathcal{M}$. Each instrument satisfies: (a) $Z_{im} \perp\!\!\!\perp (\varepsilon_i, \bar{\mu}_{i,m})$; (b) $\bar{w}_m\Delta x_{im} = h_m(Z_{im}, V_{im})$ for some unknown $h_m(\cdot)$ and scalar V_{im}, with $\frac{\partial h_m}{\partial V_{im}} > 0$ for all $m \in \overline{\mathcal{M}}(i)$; and (c) there exists variation in Z_{im} at almost every V_{im} for all $m \in \overline{\mathcal{M}}(i)$.*

Parts (a) and (c) of this assumption are analogous to standard requirements for IV estimation—they embody the familiar requirements of exogeneity and relevance, respectively.

Condition (a) requires the instruments to be independent of both the residual ε_i and the wedge $\bar{\mu}_{i,m}$.[12] One *a priori* concern about the validity of part (c) is the presence of adjustment costs, as discussed above. Just as such costs will, all else equal, reduce the power to detect misallocation of inputs that rarely adjust, they may also hinder the ability to estimate the distribution of wedges on such inputs because it may prove challenging to find sufficiently strong instruments for them. However, as discussed in Masten and Torgovitsky (2016), condition (c) is testable.

Part (b) is unique to IVCRC. Masten and Torgovitsky (2016) refer to this as a "first-stage rank-invariance" condition. It requires that the ranking of firms in terms of their input changes, for any input type m, when all firms (hypothetically) receive a low value of Z_{im} is the same as that when all firms receive a high value of Z_{im}. Put differently, firms can respond to Z_{im} in heterogeneous ways, but not in such a way that alters their rank in the conditional-on-Z_{im} distribution of changes of input m. This is more likely to be satisfied when firms' "background" reasons for adjusting inputs (i.e., those driven by V_{im}) have large variance relative to the heterogeneity in their responses to Z_{im}. We return to this point in the context of our application in Section 6.

Masten and Torgovitsky (2016) show that, under Assumption 4, a consistent estimator of $\mathbb{E}[\bar{\mu}_m]$ can be constructed for any $m \in \mathcal{M}$—that is, for the expected value of firms' wedges $\bar{\mu}_{i,m}$ on input m, with the expectation taken across firms i—in equation (15).[13] We augment this procedure to obtain an estimate of the analogous expectation when it is weighted according to an arbitrary vector of firm-specific weights $\{\alpha_m\}$—that is, the set of values α_{im} for all firms i. We denote this expectation by $\mathbb{E}_{\alpha_m}[\bar{\mu}_m]$. This can be achieved by simply using the regressor $\frac{1}{N}\bar{w}_m\Delta x_{im}/\alpha_{im}$ instead of $\bar{w}_m\Delta x_{im}$.[14] Higher-order moments of the wedge distributions can also be estimated consistently by a simple extension. As Masten

[12] Our wedge estimation procedure is less vulnerable to concerns about the endogeneity of ε_i than was the test in Proposition 1 because here it is straightforward to control for higher-order terms in $\bar{w}_m\Delta x_{im}$ (to the extent that separate instruments are available for small and large changes in inputs) as well as new inputs.

[13] Since Assumption 4 invokes conditions about inputs $m \in \overline{\mathcal{M}}(i)$, the identification of $\mathbb{E}[\bar{\mu}_m]$ for any given $m \in \overline{\mathcal{M}}$ refers to the expectation over firms i for which $m \in \overline{\mathcal{M}}(i)$.

[14] In the special case of one regressor, the weighted expectation can also be approximated by interacting the regressor with a set of indicators for groups based on quantiles of the weighting variable α_i and then constructing the group-weighted average of group-specific estimates. This approximates $\mathbb{E}_\alpha[\bar{\mu}]$ increasingly well as the number of groups grows.

and Torgovitsky (2016) discuss, the square of equation (15) implies

$$(\bar{p}_i \Delta y_i)^2 = \sum_{m \in \mathcal{M}} \bar{\mu}_{i,m}^2 (\bar{w}_m \Delta x_{im})^2 + 2 \sum_{\substack{m,m' \in \mathcal{M}, \\ m \neq m'}} \bar{\mu}_{i,m} \bar{\mu}_{i,m'} \bar{w}_m \Delta x_{im} \bar{w}_{m'} \Delta x_{im'} + 2\varepsilon_i \sum_{m \in \mathcal{M}} \bar{\mu}_{i,m} \bar{w}_m \Delta x_{im} + \varepsilon_i^2. \tag{16}$$

This, too, is an IVCRC model. As such, a repetition of the previous argument can be used to construct a consistent estimator of second-order moments, $\mathbb{E}[\bar{\mu}_m^2]$ and $\mathbb{E}[\bar{\mu}_m \bar{\mu}_{m'}]$.[15] By a similar argument to that stated above, weighted moments such as $\mathbb{E}_{\alpha_m}[\bar{\mu}_m^2]$ are also identified. Extensions to third- and higher-order moments are straightforward.

We summarize the preceding discussion in the next proposition:

Proposition 2. *Suppose that Assumptions 1-4 hold. Then all (finite-order) weighted moments of the distribution of wedges $\{\bar{\mu}_{i,m}\}$ are identified among $i \in \bar{\mathcal{I}}$ and $m \in \overline{\mathcal{M}}(i)$.*

Finally, we return to the discussion of the multi-product firm from equation (14). The new element here is the presence of potential additional coefficients $\frac{\bar{\mu}_{ij,m_0}}{\bar{\mu}_{ij_0,m_0}}$ on the regressors $\bar{p}_{ij} \Delta y_{ij}$, with $\bar{J}(i) - 1$ coefficients for the case in which firm i produces $\bar{J}(i)$ products at time-0.[16] These coefficients capture within-firm, cross-product dispersion in wedges, as measured by the ratio of the wedge (for an arbitrarily chosen input m_0) on any product j relative to that on the reference product j_0. Given suitable instruments, Proposition 2 can be applied to the moments of these distributions too, leading to the identification of all moments of the distribution of all wedges, both across firms, within firms across products, and within firms across inputs.

[15] If the number of input types is M, equation (16) suggests that $M(M+1)$ instruments are required for identification. However, an attractive feature of the Masten and Torgovitsky (2016) procedure is that only M instruments are required. This is because the method incorporates the structure of the mechanical relationships between basic and derived endogenous regressors (i.e., that the first-stage relationship $\bar{w}_m \Delta x_{im} = h_m(Z_{im}, V_{im})$ is the same no matter how the component $\bar{w}_m \Delta x_{im}$ appears in the regression).

[16] This can be executed as follows. Arbitrarily order each firms' $\bar{J}(i)$ products such that one (i.e. j_0) is chosen as the reference product, and hence populates the left-hand side of equation (14), and the remaining $\bar{J}(i) - 1$ products appear as regressors. For firms with less than $\bar{J} \equiv \max_{i \in \bar{\mathcal{I}}} \bar{J}(i)$ products, use the convention of $\Delta y_{ij} = 0$ to populate the missing $\bar{J} - \bar{J}(i)$ regressor values.

2.5 Quantifying the Cost of Misallocation

Proposition 1 has developed a nonparametric test for the existence of misallocation in production at a given observed point in time. We now go further and develop a method for calculating the aggregate welfare cost of any such misallocation.

In particular, we compare the U-AEP allocation $(\mathbf{y}^{**}, \mathbf{x}^{**})$ to the actual time-0 allocation $(\overline{\mathbf{y}}, \overline{\mathbf{x}})$ by considering—à la Harberger—the second-order expansion to an aggregate welfare function around the point $(\mathbf{y}^{**}, \mathbf{x}^{**})$.[17] As in Definition 1, this can be written as a change from the U-AEP allocation, where all wedges are equal to one, to any other vector of wedges, such as those that prevail at the time-0 allocation, denoted $\left\{\overline{\mu}_{ij,m}\right\}$. Equivalently, the welfare gain due to moving from the actual wedges of $\left\{\overline{\mu}_{ij,m}\right\}$ to the U-AEP wedges of $\{1\}$ equal what we define as the total cost of the misallocation at time-0. We also decompose this total cost into two terms: (i) a cost of misallocation due to how each input type is allocated across firms, holding aggregate input use constant at the value $\bar{X}_m$; and (ii) a cost of misallocation of the aggregate input values themselves.

Quantifying the total cost of misallocation, up to second-order, amounts to summing up a series of Harberger triangles whose heights are related to the change in wedges (i.e., to the vector $\left\{\overline{\mu}_{ij,m} - 1\right\}$) and whose bases are related to the change in quantities of each good produced as a result of the change in wedges. As usual, this latter component can itself be written as a function of the change in the wedge and a function of the relevant elasticities of supply and demand. Baqaee and Farhi (2020) have recently provided a parsimonious representation of these elasticities as functions of underlying data shares (measurable from observations about expenditures in the time-0 allocation) and structural elasticities of substitution inside households' utility functions and firms' production functions.

The Baqaee and Farhi (2020) presentation emphasizes how the cost of misallocation can be written as a set of weighted first- and second-order moments of the wedge distribution, where the structural elasticities govern the importance of different moments. Assumptions about such elasticities derive from assumptions about firms and households that are stronger than we have needed (via Assumptions 1 and 2) so far to identify weighted moments of

[17] Since Proposition 2 describes an identification argument for wedge distribution moments that are higher than second-order, our focus here on second-order expansions is done for simplicity only.

the wedge distribution using Proposition 2. However, the additional assumptions about demand and supply elasticities play a key role in shaping *which* weighted moments of the wedge distribution are important for the cost of misallocation. The remainder of this section provides two examples that illustrate this logic. Our empirical application in Section 6 then draws on these examples to put such an approach into practice.

Example #1: Multiple input types and a single sector

Consider an economy with single-product, profit-maximizing firms, each producing with a Cobb-Douglas production function that combines capital (denoted x_{iK}) and labor (x_{iL}) via $y_i = A_i(x_{iK})^{\bar{\alpha}_K}(x_{iL})^{1-\bar{\alpha}_K}$. Capital and labor are both in fixed aggregate supply to these firms, so the only source of misallocation will concern the extent to which each input is used across firms, not the overall extent of input use; that is, in this example there is no distinction between U-AEP and C-AEP. The representative consumer has CES preferences for the firms' outputs, with elasticity of substitution θ. Finally, suppose that the time-0 allocation $(\overline{\mathbf{y}}, \overline{\mathbf{x}})$ and prices $(\overline{\mathbf{p}}, \overline{\mathbf{w}})$ imply (via the definition in equation 8) that firm i's capital and labor wedges are given by $\overline{\mu}_{i,K}$ and $\overline{\mu}_{i,K}$, respectively. This is therefore the economy of Hsieh and Klenow (2009) but with a single sector.

Using the results in Baqaee and Farhi (2020), it is straightforward to show that, up to a second-order approximation, the increase in welfare (relative to the initial level) from eliminating the wedges in this economy is given by

$$\frac{\Delta W}{W} = \frac{1}{2}\bar{\alpha}_K(1-\bar{\alpha}_K)\mathbb{V}ar_{\overline{\lambda}}[\bar{\mu}_K - \bar{\mu}_L] + \frac{1}{2}\theta\mathbb{V}ar_{\overline{\lambda}}[\bar{\alpha}_K\bar{\mu}_K + (1-\bar{\alpha}_K)\bar{\mu}_L], \tag{17}$$

where $\mathbb{V}ar_a[b] \equiv \mathbb{E}_a[b^2] - (\mathbb{E}_a[b])^2$ denotes the variance of the vector $\{b_k\}$ weighted by the vector $\{a_k\}$, and $\mathbb{E}_a[b]$ denotes the expectation of the vector b weighted by the vector a. In this case, the variances are weighted by the firms' sales shares, denoted $\overline{\lambda}_i \equiv \frac{\bar{p}_i\bar{y}_i}{\sum_{i'}\bar{p}_{i'}\bar{y}_{i'}}$.

The total cost of misallocation in this example derives from misallocation of inputs across firms even while aggregate inputs are held fixed, but this misallocation can be split into two forms, as is apparent in equation (17). The first term captures the effects of within-firm substitution (which has an elasticity of one in this Cobb-Douglas case) to the potential dispersion in wedges across the two inputs within any firm. And the second term captures

the effects of across-firm substitution, on the behalf of consumers (and hence scaling by θ), to the potential differences in cost-weighted average wedges (i.e., $\bar{\alpha}_K \bar{\mu}_{i,K} + (1 - \bar{\alpha}_K)\bar{\mu}_{i,L}$) of different firms. Because both types of inputs in this economy are in fixed aggregate supply, the relevant features of the wedge distribution that matter for misallocation all concern *dispersion* rather than average levels. However, both of the variance measures that matter here are weighted by $\bar{\lambda}$, since the wedges in larger firms are more costly.

Proposition 2 highlights how all weighted, uncentered moments of the distribution of wedges across firms can be identified. This is directly applicable to the cost of misallocation in equation (17), which can be easily converted from centered second-order weighted moments to ones that are uncentered. Hence, in a setting with instruments that satisfy Assumption 4, and with knowledge of the demand parameter θ, the cost of misallocation (up to a second-order approximation) in this example can be estimated consistently.

Example #2: Multiple sectors and endogenous input supply

We continue with a setting featuring single-product, profit-maximizing firms. But we now allow each firm to have its own arbitrary production function (involving an arbitrary set of inputs) so long as the technology satisfies Assumption 1 and displays constant returns locally to the time-0 allocation.[18] The representative consumer has nested CES preferences over the products produced by these firms; in particular, the consumer's elasticity of substitution between the firms within sector s (which we denote by $\mathcal{I}(s)$) is given by θ_s and that for substitution across sector-specific bundles is given by ρ.

In contrast to the previous example, we now allow the representative consumer household to also supply the inputs x_{im} to these firms. The household is endowed with a fixed amount of time that it can freely convert into inputs or retain as leisure; it has an elasticity of substitution η between leisure and the bundle of final consumption goods. This endogenous input supply means that the total cost of misallocation will derive from both a component due to misallocation of the aggregate inputs $\bar{X}_m$ that are supplied at time-0, and from a second component due to the misallocation of those total amounts themselves. Finally, we assume that each firm has the same wedge on each of its inputs, which we denote by $\overline{\mu}_{i,m} = \overline{\mu}_i$

[18] A natural technology that fits this form is one with arbitrary overhead costs and constant marginal costs (at fixed input prices and wedges).

for all m. This would be the case if, for example, the underlying cause of potential wedges is firms' market power in their product markets, and/or taxes and subsidies on firms' sales.

Again, the tools in Baqaee and Farhi (2020) make it easy to calculate the total cost of misallocation due to wedges in this economy. As before, we let $\bar{\lambda}_i \equiv \frac{\bar{p}_i \bar{y}_i}{\sum_{i'} \bar{p}_{i'} \bar{y}_{i'}}$ denote the share of firm i's sales in total goods consumption. We also let $\bar{\psi}_s \equiv \frac{\sum_{i \in \mathcal{I}(s)} \bar{p}_i \bar{y}_i}{\sum_{i'} \bar{p}_{i'} \bar{y}_{i'}}$ denote the share of goods consumption expenditure devoted to sector s and let $\bar{\chi}_{i(s)} \equiv \bar{\lambda}_i / \bar{\psi}_s$ denote the share of firm i within sector s. Finally, we let $\bar{\omega}_C$ denote the share of the household's virtual income spent on consumption goods.[19] Then we have

$$\frac{\Delta W}{W} = \frac{1}{2}\bar{\omega}_C \sum_s \theta_s \bar{\psi}_s \mathbb{V}ar_{\bar{\chi}_{(s)}}[\bar{\mu}] + \frac{1}{2}\bar{\omega}_C \rho \mathbb{V}ar_{\bar{\psi}}\left[\mathbb{E}_{\bar{\lambda}_{(s)}}[\bar{\mu}]\right] + \frac{1}{2}\bar{\omega}_C(1-\bar{\omega}_C)\eta\left(\mathbb{E}_{\bar{\lambda}}[\bar{\mu}]-1\right)^2 \tag{18}$$

To unpack this expression, we begin by noting that all terms are multiplied by ω_C, as the only wedges in this economy are in consumption (rather than leisure). The first term captures the average effect of within-sector dispersion in wedges across firms. It therefore scales with the size of the within-sector demand elasticity of substitution, θ_s. In particular, $\mathbb{V}ar_{\bar{\chi}_{(s)}}(\bar{\mu})$ measures the amount of such dispersion within sector s, weighted by the size of each firm relative to the sector (i.e., by $\bar{\chi}_{(s)}$). The second term captures cross-sector dispersion in the average wedge within each sector (i.e., $\mathbb{E}_{\bar{\lambda}_{(s)}}[\bar{\mu}]$). This scales with ρ, the consumer's substitution elasticity across sectors.

The misallocation of the aggregates $\bar{X}_m$ across firms and sectors causes a welfare cost equal to the sum of these first two terms. By contrast, the final term arises due to misallocation of each $\bar{X}_m$ itself. In particular, this component of misallocation exists when the consumption sector-wide (sales-weighted) *average* level of wedges is different from one. Unsurprisingly, this term scales with both the elasticity of input supply (η) and the size of leisure in the economy ($1-\bar{\omega}_C$) since it is the division of aggregate inputs between firm production and leisure that is potentially misallocated. Finally, a notable feature of expression (18) is that technological features (of each firm's production function) do not enter, since within-firm dispersion is zero. However, this expression can be augmented to include such phenomena by simply adding components such as the first term in equation (17).

[19] If the household's time endowment is $\bar{T}$ and it earns the price $\bar{w}$ for selling inputs then $\bar{\omega}_C \equiv \frac{\sum_i \bar{p}_i \bar{y}_i}{\bar{w}\bar{T}}$.

3 Background and Data

As discussed above, our procedures require an instrument that is correlated with changes in firms' input use but uncorrelated with changes in firms' technologies. We now turn to an application in which we can construct such an instrument based on demand shocks from a randomized component of Ecuador's public procurement system. This section describes the procurement lottery process, the administrative data we use, and the characteristics of the firms that participate in procurement lotteries. Finally, we outline our empirical procedure for isolating the exogenous components of demand faced generated by the lotteries. Appendix B contains further details on all aspects of data construction.

3.1 Ecuador's Procurement Lottery System

Starting in 2009, contracts for public construction projects below a certain value were required to be allocated based on randomized lotteries among qualified suppliers.[20] Examples of such contracts include construction or maintenance of public buildings, small roads and town squares, schools, sewerage, and wells or water distribution channels. The contract can involve both physical construction and services done in advance of physical construction, such as those provided by architects. Procurement made through this lottery procedure represents about 4% of total public procurement during our period of analysis.

The procurement process for contracts awarded by lottery has several steps. First, any tax-compliant firm can choose to register in the system. Second, government entities initiate a given procurement contract by sending specifications and an expected budget to the national procurement office (SERCOP).[21] Third, firms that are registered to provide the designated type of service, and are in good standing in regards to past contracts, are invited by SERCOP to submit applications, which include proof of relevant qualifications.[22] Fourth, the procuring entity determines which applicants qualify for the contract. We refer to this

[20]The threshold value is 0.00007% of the central government's annual budget. The Ecuadorian economy is fully dollarized, and this threshold corresponds to $134,176 in 2009 and $240,100 in 2014.

[21]Roughly 5% of procuring entities are formally private-sector entities with a large share of state ownership.

[22]The invitation process is sometimes made over several rounds and often favors SMEs and local firms. In addition, starting in 2013, SERCOP required that the total amount of contracts a firm may enter at any given time is limited to the maximal allowed contract value (i.e., 0.0007% of the government budget).

set of firms as "lottery participants". Finally, an automatic and centralized program at SERCOP determines the winner of the contract through a lottery among the participants. As a result, even though participation in any given lottery is the result of deliberate selection on both sides, the allocation of the contract among participants will be randomly determined within the set of lotteries that have more than one participant (which will be our focus).

We compile data on the value, date and anticipated duration of each lottery contract between 2009 and 2014, the firms involved in each lottery, and the winner of the contract.[23] Restricting the sample to lotteries with at least two participants results in 18,474 lotteries with 9,393 unique firms who participate at least once over this time period. The first two panels of Table 1 report summary statistics about the contract lotteries and firm participation in these lotteries. Contracts have a mean value of about \$47,000 (and a median of \$32,000) and are usually of a short anticipated duration, with a mean (and median) of about 2 months. On average, 10 firms participate in a lottery (median 4). Lottery participation is relatively frequent among participating firms, averaging about 3.5 times a year (median once per year).

3.2 Firm Data

We match the procurement data to administrative data obtained from Ecuador's tax authority (Servicio de Rentas Internas). This begins with annually filed firm income tax returns for 2008 to 2015.[24] We observe all line items in these filings, including wages, costs, revenues, and profits. We use these line items to construct the total sales and costs measures used for our main analysis. Total sales include domestic sales, exports, and other income (e.g., received professional fees). Total costs comprise labor and non-labor costs (including, for example, capital costs, expenses on intermediaries, maintenance and repairs, and real estate rent). All our sales and costs variables are gross of taxes.

We combine these annual income tax filings with two data sources that contain information on wages, costs and sales on a monthly basis. The first is matched employer-employee social security data, which provides information on monthly earnings at the worker level

[23] We obtained these data by scraping them from the SERCOP website: https://www.compraspublicas.gob.ec/ProcesoContratacion/compras/PC/buscarProceso.cpe.

[24] Participating firms include both incorporated firms, which file the corporate income tax form (F101), and sole proprietorships, which file a combined business and individual income tax return (F102).

(available for 2007-2017). And the second is third-party reported monthly information on firms' sales filed by their clients' purchase annexes. Such purchase annexes have to be filed by all incorporated firms, government agencies and large sole proprietorships as part of the value added tax (VAT) requirements.[25] We calculate third-party reported sales by summing purchases from a given supplier across the purchase annexes of all its client entities.[26] These monthly data for sales and labor costs, though less complete than the annual data, allow us to better illustrate the dynamic paths of the treatment effects of winning a lottery.

By their nature, administrative tax data reflect reported economic activity. We can mitigate concerns about potential misreporting by validating our estimates for sales, showing below that estimated treatment effects on sales are almost identical when using either self-reported sales or third-party reported sales. Further, we find that this treatment effect on total sales is almost entirely accounted for by sales to procuring entities, which is reassuring since these are public entities that have no incentive to misreport their tax filings.

Our analysis focuses on $9,393$ firms that ever participate in a multi-participant lottery between 2009 and 2014. To account for firm entry, we use a sample of firm-year observations that begins, for each firm, when the firm first appears to be economically active. We define this status for a given firm as beginning when it first self-reports positive sales or costs in its firm income tax forms or otherwise appears in any of the above data sets (as a lottery participant, an employer in the social security data, or as a supplier in another firm's purchase annex). Once a firm is economically active, we impute zeros for any future missing data.

The third panel of Table 1 reports summary statistics by firm (for the first year a firm participates in any lottery). Sample firms are on average 11 years old and sell to 5 clients per year (median 2). Mean annual self-reported sales for participating firms are around $\$141,000$, with a lower median ($\$53,000$) as reflects the skewness of the firm size distribution. Third-party reported sales are similar, with a mean of $\$133,000$ (median $\$48,000$). Mean annual costs for participating firms are $\$124,000$ (median $\$42,000$) and mean profits are $\$17,000$ (median $\$11,000$). The mean number of employees is 4.4 (median 2).

[25]Purchase annexes include the value, date and supplier firm ID of all purchases a firm makes.

[26]The third-party sales measure is a lower bound since not all client firms file purchase annexes and sales to final consumers are not included. In practice, the two sales measures are very similar for our sample (see Table 1).

Figure 1 plots the distribution of firm sizes, in terms of annual sales, and also allows for a comparison of lottery participants to other firms in Ecuador. In particular, panel 1(a) compares lottery participants to other economically active Ecuadorian firms, whereas panel 1(b) restricts the comparison group to those within the same industry (i.e., the construction and engineering sectors). The size distributions of participants and non-participants are broadly similar. Relative to both comparison groups, the participant distribution is shifted to the right but with less mass in the upper and lower tails of the distribution.

3.3 Using Procurement Lotteries to Construct Demand Shocks

The randomized assignment of contracts provides a source of exogenous variation, coming from the demand side, in the use of inputs. Our analysis below uses this variation to construct an instrumental variable Z_{it} for each firm i and time period t in our data. We then apply this instrument to the testing and estimation procedures described in Sections 2.3 and 2.4. We need this instrument to satisfy the statistical independence assumption invoked in Assumptions 3 and 4(a). While the variation within any given lottery is akin to a simple randomized trial, pooling this variation across all firm-year observations is more complex because the timing and nature of lottery entry (in regards to lottery characteristics such as their contract value and competitiveness) may be correlated with unobserved determinants of firm growth.[27] To address this issue, we draw on the ideas in Doran et al. (2022) and Borusyak and Hull (2021) to aggregate the multiple randomized lotteries and generate quasi-experimental variation in demand from Ecuador's procurement lottery system.

Let k index all procurement lotteries, and let $\mathcal{K}_{it}$ denote the set of lotteries that firm i enters in period t. Further, let A_k denote the contract value and N_k the number of participating firms in lottery k. It follows that the amount of winnings that firm i obtains from lottery k is a random variable W_{ik} that is binomially distributed (equal to A_k with probability $1/N_k$, and zero otherwise). Hence, a firm's total winnings in any time period t is a random variable $W_{it} \equiv \sum_{k \in \mathcal{K}_{it}} W_{ik}$ that is a weighted, binomially distributed random variable, with a pth central moment—which we denote by $M_p(\{A_k, N_k\}_{k \in \mathcal{K}_{it}})$—that can be easily calculated.[28]

[27] In addition, firms can enter multiple lotteries per time period, which precludes the use of lottery fixed effects as a means for isolating purely random variation.

[28] For example, the first moment (or expected value) of firm i's winnings at time t is simply

When firms participate in different lotteries at different points in time they are therefore exposed, and potentially endogenously so, to different probability distributions of randomly-generated winnings. However, conditional on two firms participating in lotteries with the same distribution, the *realization* of the variable W_{it} will differ in a purely random manner across these two firms. Put differently, conditional on all moments of $M_p(\{A_k, N_k\}_{k\in\mathcal{K}_{it}})$, W_{it} should be independent of any pre-determined firm attributes, observed or unobserved.

One way to proceed is therefore to use W_{it} as our instrument while controlling for flexible functions of many leading moments of $M_p(\{A_k, N_k\}_{k\in\mathcal{K}_{it}})$. However, in practice, we find that controlling for anything beyond the first moment is quantitatively inconsequential—both when applied to estimators that rely only on mean independence of instruments and those that rely on full independence. In addition, as Borusyak and Hull (2021) explain, a simpler procedure that is equivalent to controlling for the first moment of winnings is to exploit the fact that, even in the absence of controls, demeaned winnings

$$D_{it} \equiv W_{it} - \mathbb{E}[W_{it} \mid \{A_k, N_k\}_{k\in\mathcal{K}_{it}}], \tag{19}$$

will be mean-independent of any firm characteristics, observed or unobserved. In the following, we refer to D_{it}, firm i's deviation from expected winnings in time t, as its *procurement winnings shock*.[29] Intuitively, D_{it} is mean independent from any firm potential outcomes because, even though firms can control expected winnings by choosing which lotteries to enter, they cannot control the random deviations from expected winnings.

We check for balance of randomization by regressing D_{it} on firms' pre-treatment characteristics. Table 2 shows that, consistent with lottery winners being randomly drawn, there is no statistically significant correlation between these characteristics and the randomly determined component of lottery winnings.[30] In addition, we will show below that none of our outcomes exhibit evidence of spurious "effects" prior to lottery realizations.

$M_1(\{A_k, N_k\}_{k\in\mathcal{K}_{it}}) \equiv \mathbb{E}[W_{it} \mid \{A_k, N_k\}_{k\in\mathcal{K}_{it}}] = \sum_{k\in\mathcal{K}_{it}} \frac{A_k}{N_k}$.

[29] Our analysis below considers data in which the time periods t are either months or years. By the linearity of expectations, D_{it} at the annual level corresponds to the sum of the monthly D_{it} in each year.

[30] These findings are consistent with Brugués et al. (2022), who show that political connections influence the distribution of regular procurement contracts in Ecuador but not of the contracts allocated by lottery.

4 Effects of Demand Shocks: Descriptive Results

This section presents descriptive evidence on average firm responses to demand shocks as well as heterogeneity in these responses. These results preview our formal tests and quantification of allocative efficiency in Sections 5 and 6.

4.1 Average Treatment Effects

We estimate effects of demand shocks on firm outcomes in an event-study framework by regressing firm outcomes on procurement winnings shocks D_{it}, the lottery-driven component of procurement contracts defined in (19). Formally, for any outcome Y_{it}, we estimate

$$Y_{it} = \alpha + \sum_{\tau=-T_{lead}}^{T_{lag}} \beta_\tau D_{i,t-\tau} + \epsilon_{it}, \tag{20}$$

where α is an intercept, β_τ is a coefficient that shows the average effect of the winnings shock $D_{i,t-\tau}$ on outcome Y_{it}, and T_{lead} and T_{lag} denote the number of included lead and lag coefficients, respectively. For example, we choose $T_{lead} = 6$ and $T_{lag} = 18$ for monthly outcomes. Throughout this section, we divide the variable D_{it} by 1,000 so that the coefficients refer to the effect of an additional \$1,000 of procurement winnings shock on the outcome of interest, and we report confidence intervals and standard errors clustered at the firm level.

Sales

We begin by estimating the treatment effect of demand shocks on firm sales. To look at dynamics at the monthly level, we use third-party reported sales (based on client entities' VAT filings). Results are presented in Figure 2. The estimated coefficients from equation (20) display a flat pre-trend at zero during the six months prior to the lottery, lending further support to the validity of the lottery randomization. There is a small spike in sales one month after the lottery, which represents the fact that many lottery contracts stipulate some up-front payments. Monthly sales rise thereafter, peaking 5 months after the lottery, and dissipate by about 14 months. There is no evidence for a temporary demand shock leaving these firms (on average) permanently larger.

We next annualize these third-party reported sales data in order to compare estimates to those of self-reported sales from firms' annual income tax returns. Columns (1) and (2) of Table 3 report the effect of $1,000 in procurement winnings shocks on sales, based on these two alternative measures, in the year of the lottery and the subsequent year.[31] The total impact on sales over this period is $708 for third-party reported sales and $669 for self-reported sales.[32]

Finally, in Figure 3, we examine whether there are heterogeneous treatment effects by contract size. We split procurement contracts into above- and below-median contract amount, and analyze firms' responses to monthly winnings shocks separately for small and large contracts. The path of monthly treatment effects—per unit of contract size—is remarkably similar. Appendix Figure A.2 presents the same results separately for small and larger firms (while holding the above/below median classification of contracts constant). We again observe very similar treatment effects across contract size within groups of small and larger firms. While it might be natural to expect that small firms would respond to large demand shocks differently from how large firms respond to small shocks, this is not the case. These findings suggest that both adjustment costs and the higher-order terms in the Taylor expansion of equation (10) are unlikely to be large in our context.

Costs

A natural question is how firm inputs scale up given the increase in firm sales seen above. We analyze this in Figure A.3, using total costs from the annual firm income tax returns, which include both labor and non-labor costs. As with sales, we see a flat pre-trend followed by a sharp increase in costs in the year of the lottery, which dissipates within two years. Figure 4 shows the impact on labor inputs, i.e., employment and labor payments measured from monthly social security data. We see a time path of impacts that is similar to that on sales, though there is a minor longer-run effect when it comes to the use of labor. Column

[31] Appendix Figure A.1 presents these results graphically, starting two years before the respective lotteries to show that there are no differential pre-trends.

[32] As discussed below, we find virtually no crowd-out of sales to other clients. The fact that cumulative sales effects from a given winnings shock are lower than the shock itself therefore derives from a mix of reductions in subsequent entry for lottery contracts posted by the same entity and *ex post* modification of contract values and scope. However, such considerations do not affect our empirical strategy since it relies only on the existence of a random source of firm input changes, not that the pass-through of contract winnings into sales growth takes any particular value (other than zero).

(3) of Table 3 reports the coefficients on the procurement winning shock variable and its one-year lag, which together imply an aggregate increase in total costs of $583 over a two-year horizon. This cost increase is driven primarily by an increase in non-labor costs (columns 4 and 5). Finally, column (6) reports the corresponding treatment effects on profits (i.e., the coefficients in column 2 minus those in column 3).

Overall, these findings concerning input use have two important implications. First, the firms in our context cannot meet additional demand simply by using existing capacity. And second, the rapid scale-up and scale-down of all observable inputs is again consistent with this being a context in which adjustment costs are limited.

Price versus quantity adjustments

The procedures developed in Section 2 rely on the ability to measure the initial period prices and changes in quantities of firm inputs and outputs. While such components are available in many settings, our data do not report prices and quantities separately. However, our data on sales and costs will still capture constant-price changes to the extent that the demand shocks have no effect on prices. Figure 5 provides evidence that is consistent with this in regards to output prices. It plots the estimated impact of procurement winning shocks on total sales (as in Figure 2), as well as on sales to four different mutually-exclusive categories of clients: (a) procuring entities for which firm i participates in at least one lottery during our study period; (b) other procuring entities (i.e., other entities that made at least one purchase through the lottery system in our study period) that firm i sells to via non-lottery means; (c) other public entities that made no purchases through the lottery system during our sample period; and (d) privately-owned firms.

Sales to the procuring entities for which firm i participates a lottery account for almost all of the effect on total sales. In particular, there is no appreciable effect on sales to the private sector: we can reject monthly effects on private sector sales larger than $\pm$ $10 for each $1,000 dollars in procurement winnings shocks at the 95% level in all periods. If the demand from private sector buyers is at all responsive to prices, as we would expect it to be, then the lack of any change in purchases from lottery winners suggests that the price at which procurement firms sell is not affected by lottery-based demand shocks.

Turning to input price adjustments, in the case of labor we can assess these directly in our data. Figure 4 has already established that the labor costs response does at least partially reflect a quantity response (in terms of the number of paid employees). Appendix Figure A.4 looks at the wage response per worker, first (in panel a) in terms of the average wage among all workers in the firm in a given month, and second (in panel b) among continuing workers who were employed at the firm already before the lottery and who stayed at the firm afterwards.[33] While the average wage paid does fall (for about eight months, before returning to previous levels), this appears to be a compositional phenomenon because the wage of continuing workers is remarkably stable in response to the demand shock, again consistent with our assumption of (input) price stability among these firms.

Implications for homogeneous wedges

The pioneering Hall (1988) study of US markups assumed: (a) that there is no wedge dispersion within firms (across products or input types, i.e. $\bar{\mu}_{ij,m} = \bar{\mu}_i$ for all j and m), which is consistent with the wedge (i.e. the markup) arising purely from output market power; and (b) that there is no markup dispersion across firms (when applied to data from a single cross-section of producers) or across time periods (when applied to data tracking one producer over time). Under these homogeneity assumptions, the single markup can be estimated from an IV regression of sales on costs while using a demand shock as the instrument. The results in Table 3 provide the reduced-form and first-stage estimates that would correspond to such an IV regression. Taking the ratio of the cumulative demand shock-driven increase in total sales in year t and $t+1$ (column 2) to that for costs (column 3) implies, under Hall's (1988) assumptions, an estimated homogeneous markup of 1.15. This already provides suggestive evidence for the fact that U-AEP—a setting in which all wedges are homogeneous and equal to 1—does not appear to hold in our context. We return to the formal version of this test, as well as to an estimate of the sales-weighted average wedge $\mathbb{E}_{\bar{\lambda}}[\bar{\mu}]$ that does not rely on the assumption of wedge homogeneity, in Sections 5 and 6, respectively.

[33]To ensure a balanced sample of workers in this event study, the analysis here tracks each firm for a window (6 months before and 18 months after) around the first lottery in which it participated.

4.2 Treatment Effect Heterogeneity

Testing for constrained allocative efficiency concerns not just the average level of the value marginal product of any given input across firms, but also the heterogeneity in such marginal products—or equivalently, the dispersion in wedges $\bar{\mu}_{ij,m}$ in equation (8). In advance of the full estimation of the distribution of such wedges in Section 6, we present here a simple exploration of treatment effect heterogeneity—that is, in the effect of price-adjusted inputs on price-adjusted outputs—based on observable firm characteristics.

First, we explore whether the previous average results are different across groups of firms with different levels of pre-treatment sales (i.e., from 2008, the year before the start of the lottery scheme).[34] Panel (a) of Figure 6 analyzes impacts of lottery-driven demand shocks on sales separately for firms with above- and below-median pre-treatment sales. These two estimates are strikingly similar, especially relative to the sampling variance indicated by the confidence intervals. Panel (b) further disaggregates firms into quintiles of pre-treatment sales and also finds very similar responses across groups. Further, the similarity of sales responses across firms also holds across groups based on other pre-treatment characteristics: number of employees, number of suppliers, labor intensity (as measured by labor costs divided by sales), and third-party reported sales (Appendix Figures A.5–A.8).

Homogeneity in firms' sales responses to a demand shock does not necessarily imply a low dispersion in firms' marginal products because firms could differ in the amount of inputs needed to achieve these output responses. However, as Figures 7 and A.9 make clear, we also see very homogeneous responses in labor and total costs across these different groups of firms. Combined, Figures 6, 7, and A.5–A.9 therefore paint the picture of a particularly strong form of marginal product homogeneity, where there is limited heterogeneity across firms in their responses to a demand shock, in either their outputs or their inputs.[35]

A key limitation of the simple evidence discussed here is that it describes firm heterogeneity based on observable characteristics. This is informative, but it may only scratch the surface of the heterogeneity that could exist in terms of unobservable characteristics. The

[34] Results in this subsection exclude firms with no sales in 2008.

[35] A finding of homogeneous output and input responses is sufficient, but not necessary, for the homogeneity of value marginal products (since marginal products concern the ratio of the two responses, so heterogeneity in output and input responses could offset one another while maintaining marginal product homogeneity).

methods in the next two sections are designed to test for, and quantify, wedge heterogeneity even when allowing for arbitrary forms of potential heterogeneity across firms, whether they align with observable firm characteristics or not.

5 Testing for Allocative Efficiency

5.1 Test Implementation

Proposition 1 has proposed a test for AEP in both its unconstrained and constrained forms. To implement this test on any given two periods (i.e., "time-0" and "time-1"), we simply evaluate $\Delta\Pi_i(\boldsymbol{\chi})$ for each firm i at the null-hypothesis value of $\boldsymbol{\chi} = \mathbf{1}$ in the case of U-AEP and at a range of values of $\boldsymbol{\chi} > 0$ in the case of C-AEP. In either case, we then test for the statistical independence of $\Delta\Pi_i(\boldsymbol{\chi})$ from an instrument Z_i (built from lottery-driven demand shocks) via the test for $f(\cdot) = 0$ in equation (12). When statistical independence is rejected, then the relevant null of allocative efficiency is rejected too.

As our test for statistical independence, we follow Ding et al. (2016), who develop procedures for tests of treatment effect heterogeneity in randomized trials.[36] This is relevant to our setting since the test for C-AEP is analogous to a test for no heterogeneous responses (of firms' outputs to an exogenously driven increase in their input use) and the U-AEP test is a nested case in which, under the null, responses are not just homogeneous but also known (to be equal to one). In particular, we estimate a quantile regression of $\Delta\Pi_i(\boldsymbol{\chi})$ on Z_i, and use as our test statistic TS the largest coefficient (in absolute value) across nine evenly spaced quantile values. If any coefficient differs substantially from zero, and hence TS is large, then the null of $f(\cdot) = 0$ is false and hence the hypothesis of independence should be rejected.

We then further take advantage of the fact that the null hypothesis is sharp, and the distribution of Z_i in our setting of randomized procurement is known, in order to use the method of randomization inference (Fisher, 1935) to calculate the probability of obtaining a given value of the test statistic under the null. This has the benefit of avoiding the need to make asymptotic distributional assumptions about the unobserved ε_i in equation (10), such

[36] Further details about our test procedures are described in Appendix C.

as how technology shocks are correlated across time and firms.

To execute this procedure, we simulate the null distribution by calculating TS_l in a set of simulations $l = 1 \dots L$, in which the identity of the winning firm of each procurement lottery at time-1 is drawn randomly from its known stochastic process. The p-value for a given test is then simply the percentage of simulations l in which the value of TS obtained when using the actual lottery winners is smaller than TS_l. Intuitively, when statistical independence of $\Delta\Pi_i(\boldsymbol{\chi})$ and Z_i is violated, we should see that there is a stronger correlation between $\Delta\Pi_i(\boldsymbol{\chi})$ and Z_i in the actual data than is likely to have occurred by chance—that is, we should see a larger value of TS when examining actual lottery realizations than the values TS_l found in many simulations l of alternative lottery realizations. Such a finding would indicate that the input reallocations caused by the lottery-driven demand shocks have generated changes in aggregate output, and hence (via Proposition 1) that the economy in question was not allocatively efficient at time-0.

This procedure can be performed on any cross-section of changes (i.e., from any "time-0" to any "time-1"). But, in practice, in order to ensure that we capture the entire two-calendar year time-path of responses seen in Table 3 (and Appendix Figure A.1), we construct changes based on differences in two-year averages. This treats the average of years $t-1$ and $t-2$ as "time-0" and that of years t and $t+1$ as "time-1". The change $\Delta\Pi_{it}(\boldsymbol{\chi})$ in equation (11) is then defined analogously.[37] Similarly, we define the instrument for firm i in year t as $Z_{it} \equiv D_{i,t} - D_{i,t-2}$. Further, since our test can be performed on any set of changes, we stack all five such changes available to us (given our data from 2008 to 2015) and perform the joint test that the null hypothesis of U-AEP is true throughout our dataset. This joint test can then be implemented via a pooled version of the procedure described above, estimating a quantile regression of $\Delta\Pi_{it}(\boldsymbol{\chi})$ on Z_{it} using all available firm-year observations.

Finally, we note one measurement challenge that arises in the context of our data source. We see changes in the values of firms' outputs and inputs, but not the corresponding changes in quantities (valued at initial prices) called for in the definition of $\Delta\Pi_i(\boldsymbol{\chi})$. We therefore use, for example, the change in labor costs (i.e., $\Delta(w_m x_{im})$, where input m corresponds to

[37] For example, the component $\bar{w}_m \Delta x_{im}$ in $\Delta\Pi_i(\boldsymbol{\chi})$ is defined as $[\frac{1}{2}(\bar{w}_{m,t-1} + \bar{w}_{m,t-2})][\frac{1}{2}(x_{im,t} + x_{im,t+1}) - \frac{1}{2}(x_{im,t-1} + x_{im,t-2})]$.

labor) as a proxy for the change in employment evaluated at initial wages (i.e., $\bar{w}_m \Delta x_{im}$), which means that the pure price-change component (i.e., $\bar{x}_{im} \Delta w_m$) will appear in ε_i in equation (10). However, as discussed in Section 4.1, multiple pieces of evidence suggest that the demand shocks we use to construct Z_{it} do not cause appreciable output or input price changes, so the violation of Assumption 3 caused by this measurement issue is likely to be small.

5.2 Results

Figure 8 reports the p-value of the test described above when $\Delta\Pi_{it}(\boldsymbol{\chi})$ is evaluated at a range of values of $\boldsymbol{\chi}$. However, we confine attention to the case in which all inputs m have the same value of χ_m (and denote this common value by the scalar χ) so that one dimension of the $\boldsymbol{\chi}$ space can be easily illustrated. Beginning with the test for U-AEP ($\boldsymbol{\chi} = \mathbf{1}$), this corresponds to the point $\chi = 1$ on the x-axis. The p-value at this point is extremely low (below 0.001), which indicates that the null of unconstrained allocative efficiency is resoundingly rejected by these data. This implies that the treatment effect of demand-driven cost increases on sales is (at standard levels of significance) different from one, a finding that was already anticipated (qualitatively) by the discussion in Section 4.1.

The previous result has demonstrated that the firms in our sample do not operate at the allocation of U-AEP—namely, one where firms not only have the same value marginal products of each input but where, further, any input's value marginal product is equal to its price and hence all wedges $\bar{\mu}_{ij,m}$ are equal to one. However, the null of C-AEP allows for the possibility that firms do have identical wedges for any given input, but where the value of that common wedge is not necessarily equal to one (i.e., $\bar{\mu}_{ij,m} = \chi_m$ for all i and j). Such a finding would be consistent with the allocation chosen by a hypothetical planner who faces a constraint on the aggregate amount of each input that can be allocated across these firms.

We now examine this wider possibility. Doing so in principle involves conducting our previous test at every value of χ_m, separately for each input type m, and checking whether there is a value of χ_m (potentially a separate one for each input type) at which the test does not reject. Figure 8 shows results from a partial version of this multi-dimensional search through the space of $\boldsymbol{\chi}$, in which χ_m is common (i.e., $\chi_m = \chi$) for all inputs. As it turns out,

even this partial search reveals a value at which the test does not reject at standard levels of statistical significance, and hence we conclude that the null of C-AEP cannot be rejected. That is, while we know from the U-AEP test that we would reject the null of C-AEP at the point $\chi = 1$, as we explore other values of χ (in the range of 0.9 to 1.4) the p-value of our randomization inference test peaks sharply at the value of $\chi = 1.140$ (where the p-value is 0.35). The null of constrained allocative efficiency is therefore not rejected in this setting.

Again, the results from this formal test for C-AEP were informally anticipated by the results in Section 4.2. There we saw how both input and output responses to demand shocks displayed striking similarities across groups based on pre-treatment characteristics, such as size. Relative to those earlier findings, the benefits of the result in Figure 8 are that: (a) it allows for sampling variation and delivers a formal p-value; (b) it tests for cross-firm wedge heterogeneity even when firms' wedges are allowed to differ along arbitrary, unobserved dimensions; and (c) it evaluates heterogeneity in the ratio of firms' output responses to their input responses, as is consistent with the definition of the value marginal product, rather than exploring heterogeneity in output responses and input responses separately.

6 The Cost of Misallocation

We have seen in the previous section that our test does reject the null of U-AEP in the context of Ecuador's construction sector. This result tells us that the extent of misallocation is large enough to trigger rejection, but it does not tell us how large the cost of misallocation is in economic terms. Equally, we have seen that our test for C-AEP does not reject, but it is possible that this reflects a lack of power to detect meaningful departures from constrained efficiency. For these reasons, our final analysis goes beyond testing and aims to arrive at an estimate of the total welfare cost of misallocation—that is, the cost of departures from U-AEP—in our context. We shall also provide a decomposition of this total cost into components stemming from (a) departures from C-AEP, which holds constant the aggregate inputs $\bar{X}_m$, and (b) the misallocation of these aggregate input levels themselves.

6.1 Assumptions and Parameter Choices

As described in Proposition 2, the joint distribution of wedges across firms, products, and inputs (weighted by any set of observable weights) is nonparametrically identified given an appropriate set of instrumental variables Z_i. And as outlined in Section 2.5, moments of these weighted distributions can be used to provide critical inputs into calculations about the cost of misallocation, as in the simple examples discussed there. We now execute one such set of calculations for our empirical context. In particular, we draw on the second example in Section 2.5, with the total cost of misallocation as given by equation (18). But we specialize to a version in which there is a single sector (i.e., we set $\rho = 0$, $\theta_s = \theta$, and $\bar{\psi}_s = 1$), as is relevant to our empirical context where we can only estimate the wedges among firms in the construction services sector. In this case, the total cost of misallocation is equal to

$$\frac{\Delta W}{W} = \frac{1}{2}\bar{\omega}_C(1-\bar{\omega}_C)\eta\left(\mathbb{E}_{\bar{\lambda}}\left[\bar{\mu}\right]-1\right)^2 + \frac{1}{2}\bar{\omega}_C\theta\mathbb{V}ar_{\bar{\lambda}}\left[\bar{\mu}\right]. \tag{21}$$

The two terms in this expression also provide our desired decomposition into two sources of misallocation. The second term measures the cost of departures from C-AEP, while holding the aggregate inputs $\bar{X}_m$ constant, which is a function of the sales-weighted variance in wedges across firms. The first term, on the other hand, quantifies the cost of an inefficient total amount of inputs, and is a function of the sales-weighted average of firms' wedges.

Recall that this expression was derived for a setting in which there is assumed to be no within-firm dispersion of wedges across inputs. This is necessary in our application because our demand-driven source of exogenous variation is unsuited to the recovery of such within-firm dispersion.[38] It is therefore reassuring that misallocation due to within-firm wedge dispersion is typically found to be relatively small (e.g., about 15% of the estimated total misallocation cost for China obtained by Hsieh and Klenow (2009, p. 1442)).

[38] As Proposition 2 explains, such dispersion can be identified in settings where a separate instrument Z_{im} is available for each type of input, m. It is possible, in principle, to use demand shocks as the source of such input-specific instruments, to the extent that firms' technologies are non-homothetic and the econometrician has access to observable proxies for such features. However, we lack the survey data on firm production required to implement this idea in practice. If wedges are dispersed across inputs and our instrument is uncorrelated with changes in relative input prices then the single-instrument version of our procedure will recover moments of the distribution of each firm's cost share-weighted average of its input-specific wedges.

Calculating the cost of misallocation via equation (21) requires three sets of components. The first is an estimate of the two weighted moments of the wedge distribution, $\mathbb{E}_{\bar{\lambda}}[\bar{\mu}]$ and $\mathbb{V}ar_{\bar{\lambda}}[\bar{\mu}]$. The next subsection describes how we use the IVCRC estimation method, following Proposition 2, to estimate these moments. Second, we require a value for the two elasticity parameters: that across firms in consumption (θ) and that for aggregate input supply to these firms (η). We choose these conservatively, and also in a way that is consistent with prior work. Specifically, we follow Hsieh and Klenow (2009) and set $\theta = 3$; and we set $\eta = 3$, as a value at the upper end of estimates in the literature on labor supply. Finally, equation (21) also depends on $\bar{\omega}_C$, the share of the representative household's shadow income that is spent on consumption of goods. We begin with the midpoint value of $\bar{\omega}_C = 0.5$ but, as we document below, our conclusions are not very sensitive to the value chosen.

6.2 Estimating Moments of the Wedge Distribution

IVCRC Implementation

We now apply the result in Proposition 2 to estimate the two unknown (weighted) moments of the wedge distribution in equation (21): $\mathbb{E}_{\bar{\lambda}}[\bar{\mu}]$ and $\mathbb{V}ar_{\bar{\lambda}}[\bar{\mu}]$. In a setting with no within-firm wedge dispersion, as assumed here, equation (15) reduces to

$$\bar{p}_i \Delta y_i = \bar{\mu}_i \sum_{m \in \mathcal{M}} \bar{w}_m \Delta x_{im} + \varepsilon_i. \tag{22}$$

We begin by applying the procedure described in Section 2.4 to this equation in order to estimate the sales-weighted first moment of the distribution of coefficients $\bar{\mu}_i$ on the single endogenous regressor $\sum_{m \in \mathcal{M}} \bar{w}_m \Delta x_{im}$, while using a single instrument Z_i constructed from lottery-based demand shocks. Provided that the instrument satisfies Assumption 4, a point to which we return below, this delivers a consistent estimate of $\mathbb{E}_{\bar{\lambda}}[\bar{\mu}]$.[39]

[39] As discussed in Section 2.4, we augment the IVCRC procedure to obtain the weighted expectation of the distribution of coefficients $\bar{\mu}_i$ by interacting the regressor with indicators for five bins based on firms' initial-year sales shares (the quintiles from Figure 6(b)). Our sensitivity analysis below explores this choice.

Further, squaring equation (22) yields a second regression specification

$$(\bar{p}_i \Delta y_i)^2 = (\bar{\mu}_i)^2 \left(\sum_{m \in \mathcal{M}} \bar{w}_m \Delta x_{im} \right)^2 + 2\bar{\mu}_i \varepsilon_i \sum_{m \in \mathcal{M}} \bar{w}_m \Delta x_{im} + (\varepsilon_i)^2. \tag{23}$$

Applying our procedure to this equation, using the same instrument, the coefficient on the regressor $\left(\sum_{m \in \mathcal{M}} \bar{w}_m \Delta x_{im}\right)^2$ allows us to obtain a consistent estimate of $\mathbb{E}_{\bar{\lambda}}[(\bar{\mu})^2]$. We then calculate the centered second moment from $\mathbb{V}ar_{\bar{\lambda}}[\bar{\mu}] = \mathbb{E}_{\bar{\lambda}}[(\bar{\mu})^2] - (\mathbb{E}_{\bar{\lambda}}[\bar{\mu}])^2$, though one limitation of this procedure is that it does not impose that estimates satisfy $\mathbb{E}_{\bar{\lambda}}[(\bar{\mu})^2] \geq (\mathbb{E}_{\bar{\lambda}}[\bar{\mu}])^2$ and hence that the estimated variance is positive.

Many details of our implementation are analogous to those from the test procedure described in Section 5. First, because our data does not report prices and quantities separately, we proxy for the price-constant changes in revenues (and costs) in equations (22) and (23) with observed changes in revenues and costs. However, for reasons discussed above, we believe the bias caused by any resulting proxy error is unlikely to be large. Second, while it is possible to estimate equations (22) and (23) using data from any single cross-section of ("time-0" to "time-1") changes, we use the stacked set of changes from 2008-15, along with the variable construction based on changes in two-year averages, as described in Section 5.1.

A third detail that is specific to IVCRC is the need to choose a bandwidth and kernel, which are used in a sub-component of the estimation routine that obtains a smoothed estimate of the expected value of the coefficient at each rank of the conditional first-stage distribution. Our baseline analysis follows the defaults in Benson et al. (2022) and uses the rule-of-thumb bandwidth proposed by Fan and Gijbels (1996) and an Epanechnikov kernel, but we explore alternatives below. Fourth, because the IVCRC estimation problem cannot be represented as a sharp null hypothesis, randomization inference approaches (such as that applied in Section 5) cannot be used for confidence intervals. We therefore apply a block-bootstrap procedure (with 100 bootstrap samples created by drawing firms with replacement) to assess the uncertainty in our estimates. Fifth, because we expect the quadratic form in equation (23) to be sensitive to outliers, we trim the sample at the 2% most extreme values of the ratio of (in absolute value) change in revenues to change in costs; but we again report a range of alternatives below and believe this to be a conservative approach.

Finally, as discussed in Section 2.4, the Masten and Torgovitsky (2016) method for estimating moments of coefficient distributions in IVCRC models relies on access to an instrument that satisfies Assumption 4. The instrument Z_{it}, built from lottery-driven demand shocks, plausibly satisfies the independence and relevance components (a) and (c) of this assumption. But it is not *a priori* clear whether it satisfies component (b), first-stage rank-invariance. To explore the accuracy of this assumption, Appendix D presents a simulated version of the economy assumed here, but where firms have a known technology (which, following Hsieh and Klenow (2009), we assume exhibits constant returns-to-scale). We then calibrate this simulation to match our data and calculate the extent of first-stage rank reversals in response to lottery-driven demand shocks, according to the metric proposed by Gollin and Udry (2021). This simulation demonstrates that such reversals are rare—a result that follows because firm heterogeneity in input growth due to non-lottery factors is much greater than that which results from any given amount of lottery-driven winnings. Hence, violations of Assumption 4(b) seem unlikely to be consequential for our conclusions.[40]

Results

Our baseline IVCRC estimates of $\mathbb{E}_{\bar{\lambda}}[\bar{\mu}]$ and $\mathbb{V}ar_{\bar{\lambda}}[\bar{\mu}]$ are reported in the first row of Table 4 (panel a). We obtain a value of 1.126 for the former (column 1) and 0.014 for the latter (column 2). These values should not be too surprising given the results of our tests in the previous section. We have seen in Section 5 how, at the common wedge value of $\chi = 1.140$, the test of C-AEP fails to reject. So we should expect a value for an average wedge of approximately that χ value, with little variation in wedges around it. And that is exactly what our point estimates of $\mathbb{E}_{\bar{\lambda}}[\bar{\mu}]$ and $\mathbb{V}ar_{\bar{\lambda}}[\bar{\mu}]$ imply.

Table 4 also reports the block-bootstrapped confidence intervals—a two-sided interval in the case of $\mathbb{E}_{\bar{\lambda}}[\bar{\mu}]$, but a one-sided version in the case of $\mathbb{V}ar_{\bar{\lambda}}[\bar{\mu}]$ since we know that the true value of this variance cannot be negative—that we obtain for each of these estimates. The implied uncertainty about $\mathbb{E}_{\bar{\lambda}}[\bar{\mu}]$ is minimal, and we can hence reject an average wedge equal to one at standard levels of statistical significance. But the confidence interval for

[40] As discussed in Section 2.4, an IV that satisfies first-stage rank-invariance when estimating equation (22) will also satisfy it for equation (23). This is because the first regressor is an endogenous variable but one that is derived from a known function of the second regressor, which is itself the only regressor in (22).

$\mathbb{V}ar_{\bar{\lambda}}[\overline{\mu}]$ is considerably wider. To explore this, panel (a) of Appendix Figure A.10 plots the 100 bootstrap sample estimates of $\mathbb{V}ar_{\bar{\lambda}}[\overline{\mu}]$ that underpin this confidence interval. The vast majority of estimates are very low, or even negative, but there is also a tail of large estimates in absolute value. This is perhaps unsurprising given the quadratic nature of the terms in the estimating equation (23).

The next five rows of Table 4 describe the sensitivity of our estimates to alternative choices one could make when carrying out our estimation procedure. For example, rather than trimming at the 2% level, we can either not trim at all (row 2) or do so at the 5% level (row 3). Or, rather than using a Epanechnikov kernel when applying the IVCRC estimator, we can use a Gaussian or uniform kernel (rows 4 and 5). Finally, rather than calculating the weighted moments $\mathbb{E}_{\bar{\lambda}}[\overline{\mu}]$ and $\mathbb{V}ar_{\bar{\lambda}}[\overline{\mu}]$ by using five bins of firm sales $\bar{\lambda}$, we can use ten bins. The estimates of $\mathbb{E}_{\bar{\lambda}}[\overline{\mu}]$ are highly insensitive to these variations, with point estimates ranging from approximately 1.11 to 1.13. The various estimates of $\mathbb{V}ar_{\bar{\lambda}}[\overline{\mu}]$—which we truncate at zero if the point estimate is negative—are more sensitive but are always small in the sense that they are less than or equal to our baseline estimate of 0.014.

6.3 Estimating the Cost of Misallocation

We now use the estimates of $\mathbb{E}_{\bar{\lambda}}[\overline{\mu}]$ and $\mathbb{V}ar_{\bar{\lambda}}[\overline{\mu}]$ in columns (1) and (2) of Table 4 to calculate the total cost of misallocation in our context, following the formula in equation (21) and the choices of parameter values for θ, η and $\bar{\omega}_C$ discussed in Section 6.1. The results are reported in column (3) of Table 4 (panel a). Our baseline point estimate is $\Delta W/W = 0.016$—which implies that the total cost of misallocation is 1.6% of the overall expenditure on the products in our model economy (the goods and services produced by the firms taking part in procurement lotteries, and the leisure consumed by households). Column (3) also reports that the block-bootstrapped 95% one-sided confidence interval on our estimate for the cost of misallocation $\Delta W/W$ spans the range from zero to 26.1%, so we cannot rule out considerably larger values at standard levels of confidence. However, as with the case of $\mathbb{V}ar_{\bar{\lambda}}[\overline{\mu}]$ discussed above, the distribution of bootstrap estimates (reported in panel (b) of Appendix Figure A.10) demonstrates that this result is highly affected by the tails of this distribution; for example, 80% of the bootstrap values of $(\Delta W/W)$ fall below 7%.

The remaining rows of Table 4 (panel a) describe a range of sensitivity checks on our conclusions about the total cost of misallocation $\Delta W/W$. The first five rows calculate the consequences for $\Delta W/W$ of the sensitivity checks on $\mathbb{E}_{\bar{\lambda}}[\overline{\mu}]$ and $\mathbb{V}ar_{\bar{\lambda}}[\overline{\mu}]$ discussed earlier. Two additional rows consider alternative values of $\bar{\omega}_C$ (of 0.75 and 0.25) centered around our baseline value (of 0.5). These seven alternatives suggest that our estimation choices have been conservative. The only exception is $\bar{\omega}_C = 0.75$, but even setting this parameter to its maximal value of $\bar{\omega}_C = 1$ implies losses from misallocation equal to just 2.1%.

As discussed above, the total cost of misallocation $\Delta W/W$ can be decomposed into two components due to: (a) departures from C-AEP while holding constant the availability of aggregate inputs, the second term in equation (21); and (b) misallocation of the aggregate input amounts themselves, the first term in equation (21). Component (b) can be calculated by a transformation of the values in column (1). Doing so, we estimate it to be 0.6% throughout the alternatives that we have explored. And component (a), which is a transformation of the values in column (2), ranges across alternatives from zero to 1.5%.

Put together, the collection of estimates in Table 4 (panel a) implies that the allocation of inputs in our context, both to firms and across them, appears to be close to the efficient point. However, in interpreting this finding, it is important to recall that it refers to the efficiency of the actual allocation, rather than the mechanisms through which this allocation arises. For example, given active government involvement in this sector, our findings do not necessarily imply that a *laissez-faire* policy stance would achieve near-efficiency.

6.4 Comparison to a Parametric Alternative

The results in Section 6.3 may seem surprising, especially when compared to previous work. One possibility is that firms in Ecuador's construction services sector are simply different from those in other contexts. However, as discussed in the Introduction, existing methods typically assume that firms use similar production technologies, which could create a biased impression of misallocation—especially the heterogeneity in wedges inherent to departures from C-AEP—if firms' actual technologies are more heterogeneous than assumed.

One way of seeing this is to start from the definition of wedges in equation (3) when applied to a single-product firm whose production function is $y_i = \widetilde{F}^{(i)}(\mathbf{x}_i)$. Then, multiplying

both sides by $(\bar{w}_m\bar{x}_{im})/(\bar{p}_i\bar{y}_i)$, summing across all inputs $m \in \mathcal{M}$, and focusing on the case with no intra-firm wedge dispersion (i.e., $\bar{\mu}_{i,m} = \bar{\mu}_i$ for all m), we obtain

$$\bar{\mu}_i = \left(\frac{\bar{p}_i\bar{y}_i}{\sum_{m\in\mathcal{M}}\bar{w}_m\bar{x}_{im}}\right)\left(\sum_{m\in\mathcal{M}}\frac{\bar{x}_{im}}{\bar{y}_i}\frac{\partial\widetilde{F}^{(i)}(\bar{\mathbf{x}}_i)}{\partial x_{i,m}}\right) = \left(\frac{\bar{p}_i\bar{y}_i}{\sum_{m\in\mathcal{M}}\bar{w}_m\bar{x}_{im}}\right)\gamma_i(\bar{\mathbf{x}}_i), \tag{24}$$

where $\gamma_i(\bar{\mathbf{x}}_i) \equiv (\lambda/\bar{y}_i)(\partial\widetilde{F}(\lambda\bar{\mathbf{x}}_i)/\partial\lambda)$ is firm i's scale elasticity at $\bar{\mathbf{x}}_i$. That is, firm i's wedge is simply equal to the product of its profit margin (the ratio of total sales, $\bar{p}_i\bar{y}_i$, to total costs, $\sum_{m\in\mathcal{M}}\bar{w}_m\bar{x}_{im}$) and its scale elasticity.

Intuitively, knowledge of the scale elasticity $\gamma_i(\bar{\mathbf{x}}_i)$ allows an analyst to convert estimates of average products (such as the profit margin) into estimates of marginal products (which is what wedges depend on). Similarly, when scale elasticities are assumed to be common across firms, then any cross-firm dispersion in average products is mapped one-to-one into conclusions about dispersion in marginal products across firms, and hence about misallocation. A common version of this is the assumption that all firms use technologies that are globally constant returns-to-scale, or $\gamma_i(\bar{\mathbf{x}}_i) = 1$ at all $\bar{\mathbf{x}}_i$, a prominent example of which appears in Hsieh and Klenow (2009).[41]

Table 4 panel (b) explores the implications of such parametric restrictions in our context. We do so by imposing the assumption that $\gamma_i(\bar{y}_i, \bar{\mathbf{x}}_i) = \gamma$ for all firms i. This continues to allow for firms' technologies to combine inputs in arbitrarily heterogeneous ways, but requires that firms do share a common scale elasticity.[42] We begin with the constant returns-to-scale case in which $\gamma = 1$. On the basis of this assumption, computing the wedge $\bar{\mu}_i$ for each firm is a straightforward application of equation (24): each firm's wedge $\bar{\mu}_i$ is simply equal to its profit margin. We then calculate the sales-weighted moments $\mathbb{E}_{\bar{\lambda}}[\bar{\mu}]$ and $\mathbb{V}ar_{\bar{\lambda}}[\bar{\mu}]$ of the distribution of such wedges across all firms in 2008.[43] Row 1 reports these estimates, along

[41] This assumption is stronger than those invoked to arrive at the estimates in Table 4 (panel a), in two respects. First, our estimates of moments of the wedge distribution in columns (1) and (2) require only that firms use locally differentiable technologies, an assumption that has no direct relation to the degree of returns-to-scale. Second, the formula for $\Delta W/W$ in equation (21) used in column (3) invokes the assumption that all firms have *locally* constant returns-to-scale technologies, which has no direct relation to the notion of globally constant returns-to-scale; for example, the local version allows for technologies with arbitrary overhead costs whereas the global version rules out overhead costs altogether.

[42] Formally, this requires that $\widetilde{F}^{(i)}(\bar{\mathbf{x}}_i) = G(g^{(i)}(\bar{\mathbf{x}}_i))$ where $G(\cdot)$ is homogeneous of degree γ and $g^{(i)}(\cdot)$ is homogeneous of degree one but otherwise arbitrary.

[43] Following Hsieh and Klenow (2009), these calculations remove the smallest and largest 1% of wedges.

with their bootstrapped (two-sided) 95% confidence intervals.

A first clear message is that the assumption of constant returns-to-scale does not substantially affect our estimate of the first moment $\mathbb{E}_{\bar{\lambda}}[\bar{\mu}]$: this value rises, but only to 1.258 (in panel b), from our preferred value of 1.126 (in panel a). However, a second clear message is that assuming constant returns does substantially affect our estimate of the second moment, $\mathbb{V}ar_{\bar{\lambda}}[\bar{\mu}]$. This rises from 0.014 in panel (a) to 0.850 in row 1 of panel (b). As a result, the estimated total cost of misallocation in column (3) grows considerably (from 1.6% to 66.2%) as a result of the assumption that all firms use constant returns-to-scale technologies.[44] The bootstrapped confidence interval on this estimate ranges from 52.7% to 135.8%, which does not overlap with the corresponding interval implied by our preferred procedure.

The remaining rows of Table 4 explore how these conclusions change with values of $\gamma \neq 1$. Row 2 uses one corresponding to decreasing returns ($\gamma = 0.85$) and row 3 uses one of increasing returns ($\gamma = 1.15$). As is clear from equation (24), given any set of data on firms' profit margins, estimated wedges increase with the assumed value of γ. Unsurprisingly, therefore, the estimates in row 1 are straddled by those in rows 2 and 3, with the total estimated cost of misallocation ranging from 46.2% in row 2 to 91.8% in row 3. But even the lowest of these remains many times larger than our preferred estimate of 1.6%.

Put together, the findings in Table 4 suggest that the firms in our setting do have heterogeneous degrees of scale economies, and hence that a measurement approach that restricts technological heterogeneity by assumption would infer more misallocation from the data in this setting than appears to be correct. Assumptions about firm technologies (such as a common scale elasticity) are indispensable for learning each firm's wedge individually. But such assumptions may be overly restrictive when, instead, one is merely interested in moments of the distribution of firms' wedges, as is typically sufficient for the study of misallocation.

To compute sales share-weighted moments in panel (b) we use the same 2008 sales shares as in panel (a).

[44] We note, however, that equation (21) was derived under the approximation that wedges are close to one. If, instead, we derive this expression under the approximation that *log* wedges are close to zero, then the cost of misallocation would be given by $\left(\frac{\Delta W}{W}\right)^{log} = \frac{1}{2}\bar{\omega}_C(1-\bar{\omega}_C)\eta\left(\mathbb{E}_{\bar{\lambda}}[\ln\bar{\mu}]\right)^2 + \frac{1}{2}\bar{\omega}_C\theta\mathbb{V}ar_{\bar{\lambda}}[\ln\bar{\mu}]$. Applying this alternative formula to wedges inferred from equation (24) under the constant returns assumption ($\gamma = 1$), we estimate a considerably lower value of $(\Delta W/W)^{log} = 6.9\%$, with the reduction due primarily to the fact that the inferred value for $\mathbb{V}ar_{\bar{\lambda}}[\ln\bar{\mu}]$ is much lower than that for $\mathbb{V}ar_{\bar{\lambda}}[\bar{\mu}]$. Of course, if wedges are truly close to one, as our preferred estimates imply, then the distinction between the approximation based on wedges close to one and that based on log wedges close to zero would be minuscule.

7 Conclusion

In this paper we have developed new tools for assessing the allocative efficiency of production, both in the absence and presence of an aggregate input constraint, among any given set of firms. We have developed new procedures that estimate features of the distribution of wedges—ratios of the value marginal product of an input divided by its price—across all firms, products, and inputs in an economy. By drawing on sources of exogenous variation in firm input changes, these methods proceed without the need to specify production functions, demand functions, or the underlying nature of the distortions that lead to potential inefficiency. Instead, they simply seek to estimate the "treatment effects" of (price-adjusted) inputs on outputs across firms. This then allows for a comparison of such treatment effects to an idealized efficient allocation, in which they would not differ across input uses.

Our results imply that the firms in our context, Ecuador's construction sector, produce at an allocation that is strikingly close to allocative efficiency. This holds both in terms of firms' relative use of given aggregate inputs and in terms of the amount of the aggregate input levels themselves. Our analysis would arrive at a different conclusion if we were to apply commonly-used parametric approaches that assume firms use technologies with similar features (such as a common degree of returns-to-scale). This should come as no surprise given that the essence of our approach has been to allow for technological heterogeneity that could create misleading impressions of misallocation if it were ignored.

While our methods rely on researchers' access to exogenous variation in firm input use, such as the lottery-based demand shocks at our disposal, we are optimistic about the availability of requisite variation in other contexts. The studies cited in the Introduction highlight a wide range of examples of such variation coming from either the output demand side (e.g., government demand, foreign demand, or competition shocks that affect residual demand) or the input supply side (e.g., subsidized inputs, bank expansions, or immigration). Our procedures would be just as applicable to those and other ideas as they are to the setting of Ecuador's procurement lottery system.

References

Adao, Rodrigo, Paul Carrillo, Arnaud Costinot, Dave Donaldson, and Dina Pomeranz, "Imports, Exports, and Earnings Inequality: Measures of Exposure and Estimates of Incidence," *The Quarterly Journal of Economics*, 2022, *137* (3), 1553–1614.

Banerjee, Abhijit V and Esther Duflo, "Do Firms Want to Borrow More? Testing Credit Constraints Using a Directed Lending Program," *Review of Economic Studies*, 2014, *81* (2), 572–607.

Baqaee, David Rezza and Emmanuel Farhi, "Productivity and Misallocation in General Equilibrium," *The Quarterly Journal of Economics*, 2020, *135* (1), 105–163.

Bau, Natalie and Adrien Matray, "Misallocation and Capital Market Integration: Evidence from India," *Econometrica*, 2023, *91* (1), 67–106.

Beerli, Andreas, Jan Ruffner, Michael Siegenthaler, and Giovanni Peri, "The Abolition of Immigration Restrictions and the Performance of Firms and Workers: Evidence from Switzerland," *American Economic Review*, 2021, *111* (3), 976–1012.

Benson, David, Matthew A Masten, and Alexander Torgovitsky, "ivcrc: An Instrumental-Variables Estimator for the Correlated Random-Coefficients Model," *The Stata Journal*, 2022, *22* (3), 469–495.

Bergquist, Lauren Falcao and Michael Dinerstein, "Competition and Entry in Agricultural Markets: Experimental Evidence from Kenya," *American Economic Review*, 2020, *110* (12), 3705–3747.

Bloom, Nicholas, Benn Eifert, Aprajit Mahajan, David McKenzie, and John Roberts, "Does Management Matter? Evidence from India," *The Quarterly Journal of Economics*, 2013, *128* (1), 1–51.

Borusyak, Kirill and Peter Hull, "Non-Random Exposure to Exogenous Shocks: Theory and Applications," NBER Working Paper No. 27845, 2021.

Brugués, Felipe, Javier Brugués, and Samuele Giambra, "Political Connections and Misallocation of Procurement Contracts: Evidence from Ecuador," STEG Working Paper, 2022.

Busso, Matias and Sebastian Galiani, "The Causal Effect of Competition on Prices and Quality: Evidence from a Field Experiment," *American Economic Journal: Applied Economics*, 2019, *11* (1), 33–56.

de Mel, Suresh, David McKenzie, and Christopher Woodruff, "Returns to Capital in Microenterprises: Evidence from a Field Experiment," *The Quarterly Journal of Economics*, 2008, *123* (4), 1329–1372.

_ , _ , **and** _ , "Labor Drops: Experimental Evidence on the Return to Additional Labor in Microenterprises," *American Economic Journal: Applied Economics*, 2016, *11* (1), 202–235.

Ding, Peng, Avi Feller, and Luke Miratrix, "Randomization Inference for Treatment Effect Variation," *Journal of the Royal Statistical Society: Series B: Statistical Methodology*, 2016, pp. 655–671.

Doran, Kirk, Alexander Gelber, and Adam Isen, "The Effects of High-Skilled Immigration Policy on Firms: Evidence from Visa Lotteries," *Journal of Political Economy*, 2022, *130* (10), 2501–2533.

Fadic, Milenko, "Letting Luck Decide: Government Procurement and the Growth of Small Firms," *The Journal of Development Studies*, 2020, *56* (7), 1263–1276.

Fan, Jianqing and Irene Gijbels, *Local Polynomial Modelling and its Applications: Monographs on Statistics and Applied Probability*, Vol. 66, CRC Press, 1996.

Felix, Mayara, "Trade, Labor Market Concentration, and Wages," Working Paper, 2021.

Ferraz, Claudio, Frederico Finan, and Dimitri Szerman, "Procuring Firm Growth: The Effects of Government Purchases on Firm Dynamics," NBER Working Paper No. 21219, 2015.

Fisher, RA, *The Design of Experiments*, Oliver and Boyd, Edinburgh, 1935.

Giorcelli, Michela, "The Long-Term Effects of Management and Technology Transfers," *American Economic Review*, 2019, *109* (1), 121–52.

Gollin, Douglas and Christopher Udry, "Heterogeneity, Measurement Error, and Misallocation: Evidence from African Agriculture," *Journal of Political Economy*, 2021, *129* (1), 1–80.

Hall, Robert E., "The Relation between Price and Marginal Cost in U.S. Industry," *Journal of Political Economy*, 1988, *96* (5), 921–947.

_ , "Using Empirical Marginal Cost to Measure Market Power in the US Economy," NBER Working Paper No. 25251, 2018.

Haltiwanger, John, Robert Kulick, and Chad Syverson, "Misallocation Measures: The Distortion that Ate the Residual," NBER Working Paper No. 24199, 2018.

Hopenhayn, Hugo A, "Firms, Misallocation, and Aggregate Productivity: A Review," *Annual Review of Economics*, 2014, *6* (1), 735–770.

Hsieh, Chang-Tai and Peter J Klenow, "Misallocation and Manufacturing TFP in China and India," *The Quarterly Journal of Economics*, 2009, *124* (4), 1403–1448.

Hvide, Hans K and Tom G Meling, "Do Temporary Demand Shocks Have Long-Term Effects for Startups?," *The Review of Financial Studies*, 2023, *36* (1), 317–350.

Jensen, Robert and Nolan Miller, "Market Integration, Demand and the Growth of Firms: Evidence from a Natural Experiment in India," *American Economic Review*, 2018, *108* (12), 3583–3625.

Kaboski, Joseph P and Robert M Townsend, "A Structural Evaluation of a Large-Scale Quasi-Experimental Microfinance Initiative," *Econometrica*, 2011, *79* (5), 1357–1406.

Klette, Tor Jakob, "Market Power, Scale Economies and Productivity: Estimates from a Panel of Establishment Data," *The Journal of Industrial Economics*, 1999, *47* (4), 451–476.

Kroft, Kory, Yao Luo, Magne Mogstad, and Bradley Setzler, "Imperfect Competition and Rents in Labor and Product Markets: The Case of the Construction Industry," NBER Working Paper No. 27325, 2022.

Lee, Munseob, "Government Purchases, Firm Growth and Industry Dynamics: Quasi-Experimental Evidence from Government Auctions," Working Paper, 2017.

Masten, Matthew A and Alexander Torgovitsky, "Identification of Instrumental Variable Correlated Random Coefficients Models," *Review of Economics and Statistics*, 2016, *98* (5), 1001–1005.

McCaig, Brian and Nina Pavcnik, "Export Markets and Labor Allocation In a Low-Income Country," *American Economic Review*, 2018, *108* (7), 1899–1941.

Restuccia, Diego and Richard Rogerson, "Policy Distortions and Aggregate Productivity with Heterogeneous Establishments," *Review of Economic Dynamics*, 2008, *11* (4), 707–720.

_ and _, "The Causes and Costs of Misallocation," *Journal of Economic Perspectives*, 2017, *31* (3), 151–174.

Rotemberg, Martin, "Equilibrium Effects of Firm Subsidies," *American Economic Review*, 2019, *109* (10), 3475–3513.

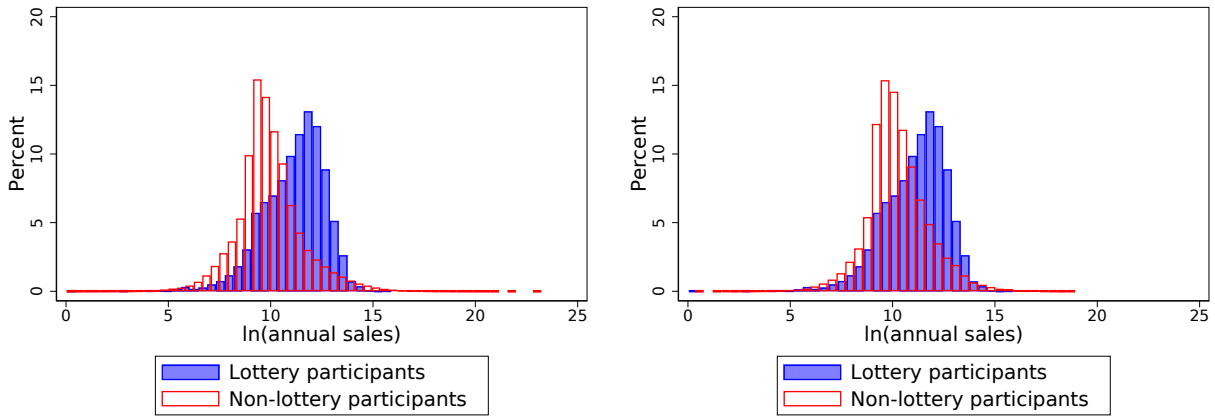
Sraer, David and David Thesmar, "How to Use Natural Experiments to Estimate Misallocation," *American Economic Review*, 2023, *113* (4), 906–938.

Figures and Tables

Figure 1: Comparison of Firm Size Distributions

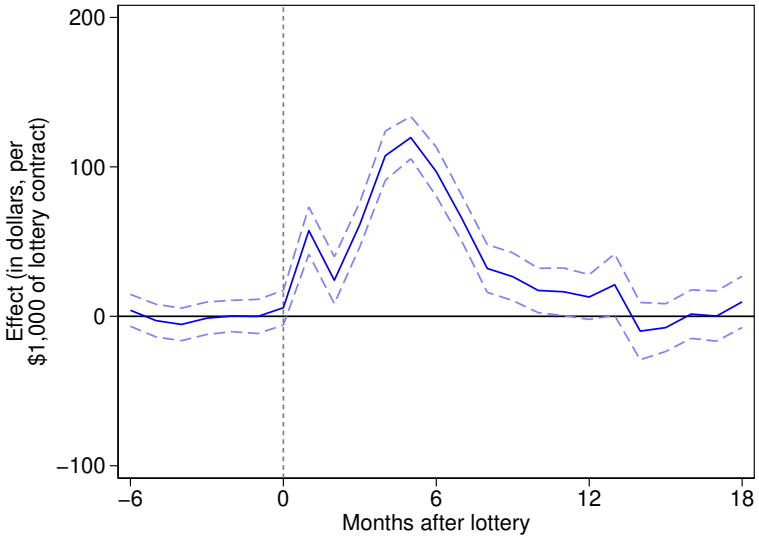
(a) Lottery Participant Firms vs. All Non-Participating Firms

(b) Lottery Participant Firms vs. Non-Participating Firms in Same Industries



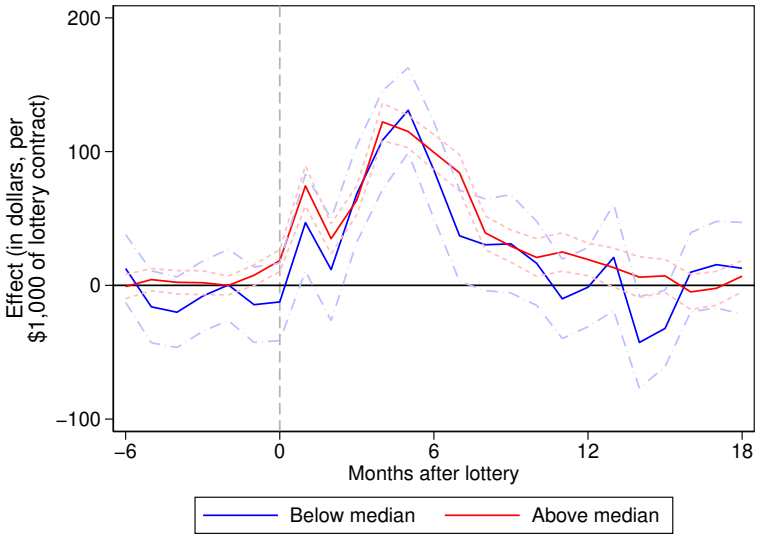
Notes: This figure plots histograms of sales in 2008, the year prior to the start of the procurement lottery system, using firms' annual income tax data (excluding firms with no sales). Lottery participants are firms that participate in at least one lottery during 2009–2014. Non-participant firms are all other firms that were economically active in 2008. Panel (b) restricts the non-participant sample to firms in the construction or engineering industries.

Figure 2: Effects on Total Sales



Notes: This figure plots estimates of the monthly effects of an additional $1,000 in procurement winnings shocks on total (third-party reported) sales following equation (20). Total sales are based on monthly purchase annexes reported by client entities' VAT filings. Dashed lines indicate 95% confidence intervals that allow for clustering at the firm level.

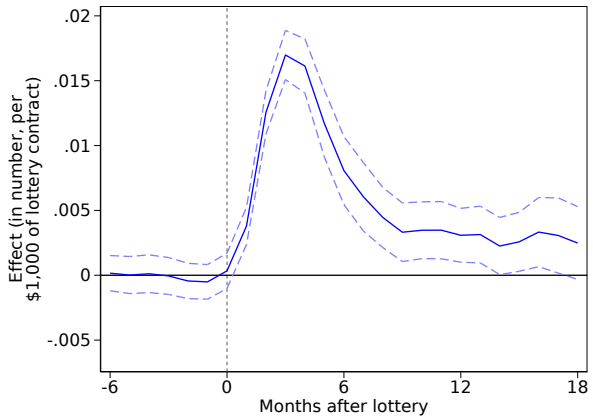
Figure 3: Effects on Total Sales by Contract Size



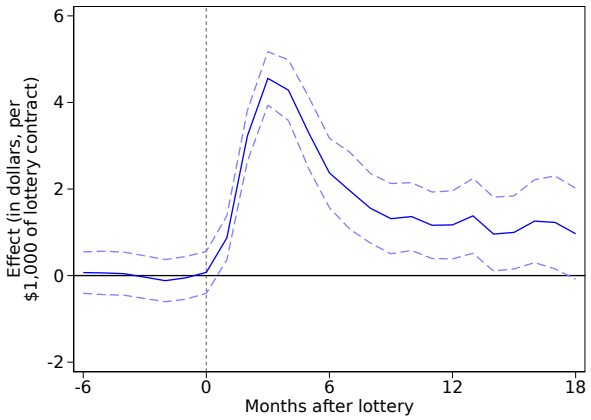
Notes: This figure extends the analysis of Figure 2, estimating monthly effects of an additional $1,000 in procurement winnings shocks on total sales separately for lotteries with large vs. small contracts (below- vs. above-median contract amount), following equation (20). Total sales are based on monthly purchase annexes reported by clients entities' VAT filings. Dashed lines indicate 95% confidence intervals that allow for clustering at the firm level.

Figure 4: Effects on Employment

(a) Number of Employees

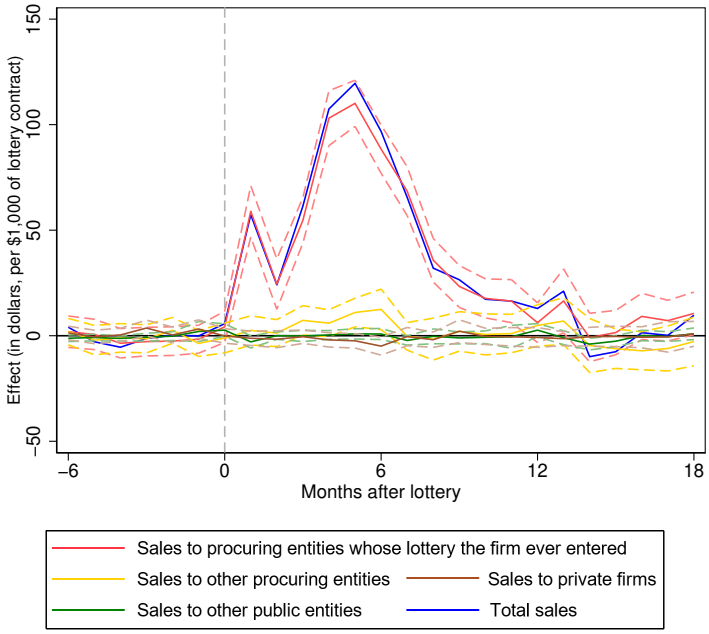


(b) Total Wage Payments



Notes: This figure plots estimates of the monthly effects of an additional $1,000 in procurement winnings shocks on employment following equation (20), using data from social security records. Panel (a) presents effects on the number of employees, and panel (b) on total wages paid. Dashed lines indicate 95% confidence intervals that allow for clustering at the firm level.

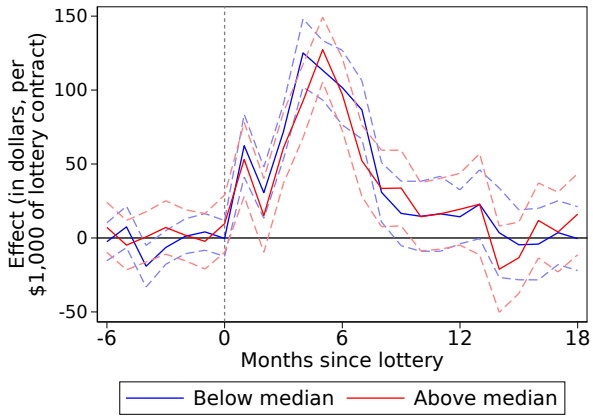
Figure 5: Effects on Sales to Different Types of Clients



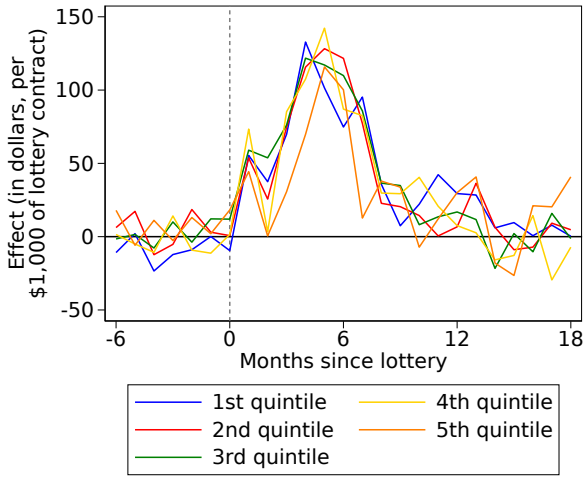
Notes: This figure extends the analysis of Figure 2, estimating the monthly effects of an additional $1,000 in procurement winnings shocks on sales to mutually-exclusive categories of clients, following equation (20). These clients are: procuring entities with at least one lottery in our study period that the firm participated in (in red); other procuring entities, i.e., other entities that made at least one purchase through the lottery system in our study period (yellow); other public entities that made no purchases through the lottery system (green); and private firms (brown). All sales measures are based on monthly purchase annexes reported by client entities' VAT filings. Dashed lines 95% confidence intervals that allow for clustering at the firm level.

Figure 6: Effects on Total Sales by Firm Size

(a) Above vs. Below Median



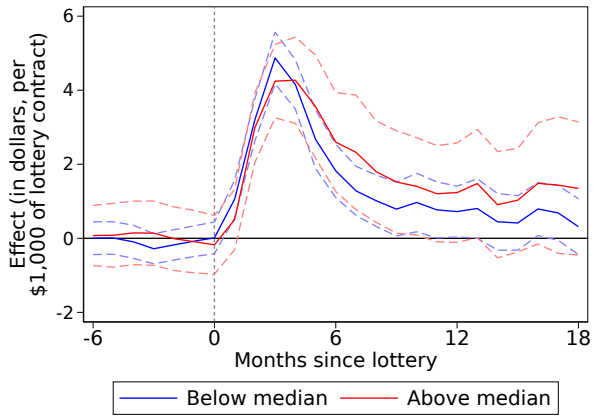
(b) Quintiles



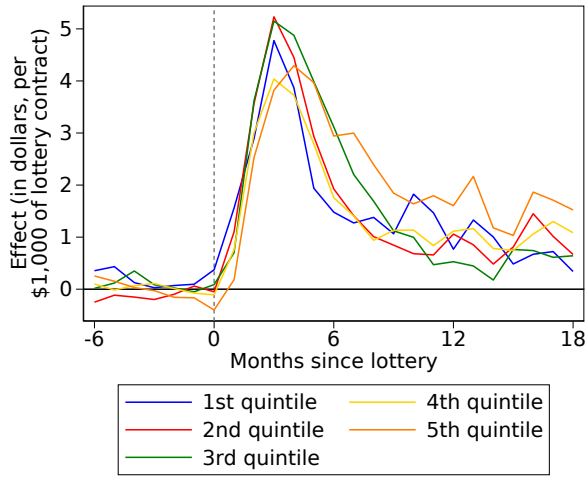
Notes: This figure extends the analysis of Figure 2, estimating the monthly effects of an additional $1,000 in procurement winnings shocks on total sales by firm size, following equation (20). Total sales are based on monthly purchase annexes reported by client entities' VAT filings. Panel (a) presents estimates for firms with above- and below-median total sales prior to the start of the lottery system (i.e., in 2008), based on firms' annual income tax filings. Panel (b) shows the same but partitioning the sample by quintiles of 2008 sales. Dashed lines in panel (a) indicate 95% confidence intervals that allow for clustering at the firm level.

Figure 7: Effects on Total Wage Payments by Firm Size

(a) Above vs. Below Median

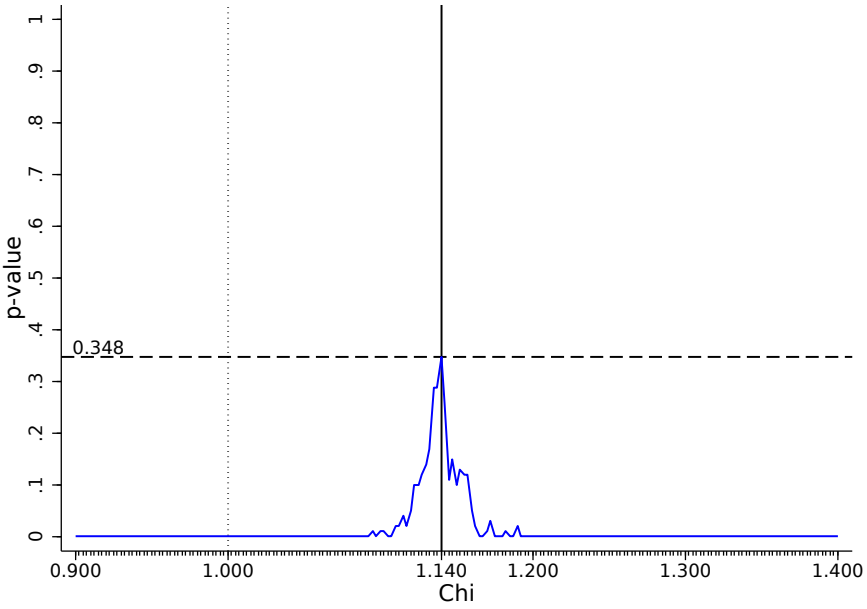


(b) Quintiles



Notes: This figure extends the analysis of Figure 4 Panel (b), estimating the monthly effects of an additional $1,000 in procurement winnings shocks on total wage payments by firm size, following equation (20), using data from social security records. Panel (a) presents estimates for firms with above- and below-median total sales prior to the start of the lottery system (i.e., in 2008), based on firms' annual income tax filings. Panel (b) shows the same but partitioning the sample by quintiles of 2008 sales. Dashed lines in panel (a) indicate 95% confidence intervals that allow for clustering at the firm level.

Figure 8: Randomization Inference Test for U-AEP and C-AEP



Notes: This figure shows the histogram of results of the randomization inference test for unconstrained allocative efficiency of production (U-AEP) and constrained allocative efficiency of production (C-AEP), as described in Sections 2 and 5. The figure plots the p-values (y-axis) under the null of $\boldsymbol{\chi} = \chi$ (i.e., $\chi_m = \chi$ for all inputs m), for different values of χ (x-axis). The dotted vertical line indicates the null value of $\chi = 1$ used for the test for U-AEP, whereas the solid vertical line indicates the value of χ that corresponds to the highest p-value (0.35) obtained in this range of χ. These results imply that the null of U-AEP is rejected at standard levels, but that of C-AEP is not.

Table 1: Summary Statistics

	Mean	Median	Std. Dev.	N
Lotteries				
Contract amount (USD)	46,523	31,597	40,989	18,474
Contract anticipated duration (days)	64.5	60	34.5	18,467
Number of participants	10.1	4	15.7	18,474
Firms' Lottery Participation				
Lotteries entered per year	3.5	1	7.2	9,393
Lotteries won per year	0.35	0	0.55	9,393
Firms' Characteristics				
Firm age (years)	11.21	10	11.06	9,393
Is incorporated	0.18	0	0.38	9,393
Number of clients (third-party reported)	4.74	2	17.72	9,393
Sales (third-party reported, USD)	132,707	47,651	271,306	9,393
Sales (self-reported, USD)	141,184	53,360	287,409	9,393
Costs (self-reported, USD)	123,907	42,058	267,508	9,393
Profits (self-reported, USD)	17,278	11,180	36,360	9,393
Employees (social security)	4.40	2	10.10	9,393
Wages (social security, USD)	7,399	2,880	25,926	9,393

Notes: This table presents summary statistics of procurement lotteries (with at least two participating firms) and of firms that participated in any such lotteries. Observations concerning firm characteristics are from each firm's first year of lottery participation. Self-reported tax variables are based on firms' annual income tax filings. Third-party reported variables are based on (annualized) monthly purchase annexes of client entities' VAT filings. 'Social security' indicates data from social security filings. Contract duration is missing for seven observations, as discussed in Appendix B.

Table 2: Balance of Randomization

Coefficient on $1000 in procurement winnings shocks	
Firm Characteritistics	
Firm age (years)	-0.0004
	(0.0012)
Is incorporated	0.0001
	(0.0000)
Lottery Variables $(t-1)$	
Lotteries entered	0.0048
	(0.0033)
Expected winnings	3.47
	(6.98)
Lotteries won	0.0003
	(0.0002)
Amount won	-6.02
	(8.56)
Social Security Data $(t-1)$	
Employees	0.0001
	(0.0014)
Wages	0.3001
	(1.97)
Tax Data $(t-2)$	
Sales (third-party reported)	-12.59
	(32.57)
Sales to procuring entities (third-party reported)	-18.10
	(27.65)
Sales to other government entities (third-party reported)	2.39
	(2.22)
Sales to private sector (third-party reported)	7.16
	(9.79)
Sales (self-reported)	-17.51
	(33.92)
Costs (self-reported)	-13.60
	(30.80)
Annual income tax (self-reported)	-0.4017
	(0.4814)
Number of clients (third-party reported)	0.0009
	(0.0018)
P-value of joint F-test:	0.29

Notes: This table presents balance tests of procurement winnings shocks (in $1,000s) on pre-treatment variables. Each displayed coefficient (and corresponding standard error in parentheses) stems from a separate regression of the outcome variable on procurement winnings shocks in year (t). For time-varying variables, we use data from year $(t\text{-}1)$. For tax variables, we use information for year $(t\text{-}2)$ to ensure that they reflect pre-lottery filings. Self-reported tax variables are based on firms' annual income tax filings. Third-party reported variables are based on monthly purchase annexes of client entities' VAT filings. Monthly variables are summed up over the year. Observations cover 2009-2014 for non-lagged variables, 2009-2013 for (lagged) lottery variables, 2008-2013 for (lagged) social security variables, and 2008-2012 for (double lagged) tax data variables. As in all our analysis, the sample only includes observations for years after the firm started to be economically active. For the joint F-test, to include the whole sample, we impute zero for missing values and add a dummy indicator. All monetary variables in USD, winsorized at the top 1%. Standard errors clustered at the firm level in parentheses. *** = $p < 0.01$, ** = $p < 0.05$, * = $p < 0.1$.

Table 3:
Effects on Sales, Costs, and Profits

	(1) Sales (Third-party reported)	(2) Sales (Self-reported)	(3) Total costs	(4) Labor costs	(5) Non-labor costs	(6) Profits
Procurement Lottery Shocks:						
Year t	438.71***	430.14***	382.47***	14.39**	368.09***	47.67***
	(37.31)	(39.19)	(36.22)	(4.70)	(33.73)	(5.80)
Year $t+1$	269.76***	239.00***	200.57***	11.03	189.54***	38.43***
	(48.70)	(51.18)	(46.49)	(6.15)	(43.05)	(7.00)
Number of observations	53,107	53,107	53,107	53,107	53,107	53,107
Number of firms	9,368	9,368	9,368	9,368	9,368	9,368

Notes: This table shows estimates of the effect of procurement winnings shocks (in \$1,000's) on firms' sales, costs and profits in the year of the shock (t) and the next (t+1), following equation (20). Third-party reported sales are based on monthly purchase annexes of client entities' VAT filings, summed up over the year. All other variables are based on firms' annual income tax filings. Observations are firm-years. Standard errors clustered at the firm level are reported in parentheses. *** = p < 0.01, ** = p < 0.05, * = p < 0.1.

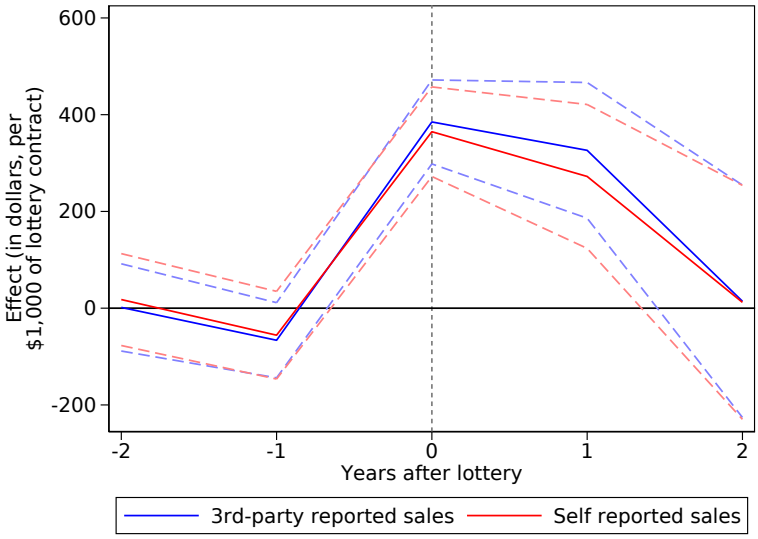
Table 4: Estimated Cost of Misallocation

	$\mathbb{E}_{\bar{\lambda}}[\bar{\mu}]$ (1)	$\mathbb{V}ar_{\bar{\lambda}}[\bar{\mu}]$ (2)	$\frac{\Delta W}{W}$ (3)
Panel (a): IVCRC estimates			
Baseline	1.126	0.014	0.016
	[1.093, 1.161]	[0, 0.341]	[0, 0.261]
No trimming	1.129	0	0.006
	[1.098, 1.188]	[0, 0.329]	[0, 0.253]
5% trimming	1.111	0	0.005
	[1.078, 1.157]	[0, 0.394]	[0, 0.301]
Gaussian kernel	1.125	0	0.006
	[1.095, 1.145]	[0, 0.040]	[0, 0.035]
Uniform kernel	1.126	0.014	0.016
	[1.093, 1.161]	[0, 0.341]	[0, 0.261]
10 sales bins	1.115	0	0.005
	[1.067, 1.158]	[0, 0.617]	[0, 0.468]
$\bar{\omega}_C = 0.75$	1.126	0.014	0.020
	[1.093, 1.161]	[0, 0.341]	[0, 0.387]
$\bar{\omega}_C = 0.25$	1.126	0.014	0.010
	[1.093, 1.161]	[0, 0.341]	[0, 0.132]
Panel (b): Alternative procedure assuming homogeneous technologies			
Constant returns-to-scale ($\gamma_i = 1$)	1.258	0.850	0.662
	[1.240, 1.290]	[0.672, 1.769]	[0.527, 1.358]
Decreasing returns-to-scale ($\gamma_i = .85$)	1.069	0.614	0.462
	[1.054, 1.097]	[0.485, 1.278]	[0.365, 0.962]
Increasing returns-to-scale ($\gamma_i = 1.15$)	1.446	1.124	0.918
	[1.427, 1.484]	[0.888, 2.339]	[0.738, 1.843]

Notes: This table reports, in columns (1) and (2), estimates of the the sales-weighted expectation and variance of the wedge distribution that enter the total cost of misallocation formula in equation (21); reported variance estimates are truncated at zero from below. Column (3) reports the corresponding estimated cost of misallocation $\Delta W/W = \frac{1}{2}\bar{\omega}_C\theta\mathbb{V}ar_{\bar{\lambda}}[\bar{\mu}] + \frac{1}{2}\bar{\omega}_C(1-\bar{\omega}_C)\eta\left(\mathbb{E}_{\bar{\lambda}}[\bar{\mu}]-1\right)^2$ implied by that formula (when using values of the parameters, $\theta = 3$, $\eta = 3$, and $\bar{\omega}_C = 0.5$, unless noted otherwise); these are calculated on the basis of the truncated variance. Panel (a) shows estimates from the IVCRC method described in Section 6.2 using different specifications (with 2% trimming, Epanechnikov kernel, and 5 sales bins, unless noted otherwise). Panel (b) follows an alternative procedure for estimating wedges based on assuming that firms use technologies with common scale elasticities γ_i, as described in Section 6.4. The ranges reported in square brackets are two-sided 95% confidence intervals in column (1) of panel (a) and all of panel (b), but one-sided intervals for columns (2) and (3) of panel (a); all such intervals are calculated on the basis of a block bootstrap procedure, with values reported in Appendix Figure A.10.

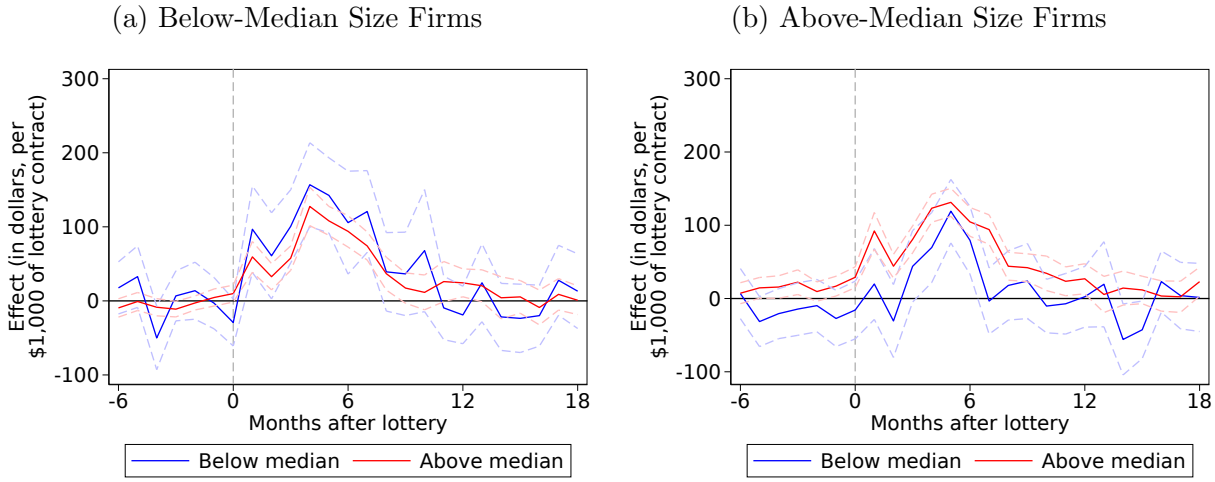
A Appendix Figures

Figure A.1: Annual Total Sales, Third-Party Reported vs. Self-Reported



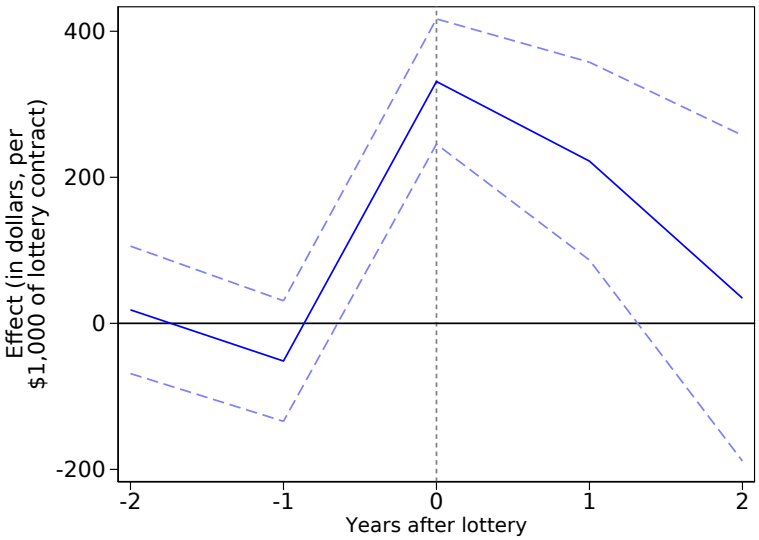
Notes: This figure plots estimates of the annual effect of an additional $1,000 in procurement winnings shocks on total sales, following equation (20). Third-party reported sales are based on monthly purchase annexes reported by the client entities' VAT filings (annualized over the year) and self-reported sales are based on firms' annual income tax filings. Dashed lines indicate 95% confidence intervals that allow for clustering at the firm level.

Figure A.2: Effects on Total Sales by Contract Size and Firm Size



Notes: This figure extends the analysis of Figures 3 and 6, estimating monthly effects of an additional $1,000 in procurement winnings shocks on total sales by firm size, separately for lotteries with large vs. small contracts (below- vs. above-median contract amount), following equation (20). Total sales are based on monthly purchase annexes reported by client entities' VAT filings. Panel (a) presents estimates for firms with below-median sales prior to the start of the lottery system (i.e., in 2008), based on firms' annual income tax filings. Panel (b) shows the same but for firms with above-median self-reported sales. Dashed lines indicate 95% confidence intervals that allow for clustering at the firm level.

Figure A.3: Effects on Total Costs

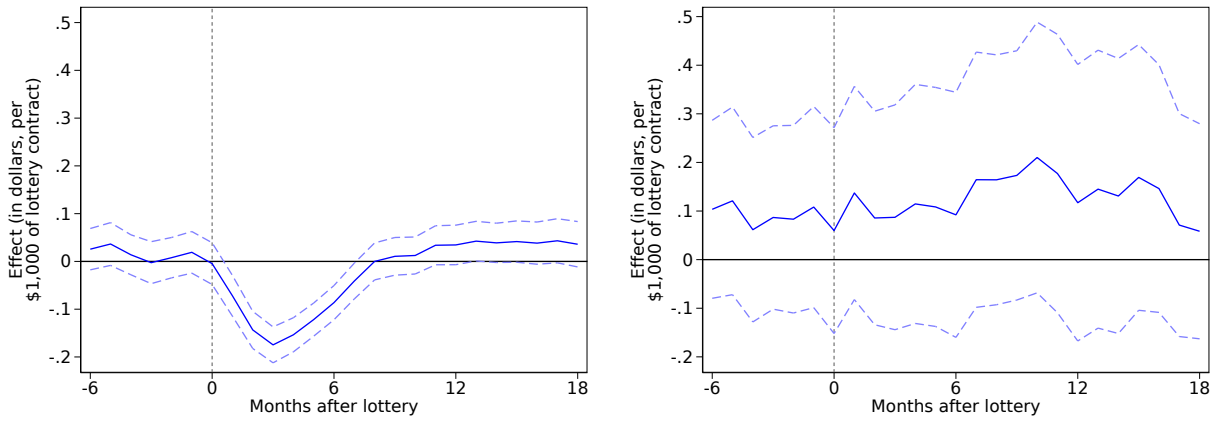


Notes: This figure plots estimates of the annual effects of an additional $1,000 in procurement winnings shocks on total costs, following equation (20). Total costs include both labor and non-labor costs and are based on firms' annual income tax filings. Dashed lines indicate 95% confidence intervals that allow for clustering at the firm level.

Figure A.4: Effects on Wages

(a) Average Monthly Wage
All Employees

(b) Average Monthly Wage
Existing Employees

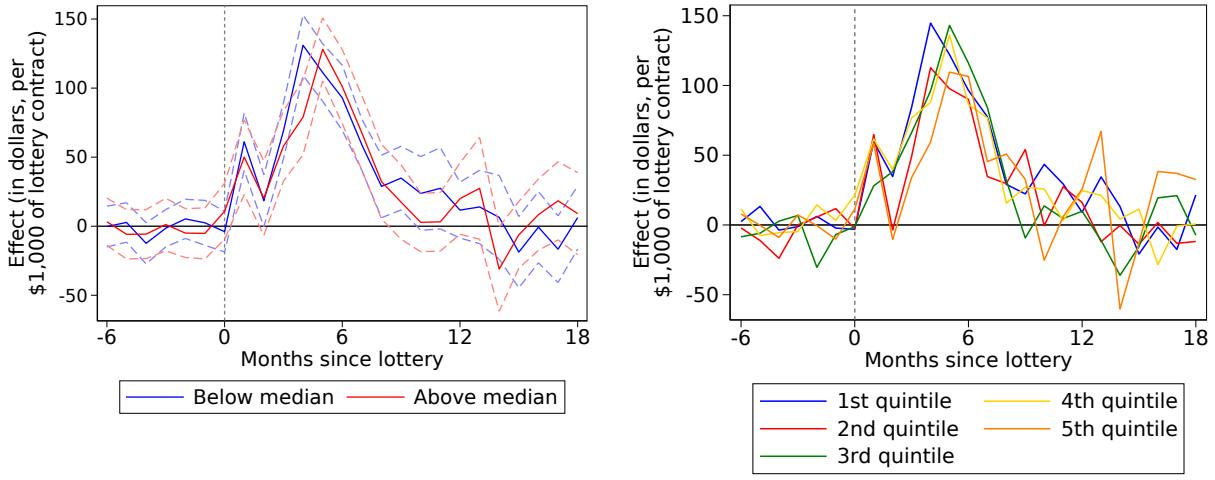


Notes: This figure plots estimates of the monthly effects of an additional $1,000 in procurement winnings shocks on the average monthly wage using data from social security records. Panel (a) follows equation (20) and presents estimates of the effect on the average monthly wage for all employees. Panel (b) shows the same for continuing workers, to avoid worker composition changes (i.e., new hires or attrition). Specifically, it only includes employees who worked continuously at the firm from 6 months before until 18 months after the first lottery that the firm participated in. Because this analysis only uses the winnings shocks of the first lottery and thus disregards any potential subsequent shocks, we estimate the coefficients in a separate regression for every lead and lag. Dashed lines indicate 95% confidence intervals that allow for clustering at the firm level.

Figure A.5: Effects on Total Sales by Number of Employees

(a) Above vs. Below Median

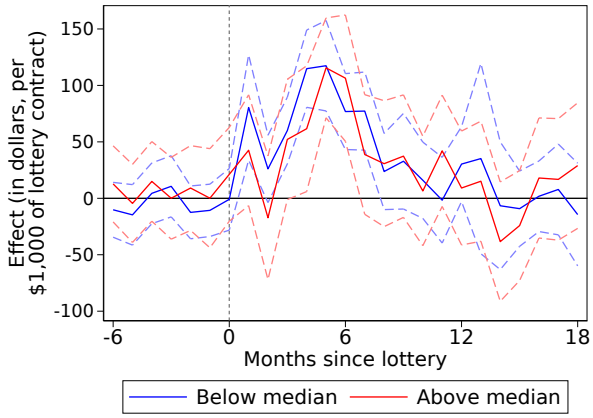
(b) Quintiles



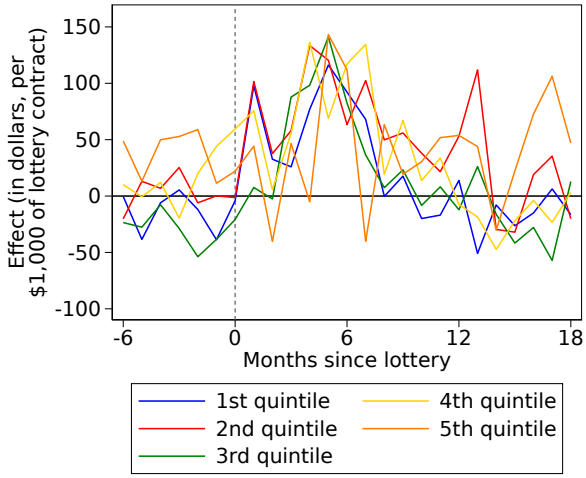
Notes: This figure extends the analysis of Figure 2, estimating the monthly effects of an additional $1,000 in procurement winnings shocks on total sales by number of employees, following equation (20). Total sales are based on monthly purchase annexes reported by client entities' VAT filings. Panel (a) presents estimates for firms with above- and below-median number of employees prior to the start of the lottery system (i.e., in 2008). Panel (b) shows the same but partitioning the sample by quintiles of the number of employees in 2008. Dashed lines in panel (a) indicate 95% confidence intervals that allow for clustering at the firm level.

Figure A.6: Effects on Total Sales by Number of Suppliers

(a) Above vs. Below Median



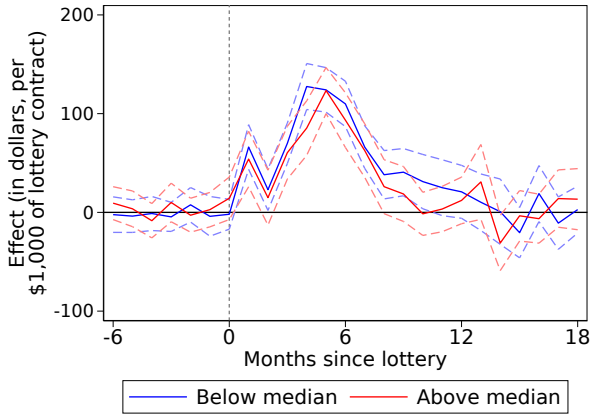
(b) Quintiles



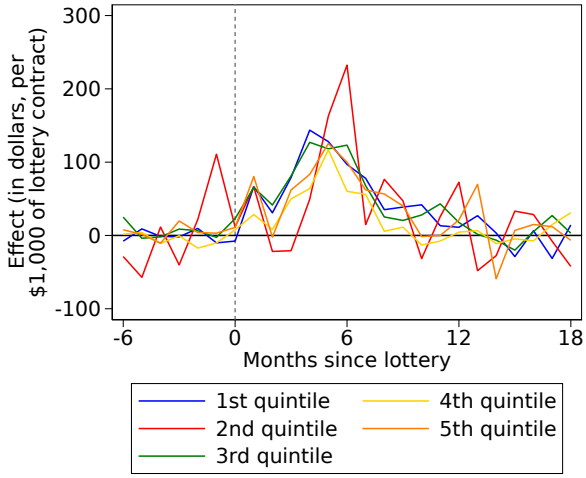
Notes: This figure extends the analysis of Figure 2, estimating the monthly effects of an additional $1,000 in procurement winnings shocks on total sales for firms with a large vs. small number of suppliers, following equation (20). Total sales are based on monthly purchase annexes reported by client entities' VAT filings. Panel (a) presents estimates for firms with above- and below-median number of suppliers prior to the start of the lottery system (i.e., in 2008), based on firms' annual income tax filings. Panel (b) shows the same but partitioning the sample by quintiles of the number of suppliers in 2008. Dashed lines in panel (a) indicate 95% confidence intervals that allow for clustering at the firm level.

Figure A.7: Effects on Total Sales by Labor Intensity

(a) Above vs. Below Median

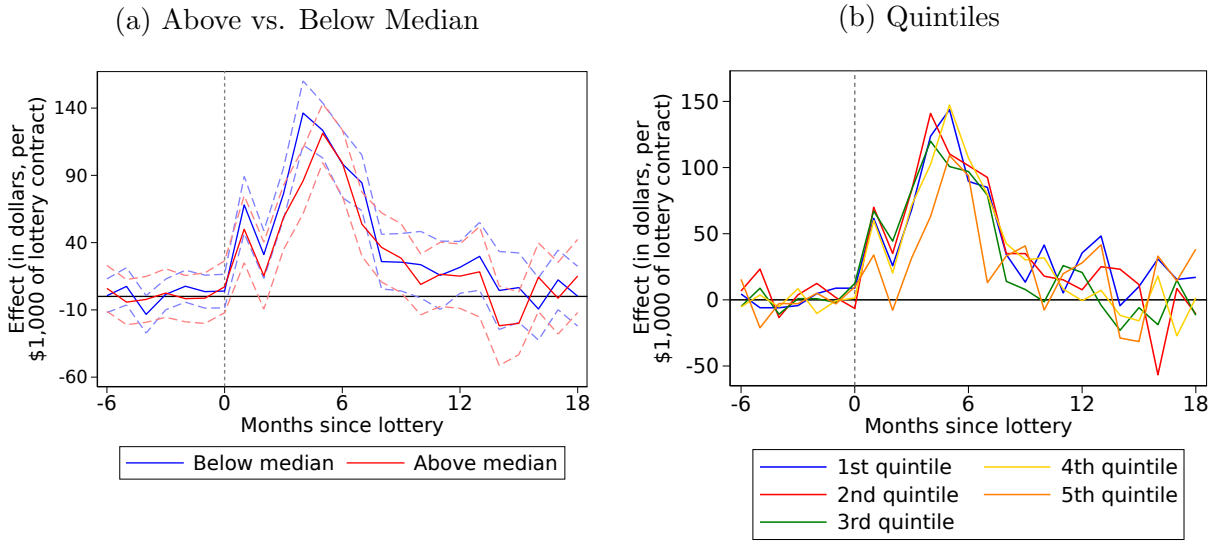


(b) Quintiles



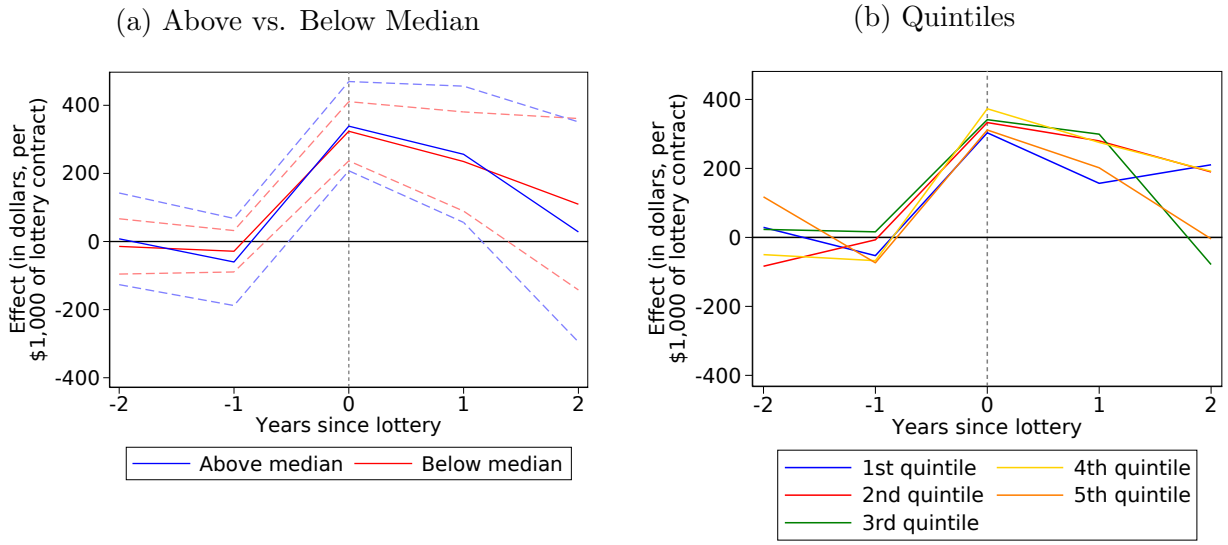
Notes: This figure extends the analysis of Figure 2, estimating the monthly effects of an additional $1,000 in procurement winnings shocks on total sales by the level of firms' labor intensity in 2008, following equation (20). Total sales are based on monthly purchase annexes reported by client entities' VAT filings. Panel (a) presents estimates for firms with above- and below-median labor intensity prior to the start of the lottery system (i.e., in 2008), based on firms' annual income tax filings. Panel (b) shows the same but partitioning the sample by quintiles of labor intensity in 2008. Labor intensity is defined as the ratio of wage payments over self-reported sales. Dashed lines in panel (a) indicate 95% confidence intervals that allow for clustering at the firm level.

Figure A.8: Effects on Total Sales by Firm Size (Based on Third-Party Reported Sales)



Notes: This figure extends the analysis of Figure 2, estimating the monthly effects of an additional $1,000 in procurement winnings shocks on total sales by the level of firms' third-party reported sales in 2008, following equation (20). Total sales are based on monthly purchase annexes reported by client entities' VAT filings. Panel (a) presents estimates for firms with above- and below-median third-party reported sales prior to the start of the lottery system (i.e., in 2008), based on firms' annual income tax filings. Panel (b) shows the same but partitioning the sample by quintiles of third-party reported sales in 2008. Dashed lines in panel (a) indicate 95% confidence intervals that allow for clustering at the firm level.

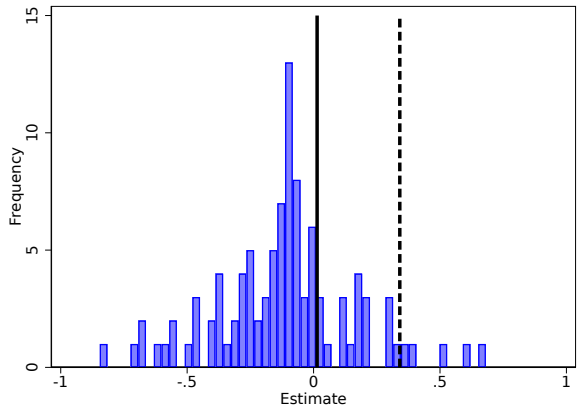
Figure A.9: Effects on Total Costs by Firm Size



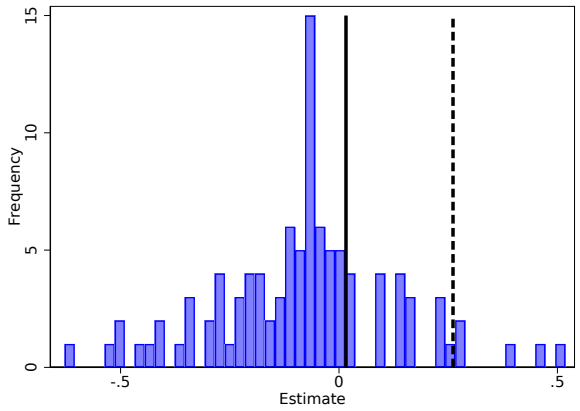
Notes: This figure extends the analysis of Figure A.3, estimating the annual effects of an additional $1,000 in procurement winnings shocks on annual total costs by firm size, following equation (20). Total costs include both labor and non-labor costs and are based on firms' annual income tax filings. Panel (a) presents estimates for firms with above- and below-median total sales prior to the start of the lottery system (i.e., in 2008), based on firms' annual income tax filings. Panel (b) shows the same but partitioning the sample by quintiles of 2008 sales. Dashed lines in panel (a) indicate 95% confidence intervals that allow for clustering at the firm level.

Figure A.10: Results from a Bootstrapping Procedure of IVCRC Estimates

(a) Sales-Weighted Variance of Wedges



(b) Welfare Cost of Misallocation



Notes: This figure shows histograms of results from a bootstrapping procedure of estimates of the sales-weighted variance of wedges, $\mathbb{V}ar_{\lambda}[\bar{\mu}]$, in panel (a), and of the welfare cost of misallocation, $\frac{\Delta W}{W}$, as described in equation (21), in panel (b). The solid vertical lines show the observed point estimates when using the original sample, as opposed to the bootstrapped samples. The dashed vertical lines show the corresponding 95% confidence limit on right-tailed tests for whether $\mathbb{V}ar_{\lambda}[\bar{\mu}]$ or $\frac{\Delta W}{W}$ are greater than zero, respectively.

B Data Appendix

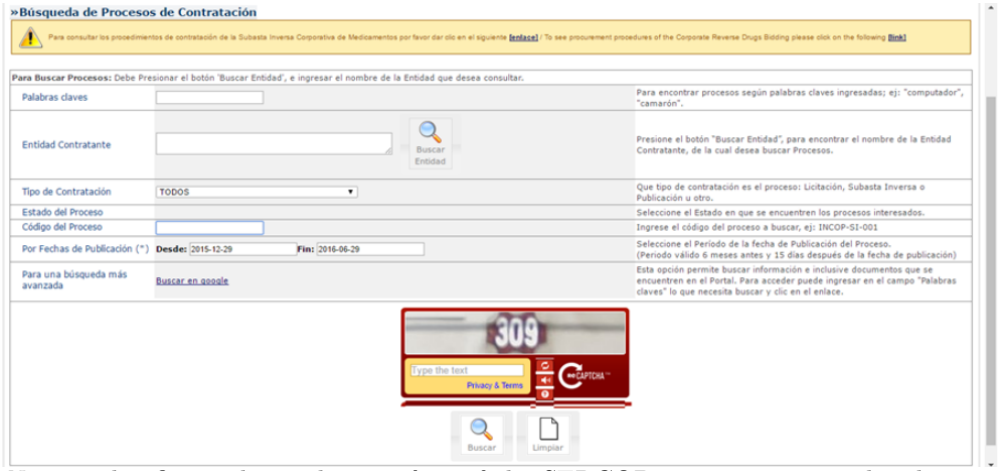
This appendix outlines the process followed to clean the data described in Section 3 and how we arrive at our final sample for analysis.

B.1 Public Procurement Lotteries and SERCOP Contracts

We scrape information from the SERCOP contracting portal between 2009 and 2014 on the application procedure for 18,474 contracts under the procurement process "Contratos de Menor Cuantía de Obras", which applied to all construction projects below a certain dollar threshold.[45] We then clean the collected information to create a dataset in which a single record corresponds to a particular firm's involvement for a particular lottery contract.

Figures B.1 and B.2 show, respectively, the website from which we scrape the information, and a sample of the data that the portal provides for each contract.

Figure B.1: SERCOP contracting portal interface



Notes: This figure shows the interface of the SERCOP contracting portal. The user inputs the contract number and a range of dates. After entering the captcha, the website displays the contract information.
Source: SERCOP contracting portal website.

We gather the following information for each contract: which firms were invited to submit proposals for it, their IDs, whether they accepted the invitation, whether they submitted

[45] This information is publicly available in the following website (working as of March 17, 2023): https://www.compraspublicas.gob.ec/ProcesoContratacion/compras/SL/view/BusquedaDeProcesos.cpe. In practice, we use an executable version of scrape.py (scrape.exe) created using py2exe. For more information, see http://www.py2exe.org/index.cgi/Tutorial.

Figure B.2: Relevant information for a sample contract

MARSEPOIN CIA. LTDA. 0890051138001	ESTRUCTURAS [DERROCAMIENTO DE INSTALACIONES DEL CAMPAMENTO DE TERMOESMERALDAS]	No cumple	No cumple	No cumple	No cumple	Cumple
MOREIRA CERVERA LUIS MIGUEL 0800426546001	SERVICIOS DE DERRIBO Y DEMOLICION DE EDIFICIOS Y OTRAS ESTRUCTURAS [DERROCAMIENTO DE INSTALACIONES DEL CAMPAMENTO DE TERMOESMERALDAS]	Cumple	Cumple	No cumple	No cumple	Cumple
PICO NAVARRO MARIA ELENA 1304932617001	SERVICIOS DE DERRIBO Y DEMOLICION DE EDIFICIOS Y OTRAS ESTRUCTURAS [DERROCAMIENTO DE INSTALACIONES DEL CAMPAMENTO DE TERMOESMERALDAS]	Cumple	Cumple	No cumple	No cumple	Cumple

Proveedores habilitados para el sorteo por parte de la Entidad Contratante

Nro.	Proveedor	Descripción	Estado
1	ALVARADO ALMEIDA EINSTIN ELIAS	cumple con todos los requisitos recnicos solicitados en los pliegos	Habilitado
2	MOREIRA CERVERA LUIS MIGUEL	cumple con todos los requisitos tecnicos solicitados en los pliegos	Habilitado
3	PICO NAVARRO MARIA ELENA	cumple con todos los requisitos tecnicos	Habilitado

Bien/Obra/Servicio Adjudicados

Categoría	Descripción Bien/Obra/Servicio	Proveedor	Cantidad Adjudicada	Precio Unitario	Subtotal	Tiempo de Entrega	Razón Adjudicación	Estado
543100011	SERVICIOS DE DERRIBO Y DEMOLICION DE EDIFICIOS Y OTRAS ESTRUCTURAS [DERROCAMIENTO DE INSTALACIONES DEL CAMPAMENTO DE TERMOESMERALDAS]	Pico Navarro María Elena 1304932617001	1	USD 167,180.14	USD 167,180.14	30	Proveedor ganador en el Sorteo MC-Obras con estos productos	Adjudicado

Notes: This figure shows information about a sample contract. The lottery winner, the value of the contract and the time to delivery, among other variables, are inside yellow squares.
Source: SERCOP contracting portal website.

a proposal, whether they were deemed eligible for the lottery, and whether they ended up winning the contract. Other variables we retrieve are: the description of the contract, its category, amount, and the delivery deadline.[46]

Starting from 38,813 scraped contracts, we kept only those with complete information and for which there exists more than one eligible participant and exactly one winner. These cleaning steps remove 20,339 (around 52.4%) of our initial observations, leaving us with 18,474 distinct contracts in total.[47] Then, we sent the remaining contracts to a third party to check that all the information was correct. Specifically, this checking involved comparing the scraped values to the actual contractual documentation (available in the same website).[48] We use the contract values as confirmed through this process.

B.2 Income Tax Forms

The firms' income tax filings provide self-reported information of firms (i.e., their revenue, costs, tax liability, among others).

[46]There are seven contracts for which the duration is extremely long and likely wrong. We discard these observations, hence the difference in the number of observations in the first three rows of Table 1.

[47]Most contracts are removed due to not having more than one eligible participant, while only three contracts are removed due to not having exactly one winner after removing submissions made without documents or by non-eligible firms.

[48]We found errors on around 1% of the contracts and manually fixed them. These errors were due to discrepancies between the contracts and the website itself, not due to our scraping procedure.

F101/102 forms

In Ecuador, all incorporated firms are legally obligated to submit an annual detailed corporate income tax form (F101), independent of how large their revenues, costs or assets are. This also applies to all publicly-owned firms. On the other hand, unincorporated firms (which mainly consist of self-employed individuals) are required to file the income tax form F102 if their annual revenue exceeds a standardized deduction amount (which was approximately $10,000 in our sample period). These forms include self-reported information of the firms' revenue and costs (broken down by certain sub-categories). Small unincorporated firms submit a simplified version of the income tax form ("short form F102") which corresponds to the personal income tax form. Large unincorporated firms have to file an extended income tax ("long form F102") consisting of two parts: a part for reporting business income corresponding to the form for incorporated firms and a part for reporting individual income mirroring the short from F102.

In very rare cases (i.e., 0.005% of firm-year observations in the annual income tax data), a firm reports both a F101 and a F102 form. Whenever this is the case and the forms were filed on the same date, we proceed as follows. If the duplicates share the same value in the same tax items, in the first instance we keep the form that the firm was expected to correctly file. In the second instance, we track the first year in which the firm uniquely filed a form and keep the one that matches the type of this first unique filing. If the duplicates have different values in the same tax item, then we keep the observation that has the highest value for each item. On the other hand, if the forms were filed on different dates, then we keep the most recent form.

Self-reported sales and costs measures

Our measures of self-reported sales and costs are constructed based on the annual income tax filings. In line with the theory developed in Section 2, we seek to capture (i) the firm's income from selling its goods or services and (ii) the costs associated with producing this output.

Accordingly, the total sales measure includes domestic sales, net exports, and other revenues related to selling goods or services. Table B.1 lists the sales items from the income

tax form included in our measure of self-reported sales. We deliberately exclude revenue from dividends, financial rents, donations and contributions, and capital gains on sales of fixed assets. The total costs measure comprises labor costs and non-labor costs as shown in Table B.2. We deliberately exclude reported costs due to losses from the sales of assets, and we also disregard tax liabilities deducted from the income tax (e.g., property taxes, excise tax, non-creditable VAT).[49]

Table B.1: Sales Line Items in Forms 101 and 102

Sales item
Domestic sales subject to 12% tax rate
Domestic sales subject to 0% tax rate
Exports
Other income from abroad
Other taxable income
Other exempted income
Change in inventory of products produced by the firm
Sales related to registered business activities*
Received fees*
Received agricultural income*
Received income from abroad*
Received wages*

This table shows the line items included in the definition of total sales for the annual income tax form. * marks line items that are only present for F102 and related to the "short form" section of the form.

[49]Specifically, we compute our sales and costs measure by subtracting the excluded line items from the reported total sales and costs measures, respectively. This has the advantage that a few firms only report information on their total costs or revenues but not on sub-categories. In the very few instances in which the adjusted total sales or costs measure is negative due to reporting errors, we replace it with a zero.

Table B.2: Cost Line Items in Forms 101 and 102

Costs items	Category
Wages, salaries and other taxable remunerations	Labor costs
Social benefits and other non-taxable compensation	Labor costs
Contribution to social security (including reserve fund)	Labor costs
Obligatory profit share passed on to employees	Labor costs
Costs related to wages*	Labor costs
Net purchases of domestic goods not produced by the firm	Non-labor costs
Imports not produced by the firm	Non-labor costs
Net domestic purchases of raw material	Non-labor costs
Imports of raw material	Non-labor costs
Professional fees and expenses	Non-labor costs
Fees to foreigners for one-time expenses	Non-labor costs
Real estate rent	Non-labor costs
Maintenance and repairs	Non-labor costs
Fuel	Non-labor costs
Marketing	Non-labor costs
Supplies and materials	Non-labor costs
Transportation	Non-labor costs
Commercial leasing	Non-labor costs
Commissions	Non-labor costs
Insurance and reinsurance intermediaries	Non-labor costs
Administrative costs	Non-labor costs
Travel expenses	Non-labor costs
Public services	Non-labor costs
Payment for other goods and services	Non-labor costs
Bank interest	Non-labor costs
Interest paid to third parties	Non-labor costs
Amortization	Non-labor costs
Change in inventory of goods not produced by the firm	Non-labor costs
Change in inventory of raw material	Non-labor costs
Depreciation of fixed assets	Non-labor costs
Provisions	Non-labor costs
Other losses	Non-labor costs
Indirect costs incurred from abroad by related parties	Non-labor costs
Costs related to registered business activities*	Non-labor costs
Costs related to professional activity or liberal occupation*	Non-labor costs
Costs related to real estate rents*	Non-labor costs
Costs related to other assets*	Non-labor costs
Costs related to agricultural income*	Non-labor costs
Costs related to other income*	Non-labor costs

This table shows the line items included in the definition of total cost for the annual income tax form, and which are included in the definition of implied labor and non-labor costs. * marks line items that are only present for F102 and related to the "short form" section of the form.

We construct these sales and costs measures to represent prices gross of sales taxes (e.g., VAT and excise tax). This attempts to consistently follow our theory which represents sales at the actual expense of the buyer (i.e., including any statutory sales tax). On the F101/102, firms report revenues and costs net-of-VAT, but gross of other sales taxes.[50] Hence, we adjust our self-reported sales measure as follows.

1. Firms report sales subject to 12% VAT in a dedicated line item.

2. We multiply the amount reported by 0.12 and add it to our measure of total self-reported sales.

Since costs are not reported separately by their VAT rate, we proceed as follows to compute the gross-of-VAT costs based on the self-reported net-of-VAT amounts.

1. We identify the line items that are subject to VAT and calculate their total amount (denoted V_i)

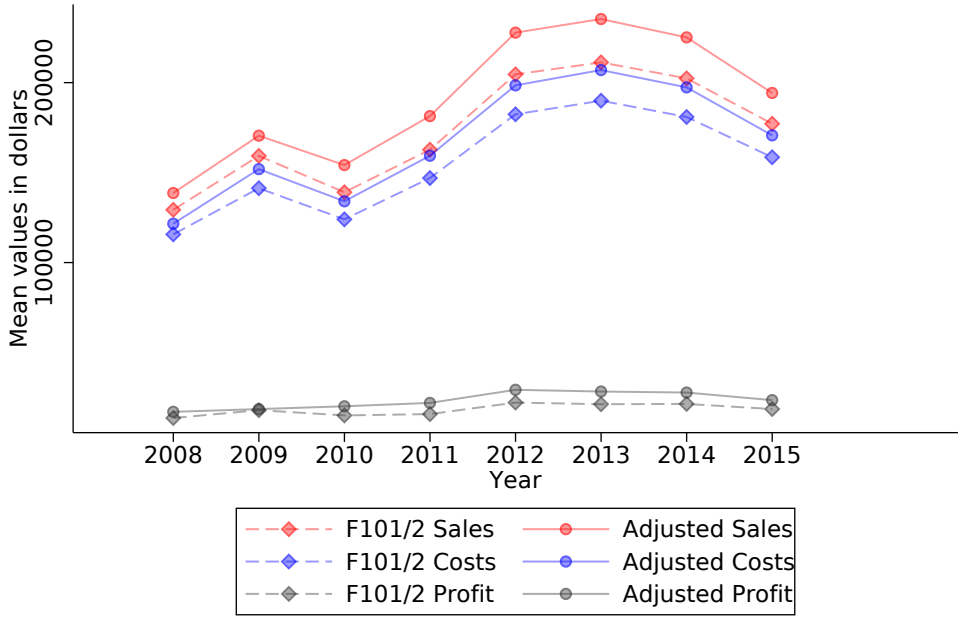
2. If a firm files purchase annexes, we calculate the (observed) share of purchases subject to 12% VAT (denoted s_i). For firms that do not file purchase annexes, we predict this share by their level of sales and incorporation status.

3. We calculate the VAT amount as $V_i \times s_i \times 0.12$ and add it to our measure of total self-reported costs.

All in all, our adjusted measures of sales and costs (and profits) are consistent with the total revenue and costs items (and profits) from the annual tax filling. They are also consistent across years. This consistency is shown in Figure B.3, which plots the evolution over time of these variables in their raw and adjusted versions.

[50] In Ecuador, sales are subject to either a 12% or 0% VAT rate, or are exempt from the VAT.

Figure B.3: Sales and Cost Measures



Notes: This figure shows the evolution across years of the sales, costs and profit measures for their raw (F101/2) and adjusted versions.

B.3 Purchase Annexes

The monthly purchase annexes provide both self-reported firms' purchases information (i.e., their own purchases) and, implicitly, third-party reported information about other firms' sales.

All incorporated firms (F101 filers) are required to submit purchase annexes every month. F102-filing firms must keep accounting records and file purchase annexes on a monthly basis if their annual revenues surpass a certain threshold (e.g., $100,000 in 2012). They are also mandated to do this if their annual costs and expenses exceed a given value ($80,000 in 2012), or if they begin economic activities with capital above a certain threshold ($60,000 in 2012).[51]

To quantify transactions between firms using the purchase annex, we follow a process similar to that described in Adao et al. (2022). For each transaction, we observe the anonymized

[51] A considerable number of firms that are not required to do so end up voluntarily filing purchase annexes. However, for these smaller firms (who represent a smaller share of the country's economic activity, naturally) records may not be complete. See Adao et al. (2022) for further details.

tax ID of both the buyer and the seller, the value of the purchase, the VAT paid, what part of the value was subject to either a tax rate of 12% or 0%, and the transaction's date. The data feature some inconsistencies. For instance, we drop transactions with a negative VAT value. When there is a positive value for VAT but the transaction value is inconsistent with it, we rectify the transaction value. In cases where there is a positive transaction value but VAT is missing, we calculate the correct VAT based on the transaction value. We also do not consider purchases whose value exceeds a value greater than two times the maximum between the buyer's annual cost and the seller's annual revenue, as such transactions are likely the result of data entry errors. Finally, we also drop transactions where the buyer and the seller are the same firm. The cleaned purchased annexes are used to compute monthly, half-yearly and annual transactions between firms.[52]

Since the transactions are reported purchases by the buyer, we compute a measure of *third-party* reported sales based on the purchase annex data.[53]

B.4 Social Security and Firm Characteristics

First, we combine two sources of data for employment. One covers the period from 2007 to 2017 and is extracted from the social security record that keeps track of all the employees whose employer filed the social security for them; the other one spans from 2009 to 2016, and has data from either the social security record or the F107 form (a tax declaration form filed by firms with information about their workers). We restrict the sample to the universe of employees that ever work for lottery firms. Then, we identify the month in which an employee is hired by a firm whenever there are no payments from the employer in the previous month. Similarly, we consider the employee to be a new employee if they were hired in the last 12 months, and we define documented employees as those that have worked for any firm in this same period.

Second, we have basic information of every firm registered with the SRI up to 2012. This

[52]We use the register date when available. Otherwise, we use the purchase date. We drop transactions where neither of these two is available.

[53]Similar to the self-reported sales measure, we construct third-party reported sales gross of sales taxes. To obtain this measure, we sum up the reported net-of-VAT transaction amount and the VAT reported for that transaction to create a measure of third-party reported sales gross-of-VAT.

dataset provides a snapshot of the registered firms' characteristics in this year for variables such as: the province where it is located, the year in which it became active and the industry code (ISIC 3.1).

B.5 Combining the Datasets

We can combine the different datasets based on consistently anonymized firm identifiers across all datasets.

To identify sales from lottery participants to different types of entities, we join the lottery data with the transaction-level purchase annexes, information from the procurement contracts, and information on firm characteristics. This allows us to identify whether the reporting firm (i.e., the buyer in the purchase annex) is a procuring entity, another public entity, or a private firm.[54] We then aggregate the reported purchases on the month-seller level to obtain measures of third-party reported sales to different types of entities as well as total third-party reported sales.

Finally, we define a firm to be economically active from the first time it self-reports positive sales or costs in its income tax forms, or appears in any of the above datasets either as a lottery participant, an employer (in the social security data), or a supplier (in another firm's purchase annex). We then exclude firm-year observations from the period before a firm is economically inactive and impute zeroes for any missing values after a firm has become economically active.

[54] We define "public entities" as firms that are either government agencies or utilize the lottery procurement system. Similarly, "private firms" are defined as firms that are neither a government agency nor utilize the lottery procurement system.

C Test of statistical independence

This appendix provides further details about the test of statistical independence used in Section 5 to test for U-AEP and C-AEP.

C.1 Testing for U-AEP

As stated in Proposition 1, the null of U-AEP can be carried out by testing for whether $f(\cdot) = 0$ in equation (12) or not. Following Ding et al. (2016), one way to do this is via a quantile regression of $\Delta\Pi_i(\mathbf{1})$ on Z_i. This corresponds to the quantile model

$$Q_{\Delta\Pi(\mathbf{1})}(\tau \mid Z) = \alpha(\tau) + \beta(\tau)Z, \tag{25}$$

for any quantile τ, where the null hypothesis of U-AEP requires that $\beta(\tau) = 0$ for all τ. We therefore use the test statistic

$$TS \equiv \max_{\tau \in \mathcal{T}} |\beta(\tau)|, \tag{26}$$

where $\mathcal{T}$ denotes some finite set of quantiles to be evaluated. The null should be rejected when TS is large.

We implement a randomization inference version of this test. Following Fisher (1935), the basic idea is to compare TS based on the actual data to the distribution of statistics TS_l that one obtains in a series of simulations $l = 1...L$ in which the winners of each lottery are randomly re-drawn from the set of actual entrants into each lottery and the sharp null value is used to generate simulated outcome data for each observation. Since the sharp null in question is one of zero treatment effect (of Z_i on $\Delta\Pi_i(\mathbf{1})$), this procedure amounts to performing a quantile regression of the actual data $\Delta\Pi_i(\mathbf{1})$ on the values of Z_i generated by simulated counterfactual lottery winnings and repeating this across many simulations. Specifically, we proceed as follows:

1. Estimate TS on the actual data and denote this value by $\widehat{TS}$. That is, estimate a quantile regression of $\Delta\Pi_{i,t}(\mathbf{1})$ on $Z_{i,t}$, obtain the quantile coefficient estimates $\widehat{\beta}(\tau)$,

and compute the corresponding $\widehat{TS}$ based on (26).[55] We set $\mathcal{T} = \{0.1, ..., 0.9\}$.

2. Re-randomize the identity of the winning firm in each lottery. Refer to the values of Z_i that are obtained under this re-randomization as Z_i^l.

3. Estimate a quantile regression of $\Delta\Pi_{i,t}(\mathbf{1})$ on $Z_{i,t}^l$. Calculate the corresponding value of TS_l based on (26), again with $\mathcal{T} = \{0.1, ..., 0.9\}$.

4. Repeat steps #2 and #3 an additional $L - 1$ times. In practice, we set $L = 100$.

5. The resulting L values of TS_l provide an estimate of the exact null distribution of the test statistic TS. Therefore, calculate the exact p-value for the null hypothesis from the share of simulations l in which $TS_l > \widehat{TS}$. If $p < \alpha$, we reject the null at the α level.

Figure 8, at the value displayed as $\chi = 1$ on the x-axis, reports the p-value from performing this test in our context.

C.2 Testing for C-AEP

Recall from Proposition 1 that the null of C-AEP is the same for that of U-AEP apart from the fact that rather than testing on the basis of $\Delta\Pi_i(\mathbf{1})$, we must repeat the test on the basis of $\Delta\Pi_i(\boldsymbol{\chi})$ for all values of $\boldsymbol{\chi} > \mathbf{0}$. As discussed in Section 5.2, we confine attention to values of $\boldsymbol{\chi}$ in which all elements of this vector are equal to each other and equal to the scalar χ (i.e., $\chi_m = \chi$ for all m). Our test for C-AEP therefore simply amounts to repeating the above test for U-AEP (which was effectively done for $\chi = 1$) at a wider range of values of χ. In practice, we do this for 201 uniformly spaced values in the interval $[0.9, 1.4]$. Figure 8 reports the p-values for each of these tests. The test for C-AEP fails to reject (at the 5% level) if any of the p-values exceeds 0.05.

[55] As described in Section 5, the lack of available data on prices of outputs and inputs in our context requires that we proxy for $\bar{p}_i \Delta y_i$ with $\Delta(p_i y_i)$, etc. We therefore have $\Delta\Pi_i(\mathbf{1}) = \Delta(p_i y_i) - \sum_m \Delta(w_m x_{im})$. Further, when computing time-differences "Δ" we use differences in two-year averages, such as $\Delta(p_{i,t} y_{i,t}) \equiv \frac{1}{2}(p_{i,t} y_{i,t} + p_{i,t+1} y_{i,t+1}) - \frac{1}{2}(p_{i,t-1} y_{i,t-1} + p_{i,t-2} y_{i,t-2})$. We then stack the cross-sections from all such available changes over 2008–2015, but only use the firm-year observations corresponding to firms that are active in 2008 so that our testing sample aligns with those firms that enter our sample for estimating sales-weighted wedge distributions in Section 6. Finally, the instrument is defined as $Z_{i,t} = D_{i,t} - D_{i,t-2}$.

D Assessing First-Stage Rank-Invariance

Section 2.4 describes the use of the IVCRC estimator developed in Masten and Torgovitsky (2016) to estimate the sales-weighted first and second moment of the distribution of wedges. As outlined in Assumption 4, one condition for consistency of IVCRC is what Masten and Torgovitsky (2016) refer to as "first-stage rank-invariance": that the first-stage relationship can be characterized by $\bar{w}_m \Delta x_{im} = h_m(Z_{im}, V_{im})$ for some unknown $h_m(\cdot)$ and scalar V_{im}, with $\frac{\partial h_m}{\partial V_{im}} > 0$ for all $m \in \overline{\mathcal{M}}(i)$ and all Z_{im}. This condition implies that (for any input type m), if we set Z_{im} to any value z, then the ranking of firms in terms of their input changes (i.e., $\bar{w}_m \Delta x_{im}$) would not depend on the value of z chosen. Put differently, firms can respond to Z_{im} heterogeneously but only to the extent that their rank in the input change distribution is unchanged by such response heterogeneity.

In order to gauge the plausibility of this assumption in our context, in this Appendix we conduct a simulation of a model inspired by features of Hsieh and Klenow (2009), calibrated to our regression sample, that suggests that first-stage rank-invariance may hold approximately in our context.

D.1 Setup

We consider a set of N firms indexed by i. Each produces a differentiated product y_i using a single input x_i according to the technology

$$y_i = a_i x_i, \tag{27}$$

where a_i denotes firm i's productivity. The firm faces two sources of demand: (a) private-sector buyers whose demand is given (in expenditure terms) by $E_i = p_i^{1-\sigma}$, where σ is a constant price elasticity of demand; and (b) procurement lottery-based demand denoted (again in expenditure terms) by W_i. The firm's total sales ($S_i \equiv p_i y_i$) are therefore given by $S_i = E_i + W_i$.

The definition of the wedge in equation (3) implies, in this model, that $\mu_i = p_i a_i / w$, where w denotes the price of the input. Starting from the "time-0" allocation (at which all

variables are denoted with a bar), we consider a small change in the input Δx_i (between time-0 and a subsequent time period) that results from an arbitrary change in a_i and W_i, but where, for the sake of simplicity in this simulation, μ_i and w are held constant. This will imply that the first-stage of our IVCRC estimation equation is given by

$$\bar{w}\Delta x_i = \left[(\sigma-1)\bar{w}^{1-\sigma}(\bar{\mu}_i)^{-\sigma}\bar{a}_i^{\sigma-1}\right]\widehat{a}_i + (\bar{\mu}_i)^{-1}\Delta W_i, \tag{28}$$

where $\widehat{a}_i \equiv \Delta a_i/\bar{a}_i$.

As discussed in Section 3.3, procurement lottery-based demand W_i satisfies $W_i = D_i + \mathbb{E}[W_i \mid \{A_k, N_k\}_{k\in\mathcal{K}_i}]$, where A_k denotes the contract value of lottery k, N_k denotes the number of participants in lottery k, $\mathcal{K}_i$ denotes the set of lotteries that firm i participates in during the given time period, and D_i denotes the deviation between actual lottery demand W_i and expected lottery demand given lottery participation. Given the definition of our instrument, $Z_i \equiv \Delta D_i$, we therefore have

$$\bar{w}\Delta x_i = b_i + (\bar{\mu}_i)^{-1}Z_i, \tag{29}$$

where we have defined

$$b_i \equiv \left[(\sigma-1)\bar{w}^{1-\sigma}(\bar{\mu}_i)^{-\sigma}\bar{a}_i^{\sigma-1}\right]\hat{a}_i + (\bar{\mu}_i)^{-1}\Delta\mathbb{E}[W_i \mid \{A_k, N_k\}_{k\in\mathcal{K}_i}]. \tag{30}$$

Input changes ($\bar{w}\Delta x_i$) therefore derive from two components. First, the component b_i arises due to factors unconnected from the instrument (the technology and lottery participation changes in equation 30). And second, the component $(\bar{\mu}_i)^{-1}Z_i$, which does depend on the instrument. In this model, first-stage rank-invariance would be satisfied if there were no heterogeneity in $\bar{\mu}_i$ because in such a case equation (29) could trivially be written in the form $\Delta x_i = h(Z_i, V_i)$ with a scalar V_i and with $\frac{\partial h}{\partial V_i} > 0$ at any value of Z_i. More generally, if there is heterogeneity in $\bar{\mu}_i$ and also in b_i, and $Corr(\bar{\mu}_i, b_i) < 0$, then first-stage rank-invariance is not guaranteed. In what follows, we use the structure of the model in this Appendix in order to calibrate the values of b_i and $\bar{\mu}_i$ implied by our data, and thereby calculate how often rank-invariance is violated.

D.2 Quantifying First-Stage Rank-Reversals

In principle, one could assess the prevalence of first-stage rank-invariance in this model by: (a) measuring b_i and $\bar{\mu}_i$ for every firm i; (b) computing each firm's rank in the distribution of $\bar{w}\Delta x_i$ implied by equation (29) at the lowest possible value of the instrument Z; and then (c) repeating this at every other possible value of Z in order to keep track of the number of times that firms' ranks (in the distribution of $\bar{w}\Delta x_i$) reverse across the full support of Z values. In practice, however, this is both computationally costly (given the number of potential calculations involved) and also conceptually unclear, given that Z is in principle a continuous variable. We therefore follow Gollin and Udry (2021) in drawing a random sample of comparisons (both pairs of firm observations whose rank is to be assessed and sets of values of Z at which we assess whether rank-reversal has occurred) from the set of all possible comparisons. To the extent that this random sample is large, it should provide a reliable impression.

We calibrate the model as follows. First, we set $\sigma = 3$ as in Section 6.4 (and, for example, in Hsieh and Klenow (2009)). Second, we let $\bar{w}$ equal the average annual wage for a worker in our dataset. Third, we pursue two versions for the calibration of $\bar{\mu}_i$ in parallel: (a) a version in which these come from the sales bin-specific average wedge estimate obtained from the IVCRC estimation procedure in Section 6.2; and (b) a version that assumes all firms have constant returns-to-scale technologies, as in Section 6.4 and, for example, Hsieh and Klenow (2009).[56] Fourth, given data on S_i and W_i, manipulations of the above supply and demand equations can be used to solve for $\bar{a}_i$ and $\widehat{a}_i$.[57] Finally, $\mathbb{E}[W_i \mid \{A_k, N_k\}_{k \in \mathcal{K}_i}]$ is directly observable. Applying these steps we therefore know the values of b_i and $\bar{\mu}_i$ (one for each separate method for calibrating $\bar{\mu}_i$) corresponding to every firm-year observation in our dataset.

Using these values of b_i and $\bar{\mu}_i$, we quantify rank-reversals using the following procedure (done twice for each separate method for calibrating $\bar{\mu}_i$):

1. Randomly partition all firm-year observations into pairs r. Let r_1 and r_2 indicate the

[56] As discussed in Section 6.4, under constant returns-to-scale we have $\bar{\mu}_i = \bar{S}_i/(\sum_m \bar{w}_m \bar{x}_{im})$.

[57] That is, combining the equations $S_i = E_i + W_i$, $E_i = p_i^{1-\sigma}$, and $\mu_i = p_i a_i / w$, we have $a_i = w\mu_i(S_i - W_i)^{1/\sigma}$.

two observations in pair r. Let there be P pairs in total.

2. Begin with pair $r = 1$. Randomly draw a value of the instrument Z from the distribution of the deviations from expected winnings and denote that value by Z_a. Evaluate the first-stage equation (29) at $Z_i = Z_a$ for observation r_1 and denote the result by $F(r_1, Z_a)$. Do the same for observation r_2 and denote the result by $F(r_2, Z_a)$.

3. Repeat step #2 for a second randomly drawn value of the instrument, denoted Z_b. Hence calculate $R(r) \equiv [F(r_1, Z_a) - F(r_2, Z_a)][F(r_1, Z_b) - F(r_2, Z_b)]$. This provides one possible comparison (between two observations and two values of the instrument).

4. Repeat steps #2 and #3 for nine additional sets of randomly drawn instrument value pairs. This yields ten values for $R(r)$ in total for the first pair, $r = 1$.

5. Repeat steps #2-#4 for all remaining pairs $r = 2...P$.

6. Rank-reversals have occurred whenever $R(r) < 0$. Calculate the share of comparisons (out of $10P$ total comparisons) in which this has happened.

Table D.1 reports the share of rank reversals in the randomly chosen comparisons illuminated by the above simulation. Unsurprisingly, the substantially smaller wedge dispersion used in the IVCRC-based version of this simulation (columns 1-3) shows far less propensity for rank-reversals than does the constant returns-based version (columns 4-6). But even in the latter case, rank reversals are quite rare, happening in less than 6.2% of pairwise comparisons.

Table D.1: Frequency of Rank-Reversals in Simulated Economy

IVCRC wedges			CRTS wedges		
Rank reversals (1)	No. of comparisons (2)	Share (3)	Rank reversals (4)	No. of comparisons (5)	Share (6)
1,256	141,820	0.0089	7,565	123,070	0.0615

Notes: This table shows results on the frequency of first-stage rank-reversals across randomly drawn pairwise comparisons of observations and instrument values. "IVCRC wedges" assigns firms' wedges based on the sales-bin specific point estimates of average wedges obtained when using the IVCRC procedure described in Section 6.2. "CRTS wedges" instead uses the formula $\bar{\mu}_i = \bar{S}_i/(\sum_m \bar{w}_m \bar{x}_{im})$, which follows from assuming that all firms have constant returns-to-scale technologies as in Section 6.4, in order to estimate wedges. The number of comparisons differs across these alternatives due to the fact that the latter version cannot be computed for firms with zero costs.