When do employers share? Rent sharing, monopsony and minimum wages

Ihsaan Bassier* and Joshua Budlender†

April 3, 2025

Click here for latest version

When firms experience productivity increases or better demand conditions, they tend to share the benefits with their workers. An increasingly favoured explanation for such rent sharing is labour market monopsony. But what happens when firms also face binding minimum wages, as is empirically common? In our theoretical model, when a monopsonistic firm's preferred wage lies just below the mandated minimum, it will neither raise wages (share rents) nor expand employment following a local revenue-productivity increase. Instead, the firm maintains the minimum wage and absorbs the additional revenue into a higher markdown. We find compelling evidence for these predictions using South African administrative data, based on a cross-sectional kink design as well as within-firm responses to internal and shift-share trade shocks. Our results demonstrate a novel monopsonistic mechanism in the labour market, and also highlight a potential disconnect between the level of rent sharing and the pass-through.

Author ordering is alphabetical; the authors contributed equally to this work. For comments and suggestions we owe particular thanks to Arindrajit Dube, as well as David Berger, Ellora Derenoncourt, Daniele Girardi, Leonard Goff, Ayanda Hlatshwayo, Michael Kilumelume, C. Friedrich Kreuser, Murray Leibbrandt, Attila Lindner, Alan Manning, Suresh Naidu, Carol Newman, John Rand, Garima Sharma, Eric Verhoogen, and Matt Woerman. We are grateful for comments from participants of several conferences (ASSA 2024, ESSA 2023), workshops (UNU-WIDER Work-in-Progress 2023, Princeton Global Minimum Wages 2024, Pretoria-Stellenbosch PhD 2024), and working groups (University of Massachusetts Amherst PhD Labor-Development 2019, Columbia University PhD 2020-2021). For assistance with data access we thank the team at the NT-SDF. We gratefully acknowledge funding from UNU-WIDER, and from the research initiative "Structural Transformation and Economic Growth" (STEG), contract reference STEG_LOA_2206_Budlender. An early version of this paper was previously circulated under the title "Rent sharing, wage floors and development."

*Economics, University of Surrey; Centre for Economic Performance, London School of Economics & Political Science; Southern Africa Labour and Development Research Unit (SALDRU), University of Cape Town (UCT). Email: i.bassier@surrey.ac.uk

†SALDRU, UCT. Email: joshua.budlender@uct.ac.za

1 Introduction

A substantial literature has sought to explain the phenomenon of firm rent sharing as an outcome of monopsonistic labour markets (Card et al. 2018; Kline et al. 2019; Lamadon et al. 2022). Another large but distinct literature seeks to understand the direct employment or welfare effects of minimum wages under monopsony (Dickens et al. 1999; Engbom and Moser 2022; Dustmann et al. 2022; Azar et al. 2024; Berger et al. 2025). A natural question then is: what does the common empirical setting of minimum wages imply for monopsonistic rent sharing? How do firms in such contexts respond to productivity increases or more favorable demand conditions?

We show that incorporating the labour market constraint of a binding minimum wage has striking implications for firm behaviour when it comes to rent sharing, employment responses, and profits and markdowns. Our question provides a demanding test for some potentially distinctive predictions of a monopsonistic model, and our results additionally highlight an important distinction between the rent-sharing *level* versus its *elasticity* (or pass-through). We first develop some novel insights from a standard model of monopsonistic competition with minimum wages, and then test these predictions using matched employer-employee administrative data from South Africa. We show that this mechanism is likely to be empirically important for a range of countries where minimum wages are substantially binding.

We start by establishing an intuitive fact: a large segment of firms with low but varying productivity pay around the minimum wage, with a concomitantly lower rent-sharing (or pass-through) elasticity for these firms. We then show that above a particular productivity threshold, there is a kink in the wage-productivity curve and the expected pass-through dynamics commence. This is consistent with our very general and simple monopsony model, which we show has distinctive implications for markdown and employment dynamics around this kink-point. Specifically, we predict that just below this kink point firms will absorb favorable revenue-productivity shocks into increased markdowns, instead of increasing employment or wages, and will only increase employment and wages with productivity once markdowns have been restored to optimal monopsony levels. We test these predictions using a cross-sectional kink design as well as within-firm responses to internal and shift-share trade shocks, and find strong evidence consistent with these predictions.

This productivity threshold divides minimum wage-constrained and -unconstrained firms. The intuition for just-constrained firms absorbing revenue-productivity increases into markdowns starts with noting that the minimum wage statically redistributes rents from firms to workers at these low productivity constrained firms. These firms therefore receive a lower markdown than they would in the absence of the minimum wage. Higher productivity unconstrained firms, in contrast, optimally markdown wages and receive excess rents. There is consequently a productivity region in which just-constrained firms absorb revenue-productivity growth as increased rents until they have reclaimed the markdown associated with the unconstrained productivity level. Firms above this productivity threshold "share" revenue-productivity gains with workers in the form of higher wages, because in a monopsony model this is the prerequisite for firm expansion. Observationally, this rent-sharing pattern is not limited to monopsony models, and we show that it comes out clearly from other models of imperfect competition such as Diamond-Mortensen-Pissarides or DMP (Mortensen and Pissarides 1999). However, we also establish a prediction more peculiar to monopsony models, which is that such just-constrained firms also do not increase employment in response to productivity increases, while their wage markdowns rise. The intuition here is that such firms cannot attract additional workers without increasing the wage, but they will not increase it above the minimum wage until their markdown has been restored to the optimal unconstrained monopsony level.

The key empirical predictions of the model, then, are differential employment and profit responses to revenue-productivity shocks along the firm productivity distribution, with a break at the point where pass-through commences.[1] After estimating firm-specific productivity using Ackerberg et al. (2015), we test this using a kink design focusing on the productivity threshold identified on the wage-productivity curve. Our test is in the spirit of the regression kink design used in several studies (Card et al. 2015, 2017; Ganong and Jäger 2018), though an important difference is that there is no direct policy treatment in our case. Our kink design shows patterns that fit the model predictions remarkably well: profit-share steeply increases with productivity below the threshold, and then is more constant above it; and employment does not change much with productivity below the threshold, but increases strongly above it.

We then complement this kink design by considering how firms on either side of the productivity threshold *respond* to revenue-productivity shocks, replicating and extending leading approaches in the literature. Following Lamadon et al. (2022), we use an "internal instrument" method, which entails constructing a stacked event study where firm-specific treatment is defined as an unusually large observed increase in firm value-added. We also use an "external instrument" based on a shift-share variable we construct from firm-specific shares of destination-country exports and imports (the "share") and destination-country GDP movements (the "shift"), similar to Garin and Silvério (2023). Our results strongly support our theoretical predictions: compared to responses in the unconstrained region, in the constrained region the wage response (the rent-sharing elasticity) is approximately 30% lower, the profit share response is almost 3 times higher, and the employment response is approximately 25% lower. Estimates by productivity bin support the prediction that the break in response size occurs around the productivity threshold. The estimates are remarkably similar using the external and internal instruments given how different the sources of variation are, which builds confidence in a causal interpretation. We conduct a variety of robustness tests.

Our primary contribution is to the rent-sharing and monopsony literatures. Theoretically and empirically, we identify a novel region where monopsonistic firms bound by minimum wages do not increase wages or employment in response to positive revenue-productivity shocks, and instead increase their profit share. A related contribution is that, as noted above, these patterns

[1] Because markdowns are unobserved, in our empirical specification we approximate markdowns with a measure of the "profit share", as one minus the labor share as it is defined in Gouin-Bonenfant (2018).

are supportive of monopsonistic mechanisms in the labour market (Dube et al. 2019; Bassier et al. 2022; Sokolova and Sorensen 2021).

The unevenness of rent sharing also highlights an interesting distinction between two related concepts in the literature (Bell et al. 2024; Kline et al. 2019; Risch 2024): the rent-sharing level (wage over marginal revenue product), versus its elasticity or pass-through (change in wage with respect to revenue-productivity). In many cases, these two concepts do closely align. However, our model and results highlight that when the minimum wage increases, the rent-sharing level rises, yet the pass-through declines. A similar distinction holds for employment: for firms just-constrained by minimum wages the level of employment is higher than otherwise, even as the employment response to productivity increases is low.

A secondary contribution is that our estimated firm-specific labour-supply and pass-through elasticities add to very few such estimates for developing countries, where higher labour surpluses and frictions may increase monopsony power (Bassier 2023). More generally, a popular approach in the literature estimating labor supply elasticities is estimate wage and employment responses to a revenue-productivity shock (Kline, 2025). We note such studies should take care to estimate these elasticities for unconstrained firms, as when shocks take place across constrained firms the estimate will not identify the labor supply elasticity.

Our findings are likely to be applicable to a broad range of countries which have similar Kaitz indices (the ratio of the median wage to the minimum wage) to South Africa. One such case is Portugal, which has an unusually high Kaitz index, and where there is a similar piecewise linear relationship between firm surplus and firm wage policies (Card et al., 2016). More generally, the mechanism we identify may be particularly important for countries seeking to develop through increasing firm productivity while protecting workers with high minimum wages (e.g. Brazil and South Africa), as well as low-income countries with relatively high wage floors associated with subsistence or efficiency wages (Breza et al. 2021; Muralidharan et al. 2023). We caution that a potentially non-trivial fraction of firms may absorb such productivity gains entirely as profits, undermining this development agenda of inclusive growth (Lewis 1954; Verhoogen 2023), and that this is more serious in more monopsonistic contexts.

The core ideas of the model are introduced in Section 2, and the data and context are discussed in Section 3. The evidence using the cross-sectional kink design is presented in Section 4, then the within-firm shock evidence in 5. Section 6 discusses implications for models of imperfect competition in the labor market, and 7 concludes.

2 Theoretical predictions

2.1 Simple model of monopsonistic firms with a minimum wage

Standard argument. We briefly recapitulate a simple model of firm responses to minimum wages under monopsony, following Dickens et al. (1999) and Manning (2003, pp. 338-345).

As is characteristic of monopsony models, we assume a firm-facing labour supply curve $w_i =$

εn_i, where lower case letters denote logs, w_i is the firm wage, n_i is firm employment, and there are many firms. The firm-facing labour supply elasticity $1/\varepsilon$ is constant across firms and is finite. Such an upwards sloping labor supply curve implies a marginal cost of labor greater than the wage for firm i:[2]

$$\text{mcl}_i = \ln(1+\varepsilon) + w_i = \ln(1+\varepsilon) + \varepsilon n_i \,. \tag{1}$$

The marginal revenue product of labour of firm i, is a simple downwards sloping labour demand curve:

$$\text{mrpl}_i = a_i - \eta n_i \,, \tag{2}$$

where a_i is a demand or productivity shifter. The elasticity of the labour demand curve under perfect competition would be $1/\eta$. This can be motivated by a production function such as $Y_i = \frac{1}{1-\eta} A_i N_i^{1-\eta}$, where additional factors such as capital can be log-additively included.

This model setup represents a very simple and general monopsonistic form, and remains agnostic as to the source of monopsony power (e.g. search frictions, amenities or concentration). As such it nests different popular approaches to modeling monopsonistic competition (Azar and Marinescu, 2024; Card et al., 2018; Manning, 2003). Setting marginal product equal to marginal cost, the unconstrained employment and wage for firm i are:

$$n_i^* = \frac{1}{\varepsilon + \eta} \left(a_i - \ln(1+\varepsilon)\right) \tag{3}$$

$$w_i^* = \varepsilon n_i = \frac{\varepsilon}{\varepsilon + \eta} \left(a_i - \ln(1+\varepsilon)\right) \,. \tag{4}$$

When a minimum wage is introduced, a firm finds itself in one of three qualitatively distinct regimes, depending on its productivity a_i. These three regimes are depicted in Figure 1 panel (a) as the three marginal revenue product of labour (MRPL) curves, corresponding to varying a_i in equation 2. The labour supply (LS) and marginal cost of labour (MCL) curves are as described above per equation 1.

If the minimum wage w_m is not binding, i.e. $w_\text{m} \leq w_i^*$, equations 3 and 4 hold and $w_i = w_i^*$ and $n_i = n_i^*$. These are *unconstrained* firms, shown in the first regime (MRPL1) in Figure 1 panel (a), and these firms are not directly affected by the minimum wage.

When the unconstrained wage is lower than the minimum wage ($w_\text{m} > w_i^*$), firms must pay a wage equal to the minimum wage ($w_i = w_\text{m}$). These are *constrained* firms. Within the constrained firms, there are two regimes. Firms for which $w_\text{m} > w_i^*$ but which have their marginal revenue product of labor *above* the minimum wage are *supply-constrained*, shown as regime two (MRPL2) in Figure 1 panel (a). We have the well-known result that for such firms, employment *increases* as a result of the minimum wage, because they attract additional workers at this higher wage, and accept all workers supplied at that wage, $n_i = (1/\varepsilon)\, w_\text{m}$. To allow for a more general framing beyond monopsonistic firms (see Section 6), we sometimes also refer to these firms as *just*-constrained firms.

[2] The intuition for this is that for a monopsonist to hire an additional worker they must increase the wage, which also applies to the wages of already-employed workers. $\frac{\partial WL}{\partial L} = \frac{\partial W}{\partial L} L + W = \varepsilon W + W$.

The other regime of constrained firms are those with $w_{\mathrm{m}} > \mathrm{mrpl}_i$, denoted *demand-constrained* and shown as regime three (MRPL3) in Figure 1 panel (a). These firms *reduce* employment in response to the minimum wage until $\mathrm{mrpl}_i = w_{\mathrm{m}}$. Firms must still pay the minimum wage, but the new employment level is now governed by the firm labour demand constraint, so that $n_i = (1/\eta)(a_i - w_{\mathrm{m}})$.

The more general model, presented in Appendix B.1, incorporates the average market-level wage as a determinant of aggregate labor supply, and allows for a firm-specific labour supply shifter (e.g. disamenities) as well as the firm-specific demand-shifter above (a_i)—but the qualitative regimes are exactly the same.

[Figure 1 here]

Novel insight. We first note that the introduction of a minimum wage changes the "latent" MCL curve to a new "effective" MCL curve indicated by the discontinuous red line shown in Figure 1 panel (a). Since wages cannot be below the minimum wage, and the cost of hiring an additional worker is simply the minimum wage paid to that worker, the marginal cost of labour when the LS curve is below the minimum wage is simply the minimum wage itself. Of particular interest to us is the discontinuity in the effective MCL at L2 on the employment axis, where firms switch from being minimum wage-constrained to unconstrained.

Firm-specific employment (on the x-axis) is set where the MRPL curve intersects with the effective MCL curve, while the wage is marked down from this point (along the y-axis) to the level on the labour supply curve. For firms with productivity MRPL2, this MRPL-MCL intersection occurs in the region of the effective MCL discontinuity. These are the supply-constrained firms. It is easy to see that for local shifts in MRPL2, the intersection with the effective MCL curve remains at the discontinuity point, and subsequently that these shifts in productivity do not change firm employment. They also do not change the wage, which is marked down to the minimum wage level. Instead, local shifts in MRPL in this region are reflected as changes in the size of the markdown from the marginal revenue productivity of labour to the (minimum) wage. The intuition is that firms with a range of MRPL intersecting the MCL at its discontinuity cannot attract additional workers without increasing the wage above the minimum wage (because the LS curve is now above the minimum wage for $n_i > L2$), but as long as their unconstrained wage (read on the labour supply curve in the region $n_i \leq L2$) is below the minimum ($w_{\mathrm{m}} > w_i^*$), there is no incentive for them to do so. Instead the additional productivity per worker is reflected in increased markdowns. The reason there is a discontinuously large cost associated with increasing the wage above the minimum is because this wage increase would also apply to their existing workers currently paid at the minimum wage.

Our main insight, then, is that for supply-constrained firms there is a range of productivity increases (decreases) which do not change the firm's wage or employment, and which are instead reflected in increases (decreases) in the markdown.

2.2 Simulations

We demonstrate these patterns by simulating the model above in Figure 1 panel (b), focusing on a fixed minimum wage with productivity on the x-axis and firm wages, employment and markdowns on the y-axis.[3] Following Manning (2003), it is useful to focus on an "adjusted productivity" term we denote v_i, which determines the regime of a firm:

$$v_i = \frac{\varepsilon a_i}{\eta + \varepsilon} \tag{5}$$

This is just the variable component of equation 4, where firm productivity a_i is adjusted by the elasticities of firm labour supply and demand.[4]

Firms with v_i above some threshold v^* will have $w_i^* \geq w_{\mathrm{m}}$ and will be unconstrained, firms with v_i below v^* but above another threshold v_1^* will be supply-constrained, and firms with v_i below v_1^* will be demand-constrained. Depending on the value of v_i, then, Figure 1 shows the three regimes:

1. Unconstrained (i.e. higher productivity, MRPL1): $\boldsymbol{v_i \geq v^*}$ The right-most region, after the second vertical line (indicating v^*), delineates firms whose optimal monopsony wages are above the minimum wage, and so are not affected directly by the minimum wage. Wages and employment increase in productivity (equations 3 and 4), wages are marked down relative to MRPL as in the standard monopsony optimization, and the markdown level is constant.

2. Supply-constrained (i.e mid/lower productivity, MRPL2): $\boldsymbol{v^* > v_i \geq v_1^*}$ The middle region between the vertical lines is our subject of interest, where the optimal wage is just below the minimum wage. Firms in this region keep wages fixed at the minimum, and do not increase employment as productivity increases. Instead, increased productivity is absorbed in higher markdowns, until the markdown is at the optimal level for an unconstrained monopsonist (at v^*).

3. Demand-constrained (i.e. very low productivity, MRPL3): $\boldsymbol{v_i \leq v_1^*}$ The left-most region, before the first vertical line (indicating v_1^*), shows firms constrained to set wages equal the minimum wage, but with too low productivity to employ all of the workers available at that wage. These firms employ more workers as their productivity increases. MRPL is equal to the minimum wage and there is no markdown.

[3] We impose that MRPL shifters a_i follow a standard normal distribution, we set the market labour supply elasticity to 0.5 and the firm-facing labour supply elasticity to 1.7 (these are based on the cross-sectional patterns, see section 4). The simulations are based on 1,000 observations, each representing a firm. Wages and employment are normalized by comparison to the average wage and employment under the perfectly competitive case—that is, no monopsony nor minimum wage. The minimum wage is set at -0.2 log units. We trim the 1% tails of productivity.

[4] In the fuller model discussed in Appendix B.1, this term v_i also includes the firm-specific amenities shifter, which we abstract from for our purposes here.

In Figure 1 panel (b), 20.8% of firms are supply-constrained and 4.2% are demand-constrained, which compare favorably with our empirical results discussed in Section 4. We provide expressions for v^* and v_1^* in Appendix B.1. The proportion of firms in the supply-constrained region of interest is increasing in the minimum wage, and with ε (more monopsony): $v^* - v_1^* = \varepsilon \ln(1+\varepsilon)/(\eta+\varepsilon)$. Appendix Figure A1 shows in cases with less monopsony (ε smaller) and/or lower minimum wages, the supply-constrained region shrinks and moves down the productivity distribution, capturing fewer firms.

Once again, using the more detailed model with firm amenities and dependence on the average wage does not affect the qualitative pattern above. Appendix Figure A2 shows the same patterns with amenities, and this is robust to alternative assumptions regarding the correlation with firm productivity.

2.3 Model predictions

What are the precise patterns predicted by the model above? In practical terms, even if the mechanisms outlined above are important, the empirical patterns will diverge from the simulations due to the influence of unmodeled factors and measurement error.

Perhaps the most important among these departures from the simulations is that we focus only on the region around v^*, the kink in the wage-productivity curve that divides the constrained and unconstrained firms. We provide more details in the next section, but a central point is to note that our institutional context involves about 40 different minimum wage systems, with correspondingly different threshold points. In our empirical exercise we pool these different regimes and re-center around the same adjusted productivity threshold to ensure enough statistical power to test our predicted patterns. We have to choose to recenter around either v^* or v_1^*, because there is little reason to expect a similar productivity range between these thresholds across minimum wage regimes. The v^* threshold is the more natural point as it contains the wage kink, and is the distinguishing point between our firms of interest (the supply-constrained firms) from the bulk of firms in the data, which will be unconstrained. The number of demand-constrained firms that we observe is likely to be small, because firms with very low productivity draws will be unobserved or under-represented in our actual firm data due to informality, fixed costs and endogenous exit (Olley and Pakes 1996; De Loecker and Syverson 2021).

How does this affect the model-predicted patterns? Figure 2 simulates a scenario much closer to our empirical context, where the simulation above (Figure 1 panel (b)) is repeated for 40 different labour markets (i.e. differing minimum wages and labour supply elasticities), and then firms are re-centered around the labour market-specific wage-kink v^*. The key implication is that we do not observe a clean kink point within the constrained region, dividing demand-constrained and supply-constrained firms, with the markdown flat and employment sharply rising to the left of this kink point. The wage-kink (v^*, dividing just-constrained and unconstrained firms) is still very apparent and we still have differential slopes on either side of this kink point for wages, employment and markdown. The constrained region does have some positive slope in employment, because it contains both demand- and supply-constrained firms, but the key point is that it is much flatter than the unconstrained slope.

[Figure 2 here]

Another point of divergence between the model predictions and empirical observation is that productivity will be measured with error, introducing noise into the simulated kink-point υ^*. In addition, many variables omitted from the model are empirically correlated with productivity and affect the outcomes of concern, such as skill composition (the model assumes undifferentiated labour). For these reasons our tests rely on *differential* slopes around the kink-point, much like, for example, the literature on UI benefit floors using regression kink designs (Ganong and Jäger 2018).

In particular, we test for the following patterns:

1. **Rent-sharing or wage-productivity curve:** The two key predictions regarding the co-movement of wages and productivity are (a) there exists a kink-point $\hat{\upsilon}^*$ on the adjusted productivity axis at which the linear wage-productivity slope changes, corresponding to the model point υ^*, and (b) the linear slope to the left of this point is smaller than the slope to the right, which is positive.

2. **Markdown-productivity curve:** At the same kink-point $\hat{\upsilon}^*$, the linear slope to the left is positive and higher than the slope to the right.

3. **Employment-productivity curve:** At the same kink-point $\hat{\upsilon}^*$, the linear slope to the left is lower than the slope to the right, which is positive.

The next sections take these testable hypotheses to the data, first in section 4 using a cross-sectional kink design and then in section 5 using within-firm shocks.

3 Data and Institutional Context

3.1 Data

For our empirical analysis we use the administrative National Treasury-South African Revenue Services (NT-SARS) tax data held at the National Treasury Secure Data Facility (NT-SDF) in Pretoria. This is restricted-access data which can only be accessed in person at the NT-SDF for approved projects.[5] The data we use consist of annual firm balance sheet information from Company Income Tax returns ("ITR14" forms) which include information such as sales, costs, profits and industry; matched worker-level annual payroll data ("IRP5" forms) which can be used to construct firm-level employment, (approximate) monthly wages, and firm geographic location; and linked firm-level monthly customs data, which provides transaction value and

[5] This research was approved under the auspices of the SA-TIED programme workstream 1.

country of origin and destination for imports and exports respectively. The data constitute a panel covering the universe of formal-sector firms in South Africa, and while each dataset covers different periods they all reliably cover at least the period from the 2010 to 2019 tax years (approximately the 2009-2018 calendar years).[6] For the trade shock analysis discussed in Section 5.1.2 we use GDP data from the World Development Indicators.

3.2 Minimum wage institutions

Prior to January 2019, a multilayered wage legislation system operated in South Africa, where minimum wages were set by the government for selected broad industry-locations ("Sectoral Determinations", SDs), or by publicly-recognized Bargaining Councils (BCs) consisting of employers and employees at the sub-industry-location level. Minimum wages can vary substantially by these sectors, and we therefore examine firms separately by their BC or SD.[7]

BCs cover industry-regions, and are constituted by trade union and employer representatives who negotiate industry-region minimum wages. Supplementary establishment-level wages can then also be negotiated above these minima, allowing for rent sharing. This is a setup common to a variety of European countries (Bhuller et al. 2022; Jäger et al. 2022), but unlike in some of these countries BC agreements are routinely extended to include non-unionized workers. We identify BC firms in the SARS-NT data by matching firms according to their industry and location, using the Bassier (2022) dataset of BC agreements. There are 39 private sector BCs; after restricting for key missing variables we identify 30 in the data, which cover approximately 26% of the (formal sector) workers in our sample. This dataset also provides a minimum wage associated with each BC for each year, but it is highly approximate: we must take the lowest BC-specified wage to be the minimum, even though BC agreements typically specify multiple occupation-specific wages, because occupations are not observed in the NT-SARS data.

SDs are government-set wage minima (and conditions of employment) for sectors not fully covered by BCs, often because they are understood as "hard to organise". There are 11 SDs, 8 of which set minimum wages for formal sector workers. SDs are defined more expansively than BCs, and sometimes overlap with BCs; in these cases the BC minimum wages apply. While SDs, like BCs, may set occupation- and location-specific wages, there is usually less heterogeneity in minimum wages than in BCs. We identify SD firms in the NT-SARS data by matching their industry and location to a dataset we create from promulgated government regulations. We identify the 8 formal sector SDs, which exclusively cover about 32% of (formal sector) workers in our sample. These are predominantly workers at the lower-end of the wage distribution, unlike BCs which have coverage concentrated in the upper half of the wage distribution (Bassier 2022). We also include minimum wages from these regulations, but these are approximate for the same

[6] See Appendix C for additional details concerning our use of this NT-SARS data.

[7] A national minimum wage (NMW) was introduced in January 2019 which supercedes a small number of these minimum wages. Only the last 2 months of our observed period overlap with the period of the national minimum wage, so we ignore this given that any dynamics in these months are likely to be irrelevant for our results in our empirical designs.

reason as the BC minima. The BC and SD monthly minima in the 2018 tax year are shown in Table A1.

4 Evidence from a cross-sectional kink design

Implicit in the model above is a measure of firm productivity. We begin by estimating this productivity, and then test for the patterns in Section 2.3 by identifying and using kinks in the cross-section of firms.

4.1 Production function estimation

In estimating firm-specific productivity we draw from a substantial Industrial Organization literature concerned with production function estimation (Olley and Pakes 1996; Levinsohn and Petrin 2003; Ackerberg et al. 2015; De Loecker and Syverson 2021). Recognising issues with OLS estimation of productivity such as simultaneity/transmission bias and selection/survival bias, we estimate productivity using the proxy variable/control function method of Ackerberg et al. (2015) (ACF) with materials as the proxy variable, probably the leading approach in the literature (De Loecker and Syverson 2021; Yeh et al. 2022). Cognisant of the Gandhi et al. (2020) critique of attempts to estimate gross output production functions using proxy variable methods, we specify a value-added production function with a flexible translog form:

$$y_{it} = \beta_l l_{it} + \beta_{ll} l_{it}^2 + \beta_k k_{it} + \beta_{kk} k_{it}^2 + \beta_{lk} l_{it} k_{it} + \omega_{it} + \varepsilon_{it} \tag{6}$$

where y_{it} is value-added for firm i in period t, l_{it} is firm employment and k_{it} is firm capital stock, all in logs, while $\omega_{it} + \varepsilon_{it}$ is the productivity residual made up of productivity shocks which are observed or predictable for the firm at time t (ω_{it}) and those which are not (ε_{it}).

We show in Appendix Figures A8 and A9 that our main results are robust to a variety of alternative methods of estimating production functions: ACF with a Cobb-Douglas functional form, the Olley and Pakes (1996) method, the Levinsohn and Petrin (2003) method, the ACF correction applied after Olley and Pakes (1996) rather than Levinsohn and Petrin (2003) estimation, and the ACF method estimated separately by various industry categories.

4.2 Empirical strategy: cross-sectional kink design

Overview. Our first step is to identify a kink ("knot") (i.e. a threshold of a piecewise function) in the wage-productivity curve. The pattern we look for is as follows: Wages (measured as firm medians) are a piecewise linear continuous function of estimated productivity, defined over two intervals, and containing a discontinuity in its derivative (the knot) at the boundary between the intervals. We identify such a knot in the observed distribution by running two OLS regressions, one to the left and another to the right, for each point in the productivity distribution, and selecting the point or threshold which maximizes the R-squared.[8] This is a candidate for the

[8]This procedure is analogous to that used by Card et al. (2016) to identify a similar kink in the distribution of AKM firm wage premia against firm log value added.

productivity threshold v^* in the model where firms move from being supply-constrained to unconstrained.

Given such a wage knot denoted $\hat{v}^*$, we test the predictions from Section 2.3. That is, at the same productivity threshold, there are significantly different slopes on either side (this is what produces the kinks) in the cross-sectional firm wage (by construction), markdown and employment curves along the productivity axis, and these differential slopes are of the predicted sign. We separately regress each of these variables on productivity to the left and right of $\hat{v}^*$, which allows us to examine whether there is indeed a change in the slope as we expect.

Estimation details. Aside from firm productivity, the markdown is of course also unobserved, and so we use the gross profit share as a proxy. This is defined as gross profits over gross profits plus the firm wagebill, and so is equivalent to one minus the labor share as it is defined in Gouin-Bonenfant (2018). We (imprecisely) also refer to this as the "capital share", borrowing from the macroeconomics literature which divides income into labor and capital income.

Due to the different minimum wages which operate in each BC/SD, the above knot-finding exercise is implemented separately for firms in each BC and SD. We estimate productivity for each firm in "pre-period" windows, and then the knot-finding exercise above is implemented only for the years after this pre-period.[9]

With the BC- and SD-specific kink-points and cross-sectional patterns having been separately estimated, we then pool the results across all the different BCs and SDs. As discussed above, in order to account for the different minimum wages and other market-level characteristics of each BC and SD, which will necessarily lead to different v^* wage-kink productivity thresholds, before pooling we re-center productivity in each BC and SD around the estimated wage-kink productivity threshold $\hat{v}^*$ in that BC/SD, so that re-centered productivity above 0 indicates an unconstrained firm and below 0 indicates a constrained firm. We then use this pooled sample containing all BCs and SDs for testing differential slopes and for plotting.[10]

Finally, though we do not focus on the demand-constrained region, we do try to isolate it from the supply-constrained region by identifying a kink on the employment-productivity curve to the left of the wage threshold (i.e. $\hat{v}_1^*$) for each BC/SD and in the pooled aggregate. In practice, this point is not well-identified, for the reasons discussed in section 2.3.

[9]See Appendix D for a description of how we construct these pre-periods. We define firm productivity as the average of firm-year-specific productivity in the pre-period, winsorized at the 1% tails. De Loecker and Syverson (2021) note, in a related context, that this averaging may reduce misspecification error.

[10]In a few cases our knot-finding algorithm does not identify a plausible interior wage-kink $\hat{v}^*$ and instead identifies a kink at extreme wage values; in order to exclude these cases we trim the estimated wage-kinks $\hat{v}^*$ at the 1st and 99th percentile of the pooled firm distribution. However these cases are few and our results are robust to whether or not we trim here.

4.3 Results from the cross-sectional kink design

Figure 3 shows the results from the pooled kink design, with no controls in panel (a) and with controls in panel (b).[11] The vertical dotted line indicates the productivity threshold corresponding to the estimated wage kink productivity threshold $\hat{v}^*$.

The pooled aggregate case clearly exhibits the predicted features of the model. To start, the wage curve (blue) has a relatively flat slope to the left and is more upwards-sloping to the right of the wage-kink at $\hat{v}^*$. Though there is a break in this relationship at the wage-kink $\hat{v}^*$ by construction, the result that the signs and differential magnitudes of these slopes match the predictions is not by construction. Moreover, the flat portion corresponds very closely to the level of the average minimum wage (indicated by the horizontal dotted line in panel (a)), despite the minimum wage not being used in the estimation of the kink point. This strongly supports the estimated kink point $\hat{v}^*$ as identifying the productivity threshold v^* below which firms are constrained by the minimum wage. Further supporting this, if we calculate a Kaitz index (the minimum wage divided by median wage, a measure of the bite of the minimum wage) for each BC/SD or a firm-specific Kaitz which depends on their granular labour market, we find that firms with a higher Kaitz index are more likely be found in the constrained rather than unconstrained region, supporting that it is the minimum wage bite which drives the productivity level of the estimated wage-kink and the region in which a firm falls (see Appendix E).

Even more strikingly, Figure 3 depicts the predicted kink patterns for firm employment (green) and the profit share (red), despite these variables not being used for kink estimation and therefore not having any by-construction relationship to the kink-point. Specifically, for firm employment (green), the slope is flat to the left of the wage-kink threshold $\hat{v}^*$, and then sharply increases to the right. Just as strikingly, the slope of the profit share (red) is increasing to the left of the wage-kink threshold, and then changes to be near-flat to the right of the threshold. The plot is remarkably similar to the simulated model prediction plot in Figure 2, and even more so for Figure 3 panel (b) which includes controls. The differences in the slopes around the wage-kink threshold are statistically significant at the 1% level (see Table 1). Because there is no by-construction relationship between firm size or the profit share and the wage-kink estimation routine, we regard these observed patterns as a compelling test and validation of the model predictions.

[Figure 3 here]

In Appendix Figure A4 panels (a) and (c) we superimpose the density of firms at each productivity point. A substantial proportion of firms are just to the left of the threshold. If we

[11] Specifically we control for detailed industry-by-location cell (2-digit industry by District Council), average AKM worker fixed effect at the firm (Abowd et al. 1999), and poaching ratio at the firm (Bagger and Lentz 2019). Like with productivity, the average AKM worker effect at the firm and the poaching ratio are estimated in a "pre-period" which is not used when calculating the cross-sectional relationship.

weight each firm equally, 24% of firms are found in the constrained region.[12] This suggests that the mechanism we discuss affects about 1 in 4 firms in South Africa, and compares favorably to our simulations in Section 2.2.

Table 1 reports the implied cross-sectional wage, employment and profit relationships, separately for the constrained and unconstrained regions. As one would expect, the wage (rent-sharing) and employment elasticities are lower in the constrained region (and indeed the employment elasticity is not statistically significantly different from zero), and the profit elasticity is markedly higher.

[Table 1 here]

While the same patterns hold for most of the individual BCs and SDs (Appendix Figure A3), they are sometimes noisy or simply not evident for particular BCs or SDs. We view this as unsurprising, given that we are testing a strong prediction, which in any case may not be detectable given the unavoidably approximate nature of our productivity estimation routine and varying BC/SD sample size. The pooled figure above is robust to several Appendix checks: Figure A4 panel (a) uses 50 productivity bins instead of 20, panel (b) re-estimates $\hat{v}^*$ based on the pooled sample rather than just re-centering separate BCs and SDs, and panel (c) fits the linear slopes on the full underlying distribution without trimming; Figure A5 shows the kink patterns occur for various percentiles of wages, employment and profits at $\hat{v}^*$, i.e. not just the medians or associated with a particular outcome level; and Figure A6 restricts to only one "event" per firm rather than the stacked event structure. It is also noteworthy that the model assumption of log-linearity appears strongly supported for all three variables with regard to productivity, on both sides of the kink (Figure 3); again, there is nothing mechanical about this as the observed empirical relationship could easily have been non-linear.

Perhaps of interest is what the magnitudes in the unconstrained region can tell us about the model parameters. We estimate the elasticities above with respect to value added in Appendix Table A2.[13] The implied pass-through or rent-sharing elasticity of 0.438 (0.0061) is *substantially* higher than most identified estimates of this parameter in prior literature, while the implied firm-facing labour supply elasticity, 0.614 (0.021) is substantially lower (see Section 5 for discussion of these prior estimates). This is not surprising: these cross-sectional estimates are analogous to estimates of the "employer size-wage effect" discussed in Manning (2003), which under-estimate the firm-specific labour supply elasticity due to confounding by unobserved worker quality and amenities. Empirically, more highly skilled workers are indeed more likely to work in more productive, larger firms, creating a positive correlation between wages and employment which is not due to the shape of the labour supply curve. This also explains why the rent-sharing

[12] Without weighting, 20% of observations falling in the constrained region, since firms which appear in the panel multiple times are disproportionately unconstrained.

[13] We estimate these elasticities by running a cross-sectional regression of the (log) outcome on (log) value-added, where value-added is instrumented with the recentered productivity value. Standard errors are clustered at the firm and event level.

elasticity may be upwardly biased.

These issues mean the slopes are likely biased (in predictable directions) if one seeks to use them to calculate rent-sharing or labour supply elasticities, and they should not be used to quantify these magnitudes. Yet the purpose of the cross-sectional kink design is not to estimate these magnitudes; rather, it is the sudden statistically significant *breaks* in the estimated magnitudes around the wage-kink threshold which are of interest, and which validate the model predictions, in the spirit of a regression kink design. The threat to the validity of this test of the model predictions would be if these breaks in the wage- employment- and profit share-productivity relationship were caused by discontinuous *breaks* in the effects or distribution of unobserved confounding variables exactly where minimum wages bind. If the confounding effect is somewhat constant across the distribution of firms, or varying but just not discontinuous at exactly the wage-kink threshold, this kink design still provides good evidence for our main theoretical predictions. And indeed it is quite difficult to come up with an explanation of why confounding effects would break directly at the point where the minimum wage binds, especially given that the kinks occur across various percentiles of wages, employment and profit share (Appendix Figure A5), and that the characteristics of constrained and unconstrained firms are not very different, especially around the wage-kink threshold (Appendix Table A3).

5 Responses to within-firm shocks

We complement the evidence from the cross-sectional kink design by looking at heterogeneity in within-firm *responses* to revenue-productivity increases, for firms along the productivity distribution. This design is also less susceptible to biases in the magnitudes discussed above, such as firm skill composition, and should produce identified estimates of the rent-sharing and firm-specific labour supply elasticity, as well the employment- and profit share-revenue productivity elasticities. To this end, we follow a large number of papers which use shocks to firm value added to identify firm rent sharing and employment responses (e.g. Amodio and De Roux 2022; Kroft et al. 2023; Saez et al. 2019). In particular, we follow the approach in Lamadon et al. (2022) (LMS) with both an "internal" and "external" instrument for value added. Also following LMS, our wage measures focus on workers who stay at the same firm over the full estimation period. This also helps address potentially confounding worker quality and other compositional issues such as tenure effects, which could otherwise bias our estimates.

5.1 Empirical strategy: within-firm shock design

Following the pooling procedure used in our kink design, we use the estimated re-centered productivity and associated threshold $\hat{v}^*$ (see Section 4) to test for heterogeneous responses to the left (constrained firms) compared to the right of the threshold (unconstrained firms).

The testable hypotheses stated at the end of Section 2 were, aside from the kink-point which is assumed from the previous section, differential slopes on either side of the kink point for each of the outcome variables with respect to productivity. In the context of firm responses to shocks to marginal revenue productivity (such as demand shocks), the hypotheses regarding differential slopes translate into differential marginal responses, i.e. higher responses to the right of $\hat{v}^*$ for wages and employment, and lower responses to the right of $\hat{v}^*$ for profit share. We test this directly by estimating the heterogenous marginal responses to value added shocks along the firm productivity distribution, using two approaches described below.

As in the cross-sectional kink design, the key test is *differential* responses around the threshold $\hat{v}^*$, rather than 0 wage or employment response in the constrained region, because practical considerations mean the theoretically-implied 0 response is unlikely to be identified in our empirical exercise. One reason is simply noise due to measurement and sampling error: our threshold between constrained and unconstrained firms is an estimated quantity rather than an exact delineation, based on estimated firm-specific productivity, so there will be some unconstrained firms in the constrained region (around the threshold), and vice versa. Another more systematic issue is that constrained firms receiving a value added shock may cross over into the unconstrained region, thus exhibiting (attenuated) behavior associated with the unconstrained region. Indeed, a back-of-envelope calculation suggests that approximately 20% of constrained firms move into the unconstrained region as a result of the internal shock.[14] Thirdly, minimum-wage compliance in South Africa is far from perfect (Bhorat et al. 2012; Bassier and Ranchhod 2024), which will be another reason that firms identified in the constrained region will actually be unconstrained and exhibit unconstrained responses. And lastly, the unconstrained region contains demand-constrained as well as supply-constrained firms, and demand-constrained firms will have employment responses to productivity shocks, and no profit share response, like firms in the unconstrained region. Fortunately, these issues would attenuate the differential responses between the constrained and unconstrained regions, making a differential response a more demanding test of the model predictions. This means that even when Section 2 provides the "correct" model, in our empirical exercise we should expect a differential response rather than the exact patterns Figure 2.

5.1.1 Internal instrument

The core of the LMS "internal" instrument method is a firm-level event study analysis where treatment is defined as an above-median increase in firm value-added between periods -1 and 0, with some additional specification and variable- and sample-definition restrictions. LMS focus on the effects of these binary value-added shocks on earnings. We extend this firstly by also

[14]To calculate this we run a firm-level cross-sectional regression of the re-centered productivity measure on the log of value-added (with various controls), finding a coefficient of between 0.32 and 0.3, depending on what controls are specified. Multiplying this by the size of the value-added shock in the constrained region (0.34 log points), we find that the value-added shock increases constrained firms' revenue-productivity measure by between 0.10 and 0.11 log points. Empirically, 19-20% of constrained firms are within 0.10 to 0.11 re-centered productivity log points of the threshold, and so approximately 1 in 5 constrained firms transition into the unconstrained region as a result of the TFP shock.

examining effects on employment and the capital share, but most importantly by examining heterogeneous responses along the firm productivity distribution. A related exercise is undertaken by de Frahan et al. (2022), who also extend LMS to examine heterogeneous effects on employment as well as earnings along the firm size distribution.[15]

Examining heterogeneous responses along the firm productivity distribution requires examining responses to comparably-sized shocks; and yet the binary shock may lead to differently-sized shocks to value-added along the distribution. We therefore normalize the size of the shock by using the binary treatment as an instrument for changes in firm value-added. This accounts for any systematic differences in value-added responses to the binary treatment in each of the heterogeneity regions.[16] This also allows us to estimate full elasticities, comparable to other estimates in the literature, rather than reduced-form semi-elasticities. The reduced-form results are relevant mainly insofar as they allow for examining the shape of the event study response, to assess issues such as parallel pre-trends, but the reduced-form results are in any case similar to our main results, statistically significant (levels and differences between constrained versus unconstrained), and consistent with the model predictions.

The sample and event definitions are important to the LMS reduced form analysis, and we closely follow them. While we outline these restrictions in detail in Appendix D, the core of the approach is to construct events which use a balanced panel of firms which have a minimum number of worker "stayers" who remain employed at the same firm for the period of the event. For our baseline specification we keep firms which have at least 2 stayers.[17] We stack four 6-year events (2010-2017, 2011-2018, 2012-2019, and 2013-2020 tax years), where we treat the first three periods as the pre-period and the latter three as the post.

For the internal LMS instrument approach, treatment is defined as an above-median increase in firm value-added between periods -1 and 0 for each event, where the median increase is weighted by firm size. As in LMS, period -2 is used as the omitted reference period to allow for some mean reversion dynamics in period -1, while period -3 is used to assess pre-period parallel trend violations. For the same reason, periods 1 and 2 are considered the post periods of interest, rather than period 0 (results are essentially unchanged if we use only period 2 instead).

[15]While not a focus of this paper, we note that we have been able to replicate the de Frahan et al. (2022) findings in our data, finding qualitatively similar results.

[16]It is equivalent to dividing the size of the reduced form response in the outcome by the size of the value-added shock that the binary treatment induces in that heterogeneity region. In our baseline specification this is 0.34 log points in the constrained region versus 0.29 log points in the unconstrained region.

[17]This is our only notable divergence from LMS, who use a 10-stayers minimum as their baseline. A 10-stayers minimum is overly restrictive when it comes to the South African firm-size distribution and a labour market context defined by high churn (Kerr 2018). We use a 2-stayer minimum to be as nonrestrictive as possible but to mitigate measurement error in one-stayer firms where the one stayer may be an owner or otherwise unrepresentative of general employer/employee dynamics. In Appendix Figure A10 we show that our main results are not sensitive to the number of stayers.

The aggregate reduced-form event study reduced form regression is:

$$y_{i,t,e} = \lambda_{i,e} + \gamma_{t,m(i),e} + \sum_{s=-3,s\neq-2}^{2} \beta_s \times \mathbb{1}[t = s] \times D_{i,e} + \varepsilon_{i,t,e} \tag{7}$$

where $y_{i,t,e}$ is the log of the outcome for firm i at time t in labor market m for event e, $\lambda_{i,e}$ is the firm-event fixed effect, $\gamma_{t,m(i),e}$ is a time-varying market-event fixed effect, $D_{i,e}$ is the treatment variable and the β_s are the coefficients of interest, relative to period -2. Standard errors are clustered at the market-event level. The market controls $\gamma_{t,m(i),e}$ constitute 81 labor markets (interacted with event and year), made up of province and 1-digit industry.[18]

While we present these reduced form results as in LMS, primarily to assess pre-trends, as noted above in our main specification we use the binary treatment to instrument for changes firm value-added. Specifically, we run long-differences regressions of the log change in the outcome between the post-period (periods 1 and 2) and period -2 on the equivalent change in value-added, with the change in value-added instrumented by the binary treatment. We include fixed effects analogous to those in equation 7 and again cluster at the market-event level.

5.1.2 External instrument

While the "internal" approach is a greater focus of their paper, LMS also implement a supplementary "external" instrument specification, where they use firms receiving a procurement contract to define treatment. We use a different external instrument, relying on shift-share trade shocks similar to Garin and Silvério (2023) and Bassier and Manning (forthcoming).

We use the South African administrative customs dataset to define *firm-specific* shift-share trade instruments based on the country composition of firm exports and imports, combined with movements in destination- and origin-country GDP, respectively. Specifically, for the export instrument, we define firm i's shock in period t as,

$$D_{i,t}^{exports-IV} = \sum_{d} \alpha_{i,d}^{export-share} \frac{\psi_{d,t}^{shift}}{\bar{\psi}_d} \tag{8}$$

where their export exposure "share" $\alpha_{i,d}^{export-share}$ is the proportion of their exports made up of exports to country d across the sample period (i.e. a firm-specific constant), and the "shock" $\psi_{d,t}^{shift}$ is the GDP of country d in that year t normalized by the mid-period GDP of that country $\bar{\psi}_d$. Such foreign GDP movements constitute positive revenue-productivity shocks for exporters, since those countries demand more goods. We define an analogous import shock $D_{i,t}^{imports-IV}$, which is defined the same as in Equation 8 except that the term $\alpha_{i,d}^{export-share}$ is replaced with $\alpha_{i,d}^{import-share}$, which is the share of firm i's imports from country d.

A notable difference with the internal instrument is that the shift-share treatment can only be defined for trading firms, and this decreases the sample size considerably, especially for

[18] Appendix Figure A11 shows results when different industry and geography variables are used, including using the 2-digit industry and the "district" geography variable, which creates 2600 labor market interactions. Our main results are essentially unchanged.

low-productivity constrained firms. This is unsurprising; it is well-known that exporters are generally higher-productivity firms (Verhoogen 2023). The scale of the reduction is however quite dramatic: while the estimation sample for the main internal instrument specification is made up of 6,167 distinct constrained firms and 34,120 distinct unconstrained firms, when this is restricted to trading firms for the external instrument approach this reduces the equivalent specification's sample to 956 distinct constrained firms and 12,017 distinct unconstrained firms. While the overall number of firms in the sample reduces by just over one third, this is particularly concentrated among the constrained firms, which drop from being about 18% of the sample in the internal instrument case to being about 8% of the sample in the external instrument setting. This dramatically reduces statistical power , rendering us unable to present an event-study or estimate results by fine-grained productivity bins.[19] Instead, we define only two bins (constrained firms versus unconstrained firms) and estimate a fixed effects regression using the full sample period (no split into pre- and post-periods).[20]

The reduced-form specification is very similar to 7 above, but without the event study terms:

$$y_{i,t,e} = \lambda_{i,e} + \gamma_{t,m(i),e} + \beta_{export} \times D_{i,t}^{exports-IV} + \beta_{import} \times D_{i,t}^{imports-IV} + \varepsilon_{i,t,e} \tag{9}$$

As for the internal instrument approach, for our main results we use these shocks (in this case $D_{i,t}^{exports-IV}$ and $D_{i,t}^{imports-IV}$) to instrument for firm value added.

5.2 Results from within-firm shocks

5.2.1 Set-up and reduced form results

The reduced-form aggregate event-study results for the internal IV are shown in Appendix Figure A7 and Appendix Table A4; with panel (a) of Figure A7 showing the aggregate response across the sample and panel (b) the responses along the recentered productivity distribution. The key outcome variables are the median wages of stayers at the firm (blue), employment at the firm (green), and profit share at the firm (red). The sample is a balanced stacked panel. When estimating responses along the firm productivity distribution (panel (b), we essentially estimate equation 7 separately by productivity bin. The key differentiation is binary: constrained versus unconstrained, which means separate regressions for firms below versus above the re-centered productivity threshold ($\hat{v}^*$, shown with a vertical dashed line in panel (b)). These estimates are shown as horizontal dashed lines with their associated 95% confidence intervals. We also present estimates by 10 approximately equally-sized productivity bins to get a sense of the shape of response along the distribution. The construction of these bins is discussed in Appendix D, as

[19] Indeed, if we implement the LMS *internal* instrument approach on the trade shock sample, we find that while magnitudes are qualitatively in line with our model predictions and similar to our main results in Table 2, the estimates are much more imprecise.

[20] In Appendix Table A6 we report results for a specification where the export and import shares are estimated only over the pre-periods used in the internal instrument approach, and GDP shocks are only considered in the equivalent post-periods. While standard errors are much wider and magnitudes vary, the qualitative conclusions are the same as in the baseline specification results we discuss in the following sections.

well as various trimming decisions (Appendix Figures A12 and A13 show our results are robust to however we do or do not trim.)

The aggregate reduced form response in Figure A7 panel (a) is figure is very similar for the comparable variables in LMS and de Frahan et al. (2022), which also suggest some mean reversion in VA and other dynamics in period -1, but which are small relative to the size of the post-period effects (and seems to dissipate between period 0 and period 1 in any case). The heterogeneous reduced form result in panel (b) is notable for two main reasons: 1) As a preliminary result, the reduced form response patterns (semi-elasticities) support the hypotheses of Section 2.3. The wage and employment responses are higher in the unconstrained region, while the profit share is lower, and these responses differentially break around the threshold $\hat{v}^*$. Differences between constrained versus unconstrained are statistically significant (Appendix Table A4). 2) Pre-trend coefficients for period -3 (lighter-shade estimates with dashed lines), in contrast, are generally not statistically significantly different from zero, and always economically much smaller than the post-period effects.

5.2.2 Main results

Figure 4 panel (a) and Table 2 panel (a) shows our main results from the internal instrument specification, with wage, employment and profit share responses as elasticities with respect to the induced changes in value-added. It shows very clearly the pattern predicted by the theory, with highly statistically significant differences between constrained and unconstrained firms in their wage, employment and profit-share responses, which break around the estimated wage-kink $\hat{v}^*$.

[Figure 4 here]

Figure 4 panel (b) shows the hires outcome using the same method, with an even starker contrast between the constrained and unconstrained firms. It is arguably at exactly the hires margin of employment that the core model mechanism of section 2 operates, as firms tend to dynamically adjust on the hires margin while the other margin, separations, is less in their control.

These patterns are robust to a number of concerns. Appendix Figures A8 and A9 show robustness to alternative methods of estimating the underlying production functions, Appendix Figure A10 shows robustness to the number of stayers, Appendix Figure A11 shows robustness to the choice of time-varying labour market fixed effects, Appendix Figure A12 shows robustness to the proportion of firms dropped around the estimated threshold, and Appendix Figure A13 shows that the patterns are not driven by firms in the tails of the recentered productivity distribution.

The magnitudes reported in Table 2 panel (a) imply that in the constrained region, relative to the unconstrained region, the wage elasticity (i.e. rent sharing) is 29% lower, the employment elasticity is 25% lower, the hires elasticity is 41% lower, and the profit share elasticity is 175%

higher. All these differences are highly statistically significant and consistent with the model predictions.

[Table 2 here]

Table 2 Panel (b) presents the results from the external IV, the shift-share trade shocks. Recall that the source of variation is completely different to the internal IV, based on GDP changes in the export destinations and import origin countries, of only the sample of trading firms. Yet the resulting elasticities are remarkably similar, providing reassurance that since two very different instruments give very similar estimates, then there is likely to be little bias. The column showing the differences in the two slopes shows that in the constrained region the elasticity for rent sharing is 0.059 log points lower (0.054 for the internal IV), for employment is 0.164 log points lower (versus 0.148), for hires is 0.414 (versus 0.280), and for profit share is 0.148 higher (versus 0.142). As noted above, the shift-share strategy has less power than the internal IV, with the demanding test for elasticity differences across regions not quite significant at the 5% level for wages ($p = 0.051$) and employment ($p = 0.074$), but for hires and profit share still statistically significant at the 5% level.[21]

Figure 5 illustrates just how close the estimates are between the internal and external IVs for the main outcomes (Appendix Figure A14 shows the same graph for hires). It is worth highlighting that not only are the differences close, but also the separate estimates for the elasticities in the constrained and unconstrained regions. These region-specific elasticities are all highly significant.

[Figure 5 here]

5.2.3 Discussion

The estimates in the unconstrained region are of independent interest since, as discussed earlier, these are interpretable as identified estimates of the corresponding elasticities. Several papers use similar methods in estimating these elasticities, but we highlight here that we are additionally able to isolate the unconstrained region (noting that such elasticities are unidentified in the constrained region). Focusing on the external IV shift-share estimates, the elasticity of the wage with respect to value-added (i.e. the rent-sharing elasticity) is 0.19 (standard error 0.01), and for employment is 0.64 (0.02). The rent-sharing elasticities are very similar to what have been found in the existing literature, for example between 0.14 and 0.16 in Card et al. (2016), and between 0.13 and 0.19 in Lamadon et al. (2022). These are much smaller than the likely biased elasticities implied by naive interpretation of the magnitudes in the cross-sectional kink

[21] Appendix Table A6 presents results for the severely under-powered but potentially cleaner "pre-post" trade shock specification discussed in footnote 20. While standard errors are much wider and magnitudes do change, it is evident that the point estimates still support model predictions.

design, as expected.

Estimates of the rent sharing *level* (firm wage bill over firm value added) are presented in Appendix Table A5. These estimates are approximate and do not exactly recover the model-consistent measures, because the rent sharing elasticities are for stayers while the wage bill is for the whole firm, but nonetheless prove interesting. The rent-sharing *levels* are much higher for constrained firms than unconstrained firms (approximately 0.9 versus 0.3); this is despite the rent-sharing *elasticities* being higher for unconstrained firms. This exactly conforms with the model predictions: rent sharing *levels* are high due to the minimum wage, even as minimum wage-bound firms only minimally share rents *dynamically* with workers, in response to revenue-productivity shocks. These differences are also reflected in estimates of the pass-through level (the dollar increase in wages for a 1 dollar increase in value added), which is the product of the rent sharing level and elasticity.

The implied firm-facing labour supply elasticity is just the employment elasticity divided by the wage elasticity. Our estimate of 3.4 (standard error 0.23) is well within the range of estimates found in the existing literature, for example 2.5 in Amodio and De Roux (2022), 4.2 in Bassier et al. (2022), 2.5 in Dal Bó et al. (2013) and 4.6 in Datta (2023).[22] Interestingly, the ratio of differences (rightmost column) between the employment and wage elasticities implies a fairly similar elasticity of 2.8. If we take the approach as in the regression kink design literature, and think of crossing the productivity threshold v^* as the "treatment", then these differences in elasticities across the regions can be thought of as a complementary identification strategy for the firm labour supply elasticity focusing on the wage-setting mechanism (activated at the kink point v^*) separately from any other potential mechanisms driving rent sharing and employment responses. Indeed, if labour supply elasticities are heterogeneous, then typical estimates get a weighted average at best while a kink design such as ours pinpoints a LATE at our threshold point v^* (Kline, 2025).

Finally, the profit share elasticity in Table 2 is also of interest especially to the monopsony literature, where there is little evidence on how much of their monopsony power and implied *potential* markdown firms actually exploit (Bloesch and Larsen 2023; Manning 2021). Using the approach above of using the differenced elasticity as an indication of the wage-setting or

[22]It is notable that the implied labour supply elasticity in the constrained region is very similar to the unconstrained elasticity, at 3.6 (1.06). As explained in Section 2, our model implies that the labor supply elasticity is in fact not identified in the constrained region. The similarity may be due to chance (we do not predict that it will be biased in a particular direction, just unidentified and indeed undefined per a literal interpretation of our model), but the striking similarity does perhaps support our suggestion that a substantial reason for positive wage and employment elasticities in the constrained region is because some unconstrained firms are actually located there, either due to noise in our estimated threshold $\hat{v}^*$, or because some firms are not truly constrained due to minimum wage non-compliance. In contrast, this suggests that the presence of demand-constrained firms is not a substantial issue, because that would increase the employment elasticity without increasing the wage elasticity, thus biasing the implied labour supply elasticity upwards. This seems quite plausible to us given the reasons we outlined in Section 2 for why low-productivity demand-constrained firms will be disproportionately unobserved in our data, which is only exacerbated by the balanced-panel stayer restrictions used in the within-firm shock specifications.

monopsony mechanism kicking in between a firm's constrained and unconstrained state, our estimates imply a markdown due to becoming unconstrained of 0.148 log points or 16%. This is not far from the monopsony markdown on marginal product implied by the firm labour supply elasticity estimated above, equal to 23%.[23] Noting the correspondence between profit share and marginal markdown is not clear, which means one should be careful about interpreting these magnitudes too closely, we do interpret this as evidence linking monopsony power and profits.

6 Theoretical implications

In this section, we discuss the theoretical implications of this evidence in light of the monopsony model in Section 2 and some alternative prominent models of imperfect competition in labour markets. We begin by outlining the baseline versions of these alternative models, restricting our attention to such models which explain rent sharing, and then compare these with the baseline monopsony model and the evidence. We do not claim to adjudicate between models–it seems likely one can make ad hoc extensions to almost any model to accommodate most patterns. Instead, we show that these other models do share the prediction that profits increase more in the just-constrained region, but in these models employment still increases in this region. We interpret this as reason to think the profit results apply to a range of theoretical models with rent sharing, but that the employment result additionally supports the relevance of a core and interesting mechanism of monopsony.

6.1 Baseline Diamond-Mortensen-Pissarides (DMP) model

The Diamond-Mortensen-Pissarides (DMP) model is the workhorse model of bilateral worker-firm bargaining (Mortensen and Pissarides 1999; Cahuc et al. 2014). Appendix B.2 derives the baseline version, with simulation details. The key features are that workers and firms are matched through a function depending on the tightness of the labour market (vacancies over unemployment), and that firms face a non-zero cost for posting vacancies. DMP delivers rent sharing because when a worker and firm match, they split the surplus which depends on productivity. Baseline simplifications include that there is no on-the-job search, workers are homogeneous, and the firm discount rate and worker reservation wage are set to zero.

Figure 6 panel (a) presents a simulation of this baseline DMP model. The region on the right shows firms unconstrained by the minimum wage: as one may expect given that surplus increases with productivity, wages and profits also increase with productivity. Vacancies increase too, since firms gain more profits from each additional match. Appendix Figure A16 shows the proportional split is near-constant in this region.

In the region on the left, where productivity is below the minimum wage, firms post no vacancies as these would be purely loss-making. Like in the monopsony model, it is the middle region which is most of interest. In this region where productivity is just above the minimum

[23] The monopsony markdown is $1 - \varepsilon/(1 + \varepsilon)$, which using $\varepsilon = 3.4$ yields 23%.

wage, firms post vacancies and are constrained to pay the minimum wage, though they would prefer to pay lower wages. The higher minimum wage means a disproportionate surplus goes to wages, and so – as in the monopsony case – productivity increases are absorbed into profits. Figure A16 shows the profit *share* rises throughout this region until it reaches its steady-state share.

However, unlike in the monopsony case, vacancies and employment *increase* in this middle region.[24] The intuition is that, as vacancies become more profitable, firms increase vacancies at a greater rate, before reaching a steady-state slope in the unconstrained region.

Overall, in addition to delivering rent sharing and employment growth with productivity when unconstrained, DMP also features a rapid increase in the markdown or profit share of just-constrained firms. These features are consistent with the evidence we present in sections 4 and 5. However, contrary to our evidence, the rate of increase in hires and employment is higher in the just-constrained region compared to the unconstrained region.

[Figure 6 here]

6.2 Baseline union bargaining model

Another prominent set of models which delivers rent sharing involves unions bargaining with firms. We focus on a baseline version of the insider-outsider model (Cahuc et al. 2014), with theory and simulation details given in Appendix B.3. "Insiders", represented by the union, bargain for a wage premium above the minimum wage. A key assumption is that the firm can hire "outsiders" at the minimum wage, conditional on insiders being retained.[25] Insiders who are fired are still given the premium as severance. This means that the marginal cost of labor to the firm is just the minimum wage, since the additional insider cost is fixed.

The insider-outsider model thus delivers a weaker form of rent sharing in that only insiders, not all workers, see wages increase with productivity. Empirically, this is consistent with many studies which find higher pass-through to incumbent workers (Cho and Krueger, 2022; Garin and Silvério, 2023; Kline et al., 2019). Figure 6 panel (b) shows in the region to the right that the insider wage premium increases with productivity, as does total employment and profits. In the region to the left, where the firm retains less than the number of insiders, the union prefers to ensure members are employed and so only maintains a minimal premium until all insiders are retained. Profits and total employment increase at a steady rate across the wage-premium kink-point, however, providing a contrast to both the DMP and monopsony models.

[24] Given exogenous separations, vacancies are proportional to employment N. This can be seen from $R = sN$, where in steady state recruits R are equal to separations, and the separation rate s is constant.

[25] One can think of this as imposing a limit on the size of the union, or as local changes in employment, where over a longer time horizon the number of insiders may change.

6.3 Discussion of models and evidence

The previous discussion shows how our key variables of interest respond in other rent-sharing models. We show there is a wage-kink point in at least two prominent alternative models, the DMP and insider-outsider models. The insider-outsider model demonstrates that such a wage-kink need not necessarily generate a corresponding kink in profit share at this point; on the other hand, DMP shows that this kink in profit share is not unique to monopsony models, with similar intuition of recovering the optimal unconstrained level of profits. However, DMP implies *higher* employment growth in this region of just-constrained firms (and the insider-outsider model has no differential employment prediction), and so the differentially *lower* employment growth of the monopsony models and empirical results is not as easy to explain with other models.

In our view, it is likely that a mixture of these, and other, models applies to firms even within the same labour market. Kline (2025) even presents a model in which bargaining and monopsony-like mechanisms operate in the same firm. We therefore would not like to claim that the differentially lower employment response is evidence *against* the other models, but rather that it suggests the relevance of the core mechanism of monopsony described in Section 2. In fact, the possibility of a mixture of models provides another explanation as to why the evidence in Section 5 still shows *some* growth in hires and employment in the constrained region. That is, it could be that DMP describes some constrained firms, and so in these firms employment and hires still increases.

7 Conclusion

This paper examines how the interaction of labor market monopsony and a binding minimum wage shapes firms' responses to revenue-productivity shocks, focusing on implications for wage (rent-sharing), employment, and profit and markdown responses.

In our theoretical model, a monopsonistic firm whose preferred wage lies just below the mandated minimum will not raise its wage (i.e. no pass-through) or expand employment when it experiences a revenue-productivity increase. Instead, the firm maintains the minimum wage and absorbs the additional revenue into a higher markdown. This behavior persists until the firm's revenue-productivity grows enough that the minimum wage is no longer binding, and normal monopsonistic rent-sharing behavior – raising wages to attract more labour – resumes.

We then test this prediction in South African administrative data, finding support for the theoretical predictions both in the cross-section using a kink design and using within-firm changes in firm value added, using both internal and external (trade shock) instruments. The results complicate and enrich the emerging conclusion from the rent-sharing literature—that firms do share rents with workers—by suggesting that this depends on how the firm judges its current level of rents relative to what it expects to receive at its unconstrained equilibrium.

Our exercise tests predictions which seem somewhat peculiar to a monopsonistic understanding of labour markets, and therefore we add to the existing evidence that monopsony is a relevant feature of labour markets across a wide range of contexts. Our results also showcase distinctions and potential disconnects between three related concepts in the literature: the level of rent sharing, its elasticity, and the pass-through. We show that the correspondence between these concepts depends on how firms respond to constraints on their monopsony rents.

How relevant are the substantive findings to workers and firms? About a quarter of formal firms are constrained in our sample, and our core mechanism implies limits on the extent to which wages and employment increase with productivity at these firms. South Africa is a particularly useful setting for testing this mechanism given that the proportion of firms affected by the supply-constrained region is larger when minimum wages are relatively high, as in our case. However, South Africa is no outlier here, as is evident from Appendix Figure A15, which shows the minimum to median wage, or Kaitz index, by cross-country gross national income. Given that the Kaitz index is generally higher for lower and middle income countries than for higher income countries (about 14 percentage points higher, with a standard error of 5), the mechanism we identify is likely to be of particular relevance to developing countries, where it may weaken a developmental path predicated on firms sharing the gains of productivity growth in the form of higher wages and expanded employment.[26] The mechanism may also help explain part of the the common developing-country complaint of "stalled" development or "jobless growth" (Kannan and Raveendran 2009; Sanyal 2014).

A similar cross-national pattern as is found for the Kaitz index is also emerging in the literature which estimates firm-specific labour supply elasticities, where a higher firm-specific labour supply elasticity decreases the size of the supply-constrained region in our model. Studies do suggest this elasticity is lower in developing economies, though there are very few such estimates. To our knowledge, a near-exhaustive list of well-identified estimates from developing countries is given by Amodio and De Roux (2022) estimating 2.4 in Colombia, Dal Bó et al. (2013) estimating 2.1 in Mexico, Naidu et al. (2016) estimating 1 in the UAE, and Sharma (2023) estimating approximately 2 in India. Our estimate of around 3 contributes to this small list.

Our paper also prompts a re-evaluation of the welfare effects of minimum wage policies in monopsonistic settings. That minimum wages can increase efficiency and improve worker welfare under static monopsony is well understood. In response to the introduction of a binding minimum wage, firms in the supply-constrained region will increase wages and employment, at the expense of reduced firm monopsony rents. However, we show that in a dynamic setting, worker welfare benefits of a minimum wage in the supply-constrained region erode as firms'

[26] Correspondingly, some higher income countries may be less affected by the dynamics associated with the supply-constrained region. For example, Berger et al. (2025) consider minimum wage effects in constrained and unconstrained firms in the United States, but find that the supply-constrained region is much smaller. We replicate this finding in our simple model from Section 2: with parameters from Berger et al. (2025) (i.e. a higher firm labour supply elasticity of 10, and a relatively lower minimum wage binding for 10% of workers), the supply-constrained region is only 4% of firms. This compares to the same model with South African data (only adjusting for a lower firm labour supply elasticity of 3 and higher minimum wage binding for 20% of workers), where the supply-constrained region is closer to 20% of firms, which compares favorably with our empirical results.

revenue productivity increases. Worker welfare in this supply-constrained region does not improve as firms' revenue productivity increases, because wages and employment remain constant. Thus the static minimum wage cannot force firms to accept lower monopsony rents in perpetuity. When the opportunity arises, firms claw back these rents by keeping employment and wages fixed and simply absorb profits, until they reach the unconstrained monopsony equilibrium. Of course, workers at the supply-constrained firms are still better off than without the minimum wage, because wages and employment do not fall below the level associated with the minimum wage. But this mechanism highlights limits as to the potential welfare benefits of the minimum wage.

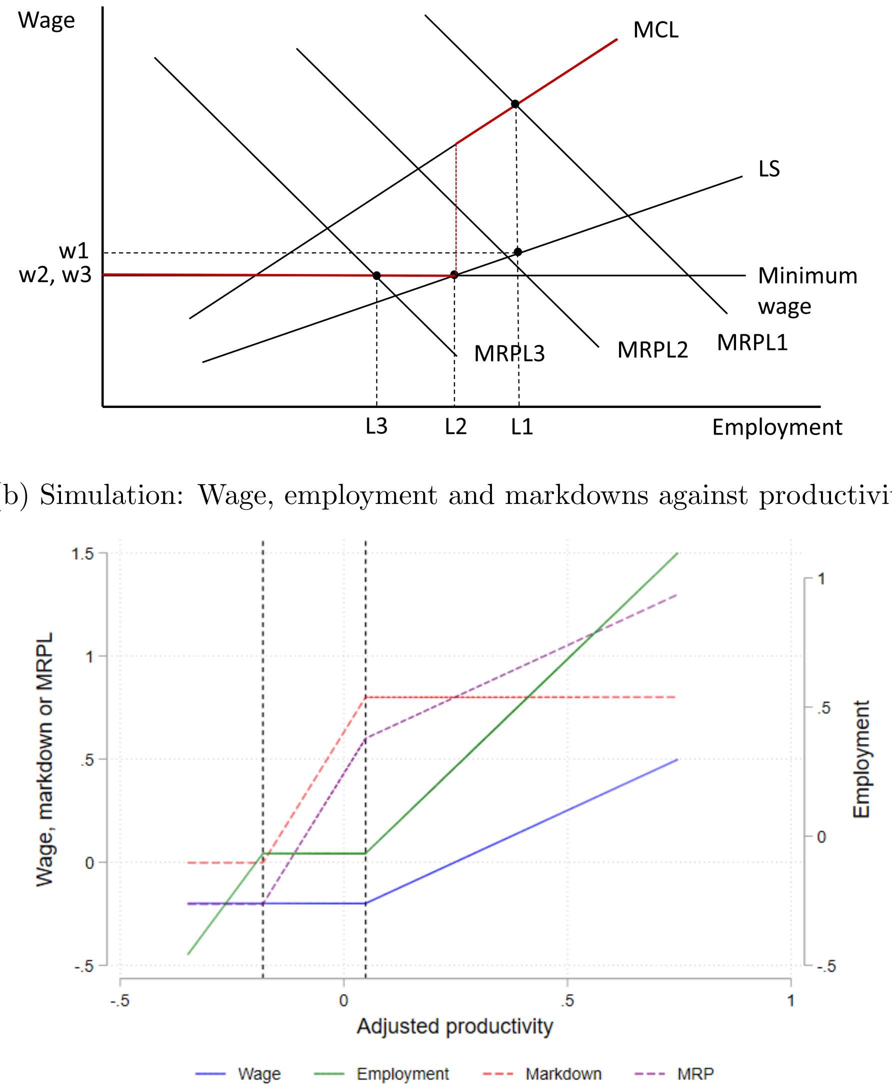
In terms of policy, ensuring that workers capture the benefits of productivity increases and growth in this context would require additional dynamic interventions, or more fundamental reforms which reduce monopsony power and empower countervailing institutions. When labour regulations aside from minimum wages or other labour supply constraints more generally bind (e.g. subsistence or efficiency wages), and firms earn profits below their desired level, there too they may choose to respond to market and productivity expansions by simply absorbing these gains as windfall profits rather than sharing benefits with workers.

Tables and figures

Figure 1: Model of monopsonistic firms of differing productivity with a minimum wage

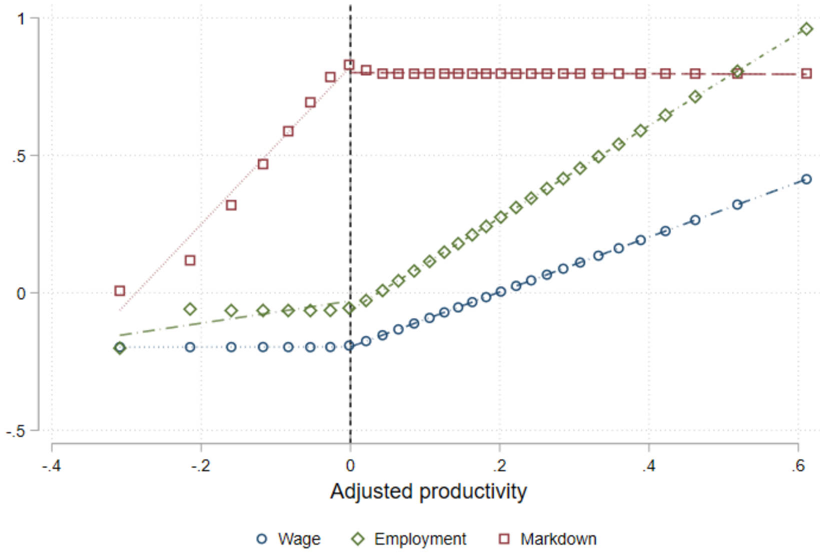
(a) Three revenue-productivity regimes in the presence of a minimum wage

(b) Simulation: Wage, employment and markdowns against productivity



Notes: Panel (a) shows the standard graphical depiction of monopsonistic firms facing minimum wages, adapted from Manning (2003, p. 343), for firms of three different productivity levels (MRPL1, MRPL2, MRPL3). The red line shows the "effective" MCL in the presence of the minimum wage, which is discontinuous at L2. Panel (b) shows a simulation of the simple model in Section 2.2 with each firm's "adjusted productivity" on the x-axis. Outcomes are the firm wage (blue), firm employment (green), markdown (red), and MRPL (purple). All outcome values are in logs. Wages and employment are normalized relative to their average values in the perfectly competitive case with no minimum wage. The dotted vertical lines indicate the boundaries of the different regions: demand-constrained firms are in the left-most region (9% firms), supply-constrained firms in the centre (19%), and unconstrained firms in the right-most region (72%). See Section 2.2 for simulation details.

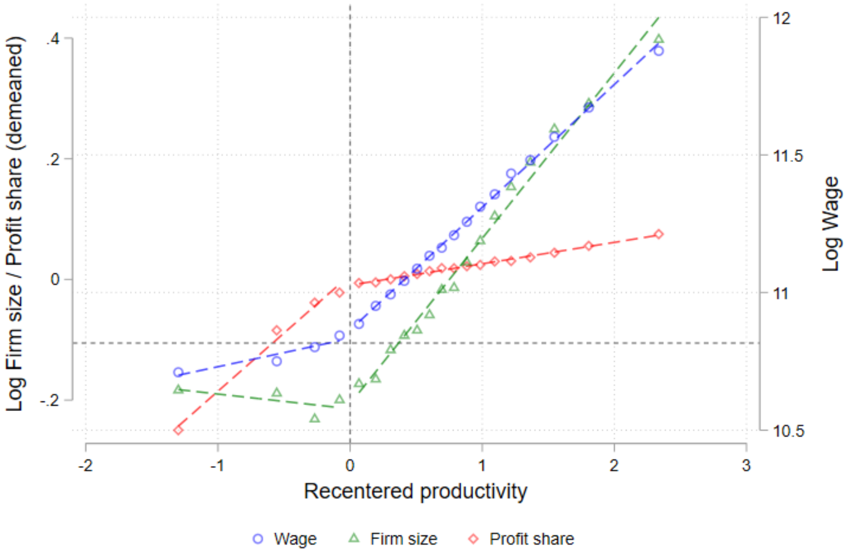
Figure 2: Pooled simulation demonstrating empirical predictions



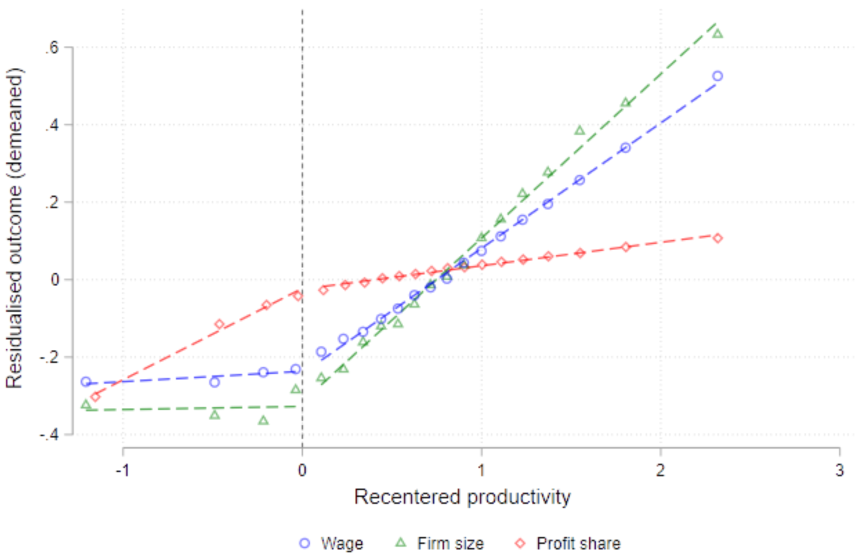
Notes: The figure shows results from pooled model simulations, where the simulation of Section 2.2 (i.e. Figure 1 panel (b)) is run separately for 40 different labour markets with randomly varying labour supply elasticities and minimum wages. Industry estimates are then pooled after re-centering their adjusted productivity v_i around the wage-kink threshold v^*. Outcome values are in logs relative to their average values in the perfectly competitive case with no minimum wage. The vertical line shows the wage-kink threshold and divides constrained firms to the left (23.5% of firms) and unconstrained firms to the right (76.5% of firms). See Section 2.3 for more details..

Figure 3: Cross-sectional kink design

(a) Raw outcomes



(b) Residualized outcomes



Notes: Panel (a) shows firm median wage (blue), employment (green), and profit share (red) (all in logs) by 20 recentered firm productivity bins (productivity estimated using the ACF method). Underlying firm productivity is recentered around the wage-kink $\hat{v}^*$, which is estimated separately for each minimum wage regime (Bargaining Council or Sectoral Determination). Recentering means recentered productivity equals zero at the value of the wage-kink (shown with the vertical dashed line). The wage-kink $\hat{v}^*$ divides minimum wage-constrained (to the left) and unconstrained firms (to the right).. The horizontal line is the average minimum wage across firms. Panel (b) shows the same graph, but with the outcomes first residualized on the average AKM worker fixed at the firm, the firm poaching ratio, and industry (2-digit) by region (district council) labour market dummies. See section 4 for details.

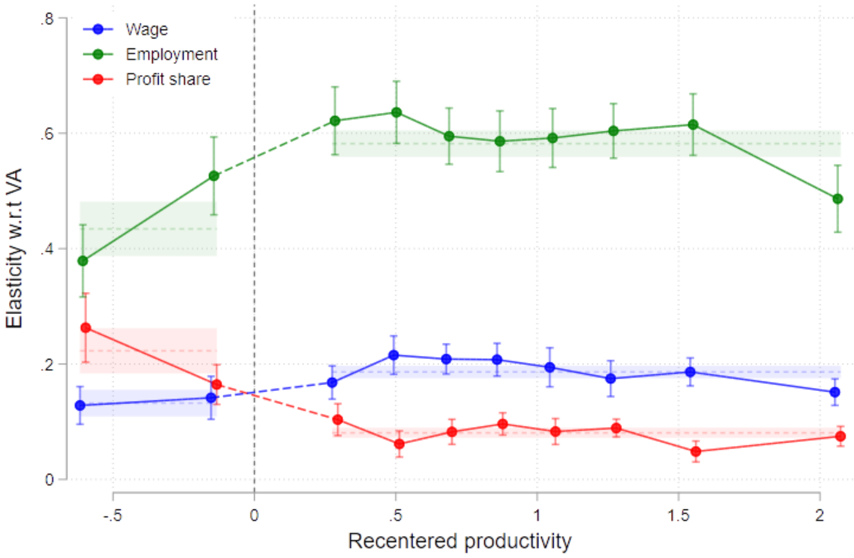
Table 1: Differential slopes from cross-sectional kink design

	No controls			**With controls**		
	Constrained	Unconstrained	Difference	Constrained	Unconstrained	Difference
Wage	0.086***	0.444***	0.358***	0.033**	0.324***	0.291***
	(0.0112)	(0.0062)	(0.0128)	(0.0102)	(0.0073)	(0.0126)
Employment	-0.027	0.273***	0.300***	0.054*	0.445***	0.391***
	(0.0228)	(0.0117)	(0.0257)	(0.0218)	(0.0100)	(0.0240)
Profit-share	0.205***	0.036***	-0.168***	0.252***	0.056***	-0.195***
	(0.0092)	(0.0030)	(0.0097)	(0.0130)	(0.0028)	(0.0133)
N	252533	1034539		189365	847963	

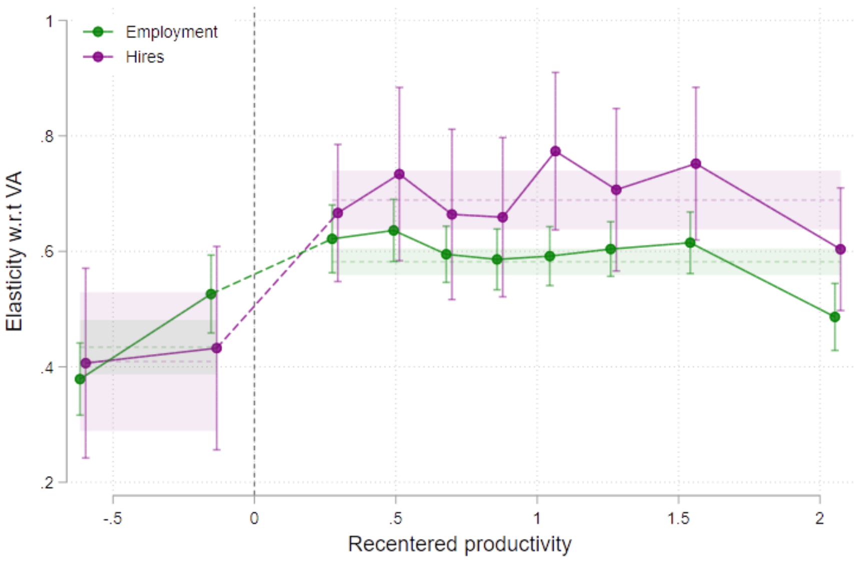
Notes: Table shows the slopes from the kink design in Figure 3 (see section 4), separately for constrained and unconstrained firms, as well as the difference in slopes between the constrained and unconstrained firms. These are the slopes of the outcome variable (wage, employment, or profit share, all in logs) on the recentered productivity term. Controls refer to the average AKM worker fixed at the firm, the firm poaching ratio, and industry (2-digit) by region (district council) labour market dummies. Standard errors are shown in parentheses and clustered at the firm and event levels; for differences these are calculated using the Delta method. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Figure 4: Internal IV main results (full elasticities), by recentered productivity bin

(a) Main outcomes



(b) Employment and hires



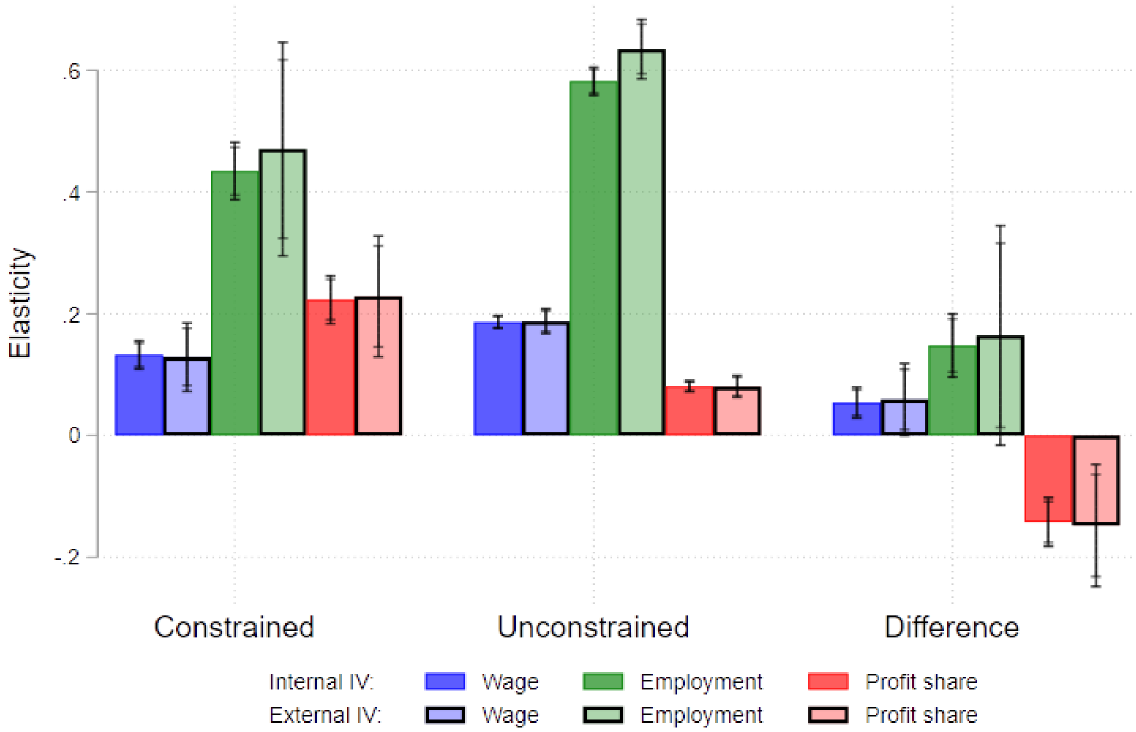
Notes: Panel (a) shows main IV results for the internal instrument specification, estimated by productivity bin. Full elasticities are shown, estimated by regressing the pre-post change in the outcome (in logs) on the pre-post change in log value-added, with the change in value-added instrumented by the binary treatment variable, as discussed in Section 5 The horizontal axis is firm productivity (estimated using the ACF method) recentered around the estimated productivity wage-kink $\hat{v}^*$ (see Section 4). Recentered productivity equals zero at the wage-kink, shown with a vertical dashed line, and this line divides minimum wage-constrained (to the left) and -unconstrained (to the right) firms. Ten approximately equally-sized productivity bins (deciles) are created. The solid lines and points show the average treatment effect across post-periods 1 and 2 by bin. 95% confidence intervals are shown with vertical bars. The horizontal dashed lines with attendant shaded regions (95% confidence intervals) show applicable post-period treatment effects estimated across the productivity bins, pooled separately below and above the wage-kink value $\hat{v}^*$. Red is for firm profit share, green firm employment, and blue the median wage of firm stayers (incumbents). Panel (b) shows the same results but with hires, shown in purple.

Table 2: Estimates from internal and external IV shocks, constrained versus unconstrained firms

	Internal IV (large VA change)			**External IV** (trade shock)		
	Constrained	Unconstrained	Difference	Constrained	Unconstrained	Difference
Panel (a)						
Rent sharing	**0.132*** **	**0.186*** **	**0.054*** **	**0.129*** **	**0.188*** **	**0.059* **
	(0.0117)	(0.0054)	(0.0129)	(0.0284)	(0.0104)	(0.0302)
Employment	**0.434*** **	**0.582*** **	**0.148*** **	**0.470*** **	**0.635*** **	**0.164* **
	(0.0238)	(0.0116)	(0.0265)	(0.0887)	(0.0247)	(0.0921)
Profit-share	**0.223*** **	**0.081*** **	**-0.142*** **	**0.228*** **	**0.080*** **	**-0.148*** **
	(0.0198)	(0.0044)	(0.0203)	(0.0502)	(0.0092)	(0.0511)
F-stat	449.9	5956.7		40.6	716.5	
N firms	6167	34120		956	12017	
Obs	13596	98234		13242	224244	
Panel (b)						
Hires	**0.409*** **	**0.689*** **	**0.280*** **	**0.271* **	**0.685*** **	**0.414** **
	(0.0609)	(0.0258)	(0.0662)	(0.1549)	(0.0594)	(0.1659)
F-stat	299.4	5013.0		32.6	583.6	
N firms	4884	29587		943	11893	
Obs	9929	77244		11367	191588	

Notes: The table shows main results for the responses of wages (rent sharing), firm employment, profit-share and firm hires (all in logs) to firm value added (VA) shocks, by constrained versus unconstrained status, as well as the difference in responses between constrained and unconstrained. Firm VA (in logs) is instrumented using two approaches: the left super-column shows estimates from the internal IV, i.e. above-median increases in firm VA between event periods -1 and 0. Only the post-period effects are reported. The right super-column shows estimates from the external IV, i.e. the shift-share trade shocks. See section 5 for sample and specification details. Note that the external shock uses a larger pooled sample to account for sample size limitations (as evidenced by the count of individual firms). Standard errors are shown in parentheses and clustered at labor market by event; for differences these are calculated using the Delta method. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

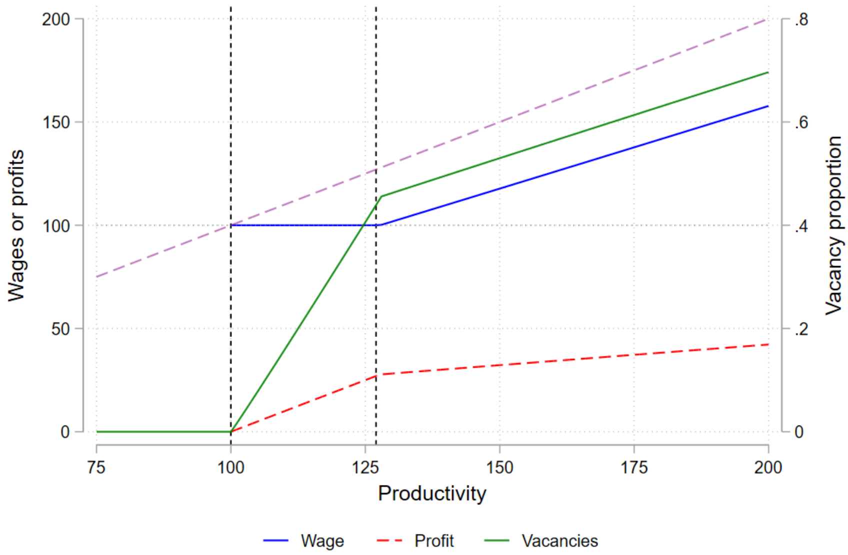
Figure 5: Comparison of estimates from internal and external IV shocks



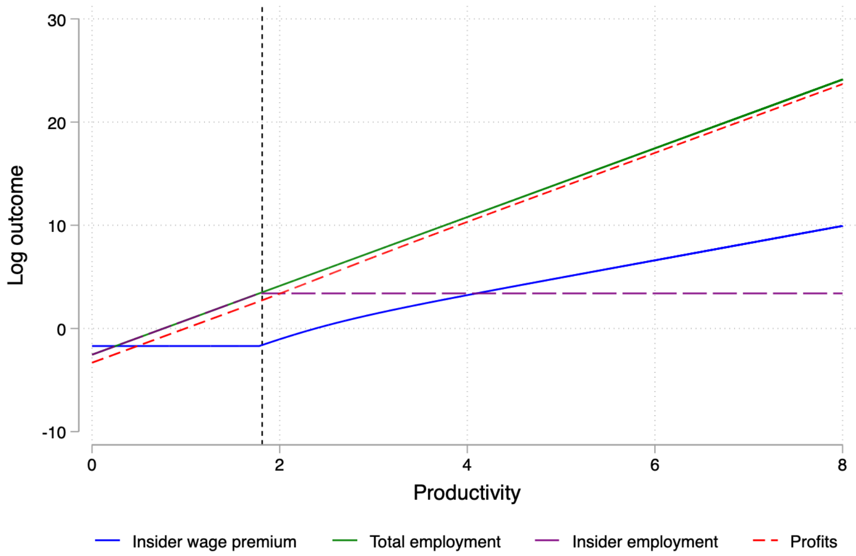
Notes: The figure plots the estimates from Table 2, where the internal IV refers to above-median increases in firm VA between event periods -1 and 0 while the external IV refers to the shift-share trade shocks. Wage responses are in blue, employment in green, and profit share in red (all are elasticities). See section 5 for sample and specification details. Vertical bars represent 90% and 95% confidence intervals, where standard errors are clustered at labor market by event and differences these are calculated using the Delta method.

Figure 6: Simulations of other models: Wages, employment and profits with productivity

(a) Diamond-Mortensen-Pissarides (DMP)



(b) Union bargaining



Notes: Panel (a) shows the DMP model (Mortensen and Pissarides 1999); see Appendix B.2 for details of the model and simulation. The minimum wage is set at 100, and the wage is missing where there are no vacancies. The purple dashed line is the 45 degree line representing productivity. The left vertical line indicates where productivity is equal to the minimum wage (i.e. vacancies are first profitable), and the right line indicates where the minimum wage is no longer binding (analogous to the v^* from the monopsony model in section 1). Vacancies are proportional to firm employment. Panel (b) shows the union bargaining model with insiders and outsiders; see Appendix B.3 for details of the model and simulation. The minimum wage is set at 1.5, with the vertical line indicating where the maximum "insider" employment is reached (closest analogous point to the v^* from the monopsony model in Section 1)

References

Abowd, J. M., Kramarz, F., and Margolis, D. N. (1999). High wage workers and high wage firms. *Econometrica*, 67(2):251–333.

Ackerberg, D. A., Caves, K., and Frazer, G. (2015). Identification properties of recent production function estimators. *Econometrica*, 83(6):2411–2451.

Amodio, F. and De Roux, N. (2022). Labor market power in developing countries: Evidence from Colombian plants. Mimeo, Preprint.

Azar, J., Huet-Vaughn, E., Marinescu, I., Taska, B., and Von Wachter, T. (2024). Minimum wage employment effects and labour market concentration. *Review of Economic Studies*, 91(4):1843–1883.

Azar, J. and Marinescu, I. (2024). Monopsony power in the labor market. In *Handbook of Labor Economics*, volume 5, pages 761–827. Elsevier.

Bagger, J. and Lentz, R. (2019). An empirical model of wage dispersion with sorting. *The Review of Economic Studies*, 86(1):153–190.

Bassier, I. (2022). Collective bargaining and spillovers in local labor markets. CEP Discussion Paper 1895, Centre for Economic Performance.

Bassier, I. (2023). Firms and inequality when unemployment is high. *Journal of Development Economics*, 161.

Bassier, I., Dube, A., and Naidu, S. (2022). Monopsony in movers: The elasticity of labor supply to firm wage policies. *Journal of Human Resources*, 57(S):S50–s86.

Bassier, I. and Manning, A. (forthcoming). Firm dynamics and the labour supply elasticity. Working paper, Centre for Economic Performance, London School of Economics.

Bassier, I. and Ranchhod, V. (2024). Can minimum wages effectively reduce poverty under low compliance? a case study from the agricultural sector in south africa. *Review of Political Economy*, 36(2):398–419.

Baum, C. F., Schaffer, M. E., and Stillman, S. (2002). IVREG2: Stata module for extended instrumental variables/2SLS and GMM estimation. Statistical Software Components, Boston College Department of Economics.

Bell, B., Bukowski, P., and Machin, S. (2024). The decline in rent sharing. *Journal of Labor Economics*, 42(3):683–716.

Berger, D., Herkenhoff, K., and Mongey, S. (2025). Minimum wages, efficiency, and welfare. *Econometrica*, 93(1):265–301.

Bhorat, H., Kanbur, R., and Mayet, N. (2012). Minimum wage violation in south africa. *International Labour Review*, 151(3):277–287.

Bhuller, M., Moene, K. O., Mogstad, M., and Vestad, O. L. (2022). Facts and fantasies about wage setting and collective bargaining. *Journal of Economic Perspectives*, 36(4):29–52.

Bischof, D. (2017). New graphic schemes for Stata: plotplain and plottig. *Stata Journal*, 17(3):748–759.

Bloesch, J. and Larsen, B. (2023). When do firms profit from wage setting power? Working paper, Available at SSRN 4614661.

Bravo, M. C. (2018). GTOOLS: Stata module to provide a fast implementation of common group commands. Statistical Software Components, Boston College Department of Economics.

Breza, E., Kaur, S., and Shamdasani, Y. (2021). Labor rationing. *American Economic Review*, 111(10):3184–3224.

Budlender, J. and Ebrahim, A. (2020). Industry classification in the South African tax microdata. SA-TIED Working Paper 134, UNU-WIDER.

Cahuc, P., Carcillo, S., and Zylberberg, A. (2014). *Labor economics*. MIT press.

Card, D., Cardoso, A. R., Heining, J., and Kline, P. (2018). Firms and labor market inequality: Evidence and some theory. *Journal of Labor Economics*, 36(S1):S13–S70.

Card, D., Cardoso, A. R., and Kline, P. (2016). Bargaining, sorting, and the gender wage gap: Quantifying the impact of firms on the relative pay of women. *The Quarterly Journal of Economics*, 131(2):633–686.

Card, D., Lee, D. S., Pei, Z., and Weber, A. (2015). Inference on causal effects in a generalized regression kink design. *Econometrica*, 83(6):2453–2483.

Card, D., Lee, D. S., Pei, Z., and Weber, A. (2017). Regression kink design: Theory and practice. Discussion paper.

Cattaneo, M. D., Crump, R. K., Farrell, M. H., and Feng, Y. (2024). On binscatter. *American Economic Review*, 114(5):1488–1514.

Cengiz, D., Dube, A., Lindner, A., and Zipperer, B. (2019). The effect of minimum wages on low-wage jobs. *The Quarterly Journal of Economics*, 134(3):1405–1454.

Cho, D. and Krueger, A. B. (2022). Rent sharing within firms. *Journal of Labor Economics*, 40(S1):S17–S38.

Correia, S. (2014). REGHDFE: Stata module to perform linear or instrumental-variable regression absorbing any number of high-dimensional fixed effects. Statistical Software Components, Boston College Department of Economics.

Correia, S. (2018). IVREGHDFE: Stata module for extended instrumental variable regressions with multiple levels of fixed effects. Statistical Software Components, Boston College Department of Economics.

Dal Bó, E., Finan, F., and Rossi, M. A. (2013). Strengthening state capabilities: The role of financial incentives in the call to public service. *The Quarterly Journal of Economics*, 128(3):1169–1218.

Datta, N. (2023). The measure of monopsony: the labour supply elasticity to the firm and its constituents. CEP Discussion Paper 1930, Centre for Economic Performance.

de Frahan, I., Lamadon, T., Meling, T., and Mogstad, M. (2022). Why do larger firms have lower labor shares? NBER Labor Studies Program Meeting, Spring.

De Loecker, J. and Syverson, C. (2021). An industrial organization perspective on productivity. In *Handbook of industrial organization*, volume 4, pages 141–223. Elsevier.

Dickens, R., Machin, S., and Manning, A. (1999). The effects of minimum wages on employment: Theory and evidence from Britain. *Journal of Labor Economics*, 17(1):1–22.

Dube, A., Giuliano, L., and Leonard, J. (2019). Fairness and frictions: The impact of unequal raises on quit behavior. *American Economic Review*, 109(2):620–63.

Dustmann, C., Lindner, A., Schönberg, U., Umkehrer, M., and Vom Berge, P. (2022). Reallocation effects of the minimum wage. *The Quarterly Journal of Economics*, 137(1):267–328.

Engbom, N. and Moser, C. (2022). Earnings inequality and the minimum wage: Evidence from brazil. *American Economic Review*, 112(12):3803–3847.

Gandhi, A., Navarro, S., and Rivers, D. A. (2020). On the identification of gross output production functions. *Journal of Political Economy*, 128(8):2973–3016.

Ganong, P. and Jäger, S. (2018). A permutation test for the regression kink design. *Journal of the American Statistical Association*, 113(522):494–504.

Garin, A. and Silvério, F. (2023). How responsive are wages to firm-specific changes in labor demand? evidence from idiosyncratic export demand shocks. *Review of Economic Studies*, page https://doi.org/10.1093/restud/rdad069.

Gouin-Bonenfant, E. (2018). Productivity dispersion, between-firm competition, and the labor share. Unpublished paper (JMP).

Jäger, S., Noy, S., and Schoefer, B. (2022). The german model of industrial relations: balancing flexibility and collective action. *Journal of Economic Perspectives*, 36(4):53–80.

Kannan, K. and Raveendran, G. (2009). Growth sans employment: A quarter century of jobless growth in india's organised manufacturing. *Economic and Political weekly*, pages 80–91.

Kerr, A. (2018). Job flows, worker flows and churning in south africa. *South African Journal of Economics*, 86:141–166.

Kline, P., Petkova, N., Williams, H., and Zidar, O. (2019). Who profits from patents? rent-sharing at innovative firms. *The Quarterly Journal of Economics*, 134(3):1343–1404.

Kline, P. M. (2025). Labor market monopsony: Fundamentals and frontiers. Working paper, National Bureau of Economic Research.

Kroft, K., Luo, Y., Mogstad, M., and Setzler, B. (2023). Imperfect competition and rents in labor and product markets: The case of the construction industry. Update of NBER Working Paper w27325, Mimeo.

Lamadon, T., Mogstad, M., and Setzler, B. (2022). Imperfect competition, compensating differentials, and rent sharing in the us labor market. *American Economic Review*, 112(1):169–212.

Levinsohn, J. and Petrin, A. (2003). Estimating production functions using inputs to control for unobservables. *The Review of Economic Studies*, 70(2):317–341.

Lewis, W. A. (1954). Economic development with unlimited supplies of labour. *The manchester school*, 22(2):139–191.

Lunt, M. (n.d.). Piecewise linear ('hockey-stick') regression. https://personalpages.manchester.ac.uk/staff/mark.lunt/loghockey.html, Last accessed on 2023-10-09.

Manning, A. (1993). Wage bargaining and the phillips curve: the identification and specification of aggregate wage equations. *The Economic Journal*, 103(416):98–118.

Manning, A. (2003). *Monopsony in motion: Imperfect competition in labor markets*. Princeton University Press, Princeton.

Manning, A. (2021). Monopsony in labor markets: a review. *ILR Review*, 74(1):3–26.

Mortensen, D. T. and Pissarides, C. A. (1999). New developments in models of search in the labor market. *Handbook of labor economics*, 3:2567–2627.

Muralidharan, K., Niehaus, P., and Sukhtankar, S. (2023). General equilibrium effects of (improving) public employment programs: Experimental evidence from india. *Econometrica*, 91(4):1261–1295.

Naidu, S., Nyarko, Y., and Wang, S.-Y. (2016). Monopsony power in migrant labor markets: evidence from the united arab emirates. *Journal of Political Economy*, 124(6):1735–1792.

Olley, G. and Pakes, A. (1996). The dynamics of productivity in the telecommunications equipment industry. *Econometrica*, 64(6):1263–1297.

Risch, M. (2024). Does taxing business owners affect employees? evidence from a change in the top marginal tax rate. *The Quarterly Journal of Economics*, 139(1):637–692.

Rovigatti, G. and Mollisi, V. (2016). PRODEST: Stata module for production function estimation based on the control function approach. Statistical Software Components, Boston College Department of Economics.

Saez, E., Schoefer, B., and Seim, D. (2019). Payroll taxes, firm behavior, and rent sharing: Evidence from a young workers' tax cut in sweden. *American Economic Review*, 109(5):1717–63.

Sanyal, K. (2014). *Rethinking capitalist development: Primitive accumulation, governmentality and post-colonial capitalism*. Routledge India.

Sharma, G. (2023). Monopsony and gender. Technical report, MIT working paper.

Sokolova, A. and Sorensen, T. (2021). Monopsony in labor markets: A meta-analysis. *ILR Review*, 74(1):27–55.

Stepner, M. (2013). BINSCATTER: Stata module to generate binned scatterplots. Statistical Software Components, Boston College Department of Economics.

Verhoogen, E. (2023). Firm-level upgrading in developing countries. *Journal of Economic Literature*, 61(4):1410–1464.

Yeh, C., Macaluso, C., and Hershbein, B. (2022). Monopsony in the us labor market. *American Economic Review*, 112(7):2099–2138.

Appendix A: Additional Tables and Figures

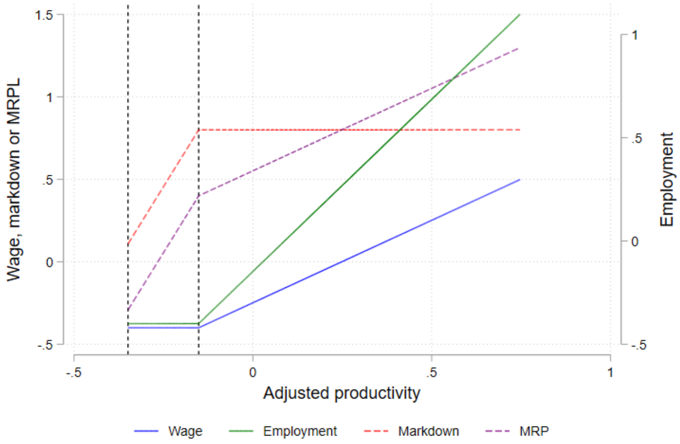
Table A1: Bargaining Council (BC) and Sectoral Determination (SD) Minimum Wages

Name	**Type**	**Minimum wage (monthly ZAR)**
Building (Bloemfontein)	BC	4174
Building (Boland)	BC	3575
Building (Cape)	BC	4810
Building (EC)	BC	6237
Building (Kimberly)	BC	6237
Chemical	BC	6255
Civil engineering	BC	6237
Clothing manufacturing	BC	4476
Electrical	BC	3995
Fishing	BC	3553
Food and restaurant	BC	3087
Furniture (KZN)	BC	2526
Furniture (WC)	BC	2713
Furniture (national)	BC	2714
Hairdressing	BC	2796
Laundry (Cape)	BC	3735
Leather	BC	4963
Meat trade	BC	3281
MEIBC	BC	7550
Motor industry	BC	3812
Restaurant catering	BC	3420
Road Freight and Logistics	BC	5066
Road passenger	BC	6071
Textile	BC	5546
Transnet	BC	7702
Tyre	BC	11402
Wood and paper	BC	5799
Contract cleaning	SD	3126
Private security	SD	3192
Farm worker	SD	2998
Forestry	SD	2998
Hospitality	SD	3169
Wholesale and retail	SD	3184

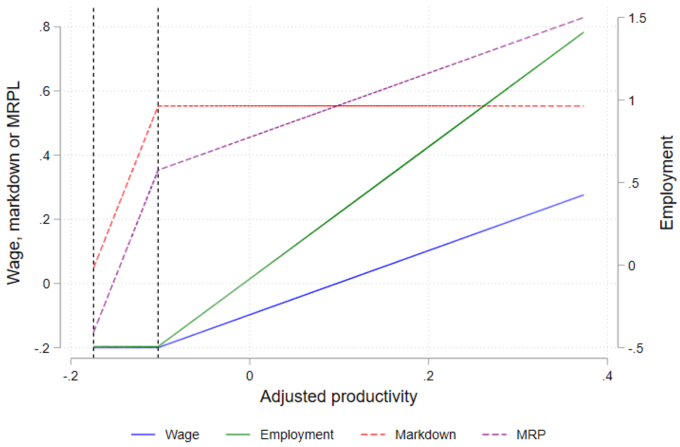
Notes: Table shows monthly minimum wages associated with each bargaining council (BC) and sectoral determination (SD) in the 2018 taxyear, in nominal Rands. The names are abbreviated. MEIBC refers to the Metal and Engineering Industries Bargaining Council.

Figure A1: Additional simulations (monopsony model)

(a) Lower minimum wage

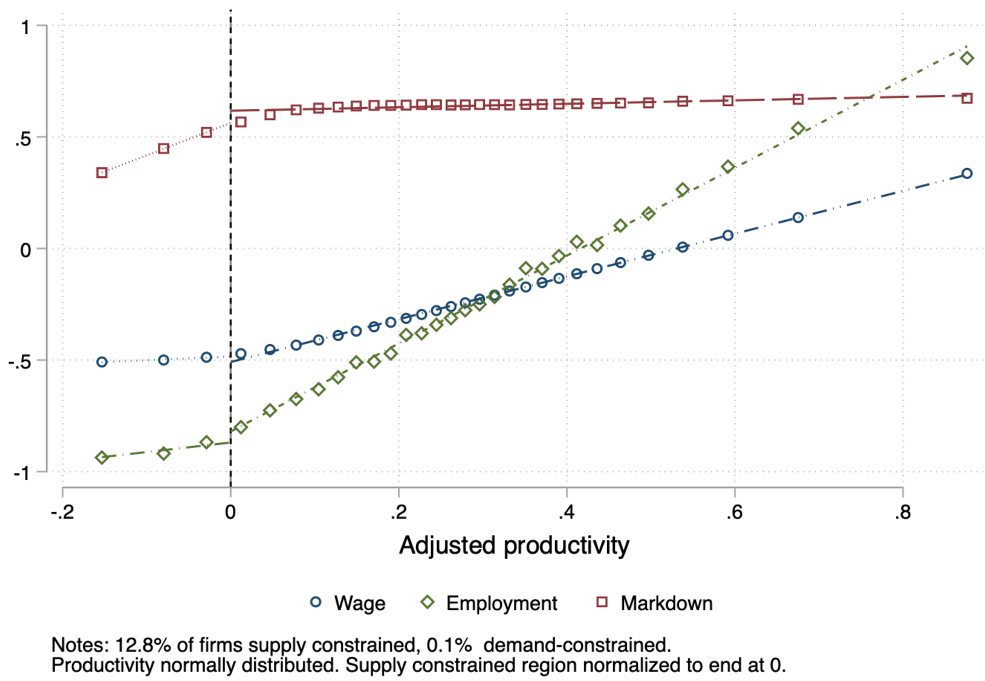


(b) High minimum wage, low labour supply elasticity



Notes: Simulations are as in Figure 1 panel (b), except the minimum wage is lower (-0.5 log points) in Panel (a) and the firm-facing labour supply elasticity is set to 4 in Panel (b). The (supply plus demand) constrained regions are smaller than the baseline in both cases, covering 8.9% and 3.2% of firms respectively.

Figure A2: Pooled monopsony model with amenities



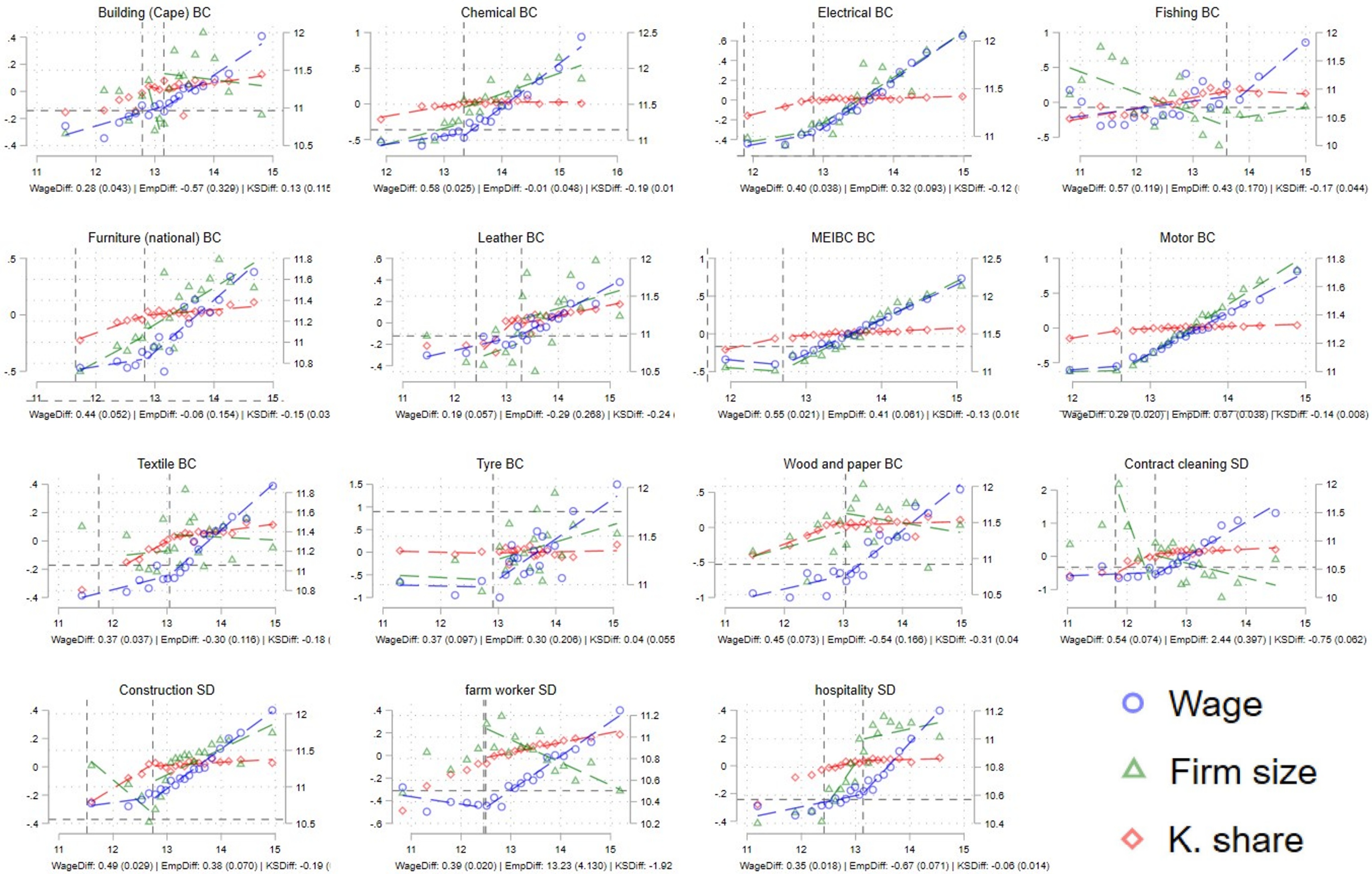
Notes: The plot shows the pooled monopsony simulation as in Figure 2, with the extension that it allows for firm-specific amenities. Amenities are set to positively covary with wages.

Table A2: Kink design implied elasticities

	No controls			**With controls**		
	Constrained	Unconstrained	Difference	Constrained	Unconstrained	Difference
Wage	0.207***	0.438***	0.231***	0.061**	0.291***	0.231***
	(0.0242)	(0.0061)	(0.0250)	(0.0179)	(0.0080)	(0.0196)
Employment	-0.064	0.269***	0.333***	0.098*	0.399***	0.301***
	(0.0502)	(0.0093)	(0.0511)	(0.0365)	(0.0060)	(0.0370)
Profit-share	0.489***	0.036***	-0.454***	0.459***	0.050***	-0.409***
	(0.0254)	(0.0025)	(0.0255)	(0.0240)	(0.0023)	(0.0241)
F-stat	231.4	15316.8		324.1	8593.7	
N	252533	1034539		189365	847963	

Notes: Table shows implied elasticities from cross-sectional kink design (Section 4) for constrained and unconstrained firms, and their differences, with and without controls. The wage (rent-sharing), employment and profit share elasticities are with respect to value-added. Elasticities are estimated by running a cross-sectional regression of the (log) outcome on (log) value-added, where value-added is instrumented with the recentered productivity value. Controls refer to the average AKM worker fixed at the firm, the firm poaching ratio, and industry (2-digit) by region (district council) labour market dummies. Standard errors are shown in parentheses and clustered at the firm and event level; for differences these are calculated using the Delta method.

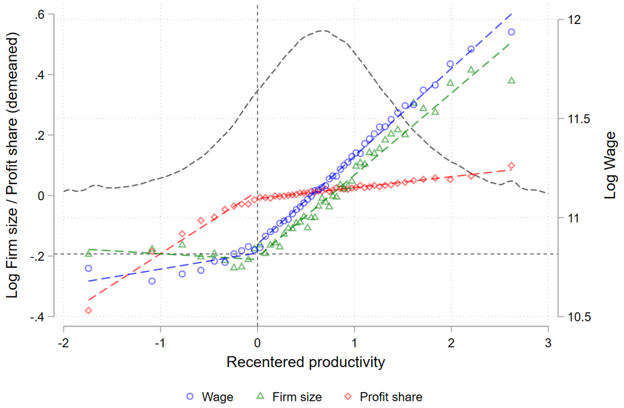
Figure A3: Cross-sectional kink design case-studies



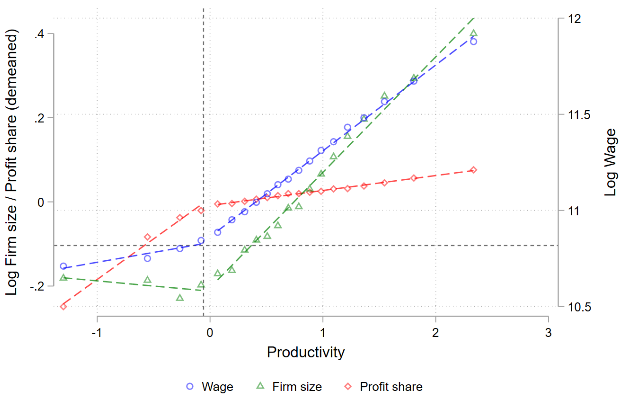
Notes: Figure shows firm median wage, employment, and profit share by 20 firm productivity bins (productivity estimated using the ACF method of Section 4; ventiles), for selected BCs and SDs. The algorithm outlined in Section 4 is used to fit underlying firm median wages as a piece-wise continuous linear function of productivity. Analogous linear fits of employment and profit share are then plotted on either side of the identified wage kink. The right-most vertical line is the estimated value of the wage-kink $\hat{v}^*$. If the there is a second left-most line, this identifies the estimated demand-constrained value $\hat{v}_1^*$. The horizontal line is the BC/SD minimum wage.

Figure A4: Robustness on cross-sectional kink design specification

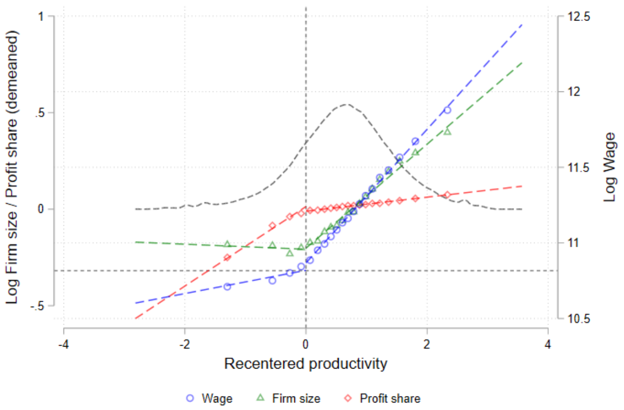
(a) 50 quantiles



(b) Kink re-estimated on pooled sample



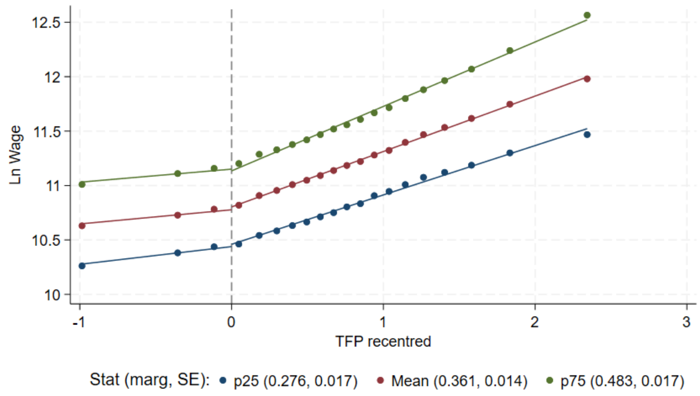
(c) Linear fits estimated on underlying data, full domain



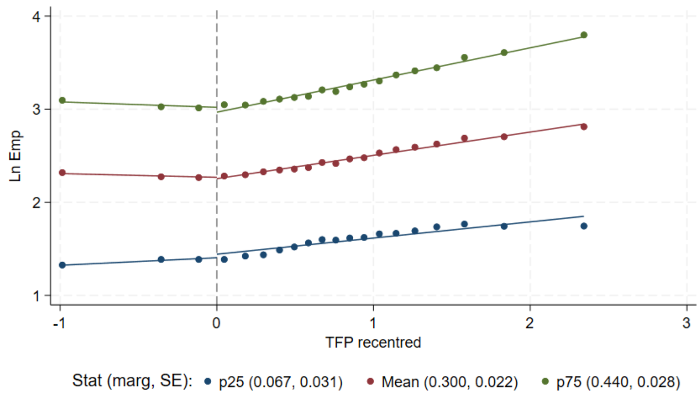
Notes: Figure shows various robustness versions of Figure 3 Panel (a). Panel (a) shows the figure with 50 quantiles rather than 20. Panel (b) shows the figure when the wage-kink algorithm is re-run on the pooled sample (the dashed vertical line shows where the wage-kink is detected). Panel (c) shows the figure when the linear fit is estimated on the underlying data rather than quantiles, and over the full domain of the data. For panels (a) and (c), the firm distribution density is overlaid.

Figure A5: Cross-sectional kink design: percentiles of outcome variable

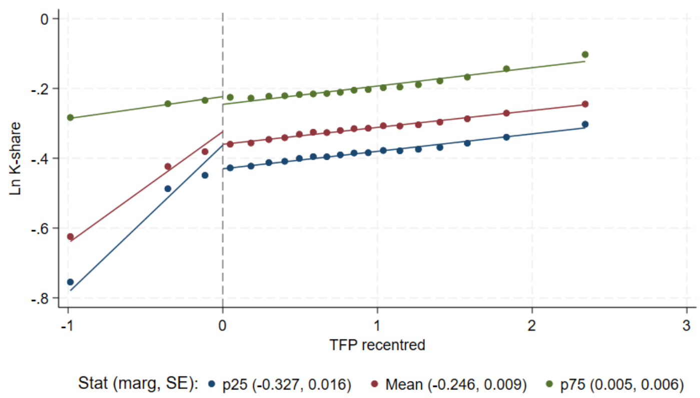
(a) Wages



(b) Employment



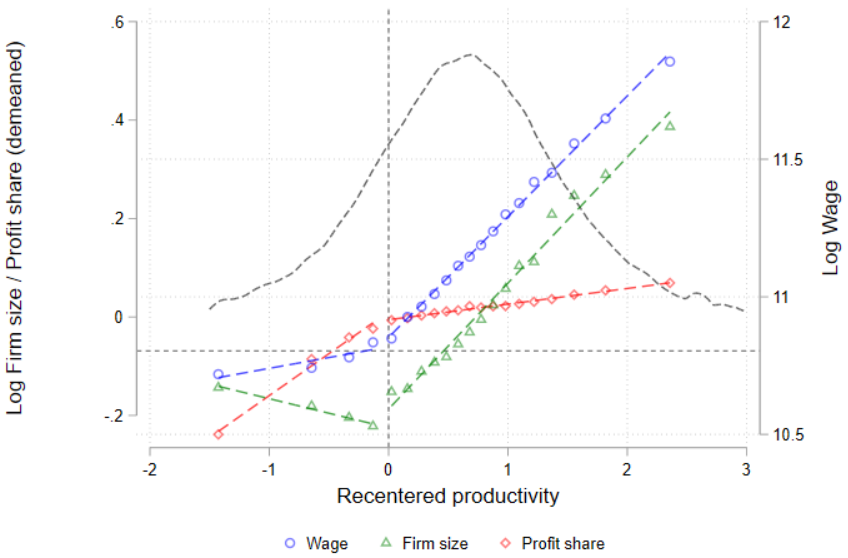
(c) Profit share



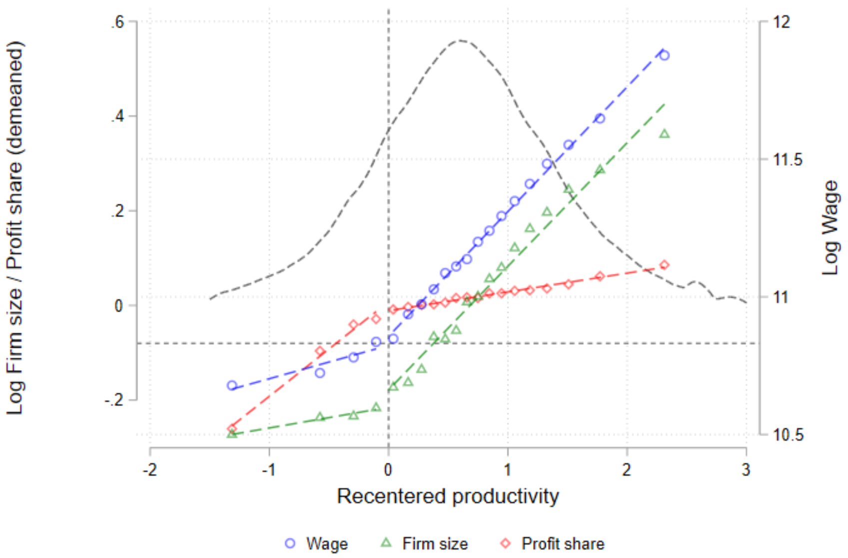
Notes: Figure shows the cross-sectional distribution of firm wages, employment and profit-share (all in logs) against recentered productivity, as in Figure 3, but for different percentiles of the outcome variable. The red markers show the mean value of the outcome for each productivity bin (the same as Figure 3). The green markers show the 75th percentile of the outcome for each productivity bin. The blue markers show the 25th percentile of the outcome for each productivity bin.

Figure A6: Cross-sectional kink design: One event per firm

(a) Only first event



(b) Only last event



Notes: Figure shows various robustness versions of Figure 3 Panel (a), where we do not pool across multiple events for any firms. Panel (a) shows the figure when only the earliest event is used for each firm. Panel (b) shows the figure when only the last event is used for each firm.

Table A3: Characteristics of constrained and unconstrained firms (means)

	All firms		**Narrow band**	
	Constrained	**Unconstrained**	**Constrained**	**Unconstrained**
Median annual wage (log)	10.78 (0.0014)	11.28 (0.0007)	10.84 (0.0041)	10.85 (0.0042)
Full-time equiv. employment	41.06 (0.8659)	43.51 (0.5732)	29.17 (0.7324)	26.85 (0.7697)
Profit share	0.67 (0.0004)	0.73 (0.0002)	0.71 (0.0010)	0.71 (0.0010)
Value-added per worker (log)	12.10 (0.0021)	12.95 (0.0009)	12.32 (0.0052)	12.35 (0.0051)
Firm-specific productivity	12.48 (0.0011)	13.62 (0.0006)	12.86 (0.0022)	12.90 (0.0021)
AKM worker effect	-0.27 (0.0010)	0.00 (0.0005)	-0.22 (0.0029)	-0.23 (0.0030)
AKM firm effect	-0.26 (0.0010)	-0.04 (0.0005)	-0.25 (0.0030)	-0.25 (0.0030)
Monthly minimum wage	4127.00 (3.2356)	4501.51 (1.8803)	4250.98 (10.6827)	4226.41 (10.7308)
Metro	0.59 (0.0010)	0.68 (0.0005)	0.59 (0.0030)	0.60 (0.0030)
Sector proportions				
Primary sector	0.09 (0.0005)	0.06 (0.0002)	0.06 (0.0015)	0.06 (0.0014)
Manufacturing	0.23 (0.0008)	0.29 (0.0004)	0.26 (0.0027)	0.25 (0.0026)
Construction	0.14 (0.0007)	0.14 (0.0003)	0.16 (0.0022)	0.16 (0.0023)
Wholesale & Retail	0.42 (0.0010)	0.37 (0.0005)	0.42 (0.0030)	0.43 (0.0030)
Infrastructure services	0.02 (0.0003)	0.04 (0.0002)	0.02 (0.0009)	0.02 (0.0009)
Bus. & Pers. services	0.10 (0.0006)	0.09 (0.0003)	0.09 (0.0017)	0.08 (0.0017)
Province proportions				
Western Cape	0.24 (0.0008)	0.21 (0.0004)	0.24 (0.0026)	0.23 (0.0026)
Eastern Cape	0.07 (0.0005)	0.06 (0.0002)	0.07 (0.0016)	0.07 (0.0016)
Northern Cape	0.02 (0.0003)	0.02 (0.0001)	0.02 (0.0008)	0.02 (0.0008)
Free State	0.04 (0.0004)	0.04 (0.0002)	0.05 (0.0013)	0.04 (0.0013)
KwaZulu-Natal	0.16 (0.0007)	0.15 (0.0003)	0.17 (0.0023)	0.17 (0.0023)
North West	0.03 (0.0003)	0.02 (0.0001)	0.03 (0.0011)	0.03 (0.0010)
Gauteng	0.33 (0.0009)	0.43 (0.0005)	0.33 (0.0029)	0.35 (0.0029)
Mpumalanga	0.06 (0.0005)	0.05 (0.0002)	0.06 (0.0015)	0.06 (0.0015)
Limpopo	0.03 (0.0003)	0.02 (0.0001)	0.03 (0.0010)	0.03 (0.0010)

Notes: Table shows descriptive statistics of the pooled cross-sectional sample, by firms' constrained status. "All firms" refers to all firms in the cross-section. "Narrow band" refers to a subset of firms which are close to the wage-kink (constrained/unconstrained boundary); specifically the top 10% of the constrained-firm recentered productivity distribution and the same number of lowest-recentered productivity unconstrained firms. AKM effects refer to wage decomposition fixed effects estimated as per Abowd et al. (1999). Firm-specific productivity is estimated using the ACF method; see Section 4. For the sectors: "Primary sector" refers to agriculture and mining; "Infrastructure services" refers to electricity, gas, and water supply as well as transport, storage, and communication; while Business and Personal Services refers to financial intermediation, insurance, real Estate, and business services as well as community, social, and personal services. "Metro" is a dummy referring to a firm being located in one of South Africa's eight metropolitan municipalities. The monthly minimum wage refers to the BC/SD minimum in 2018 Rands.

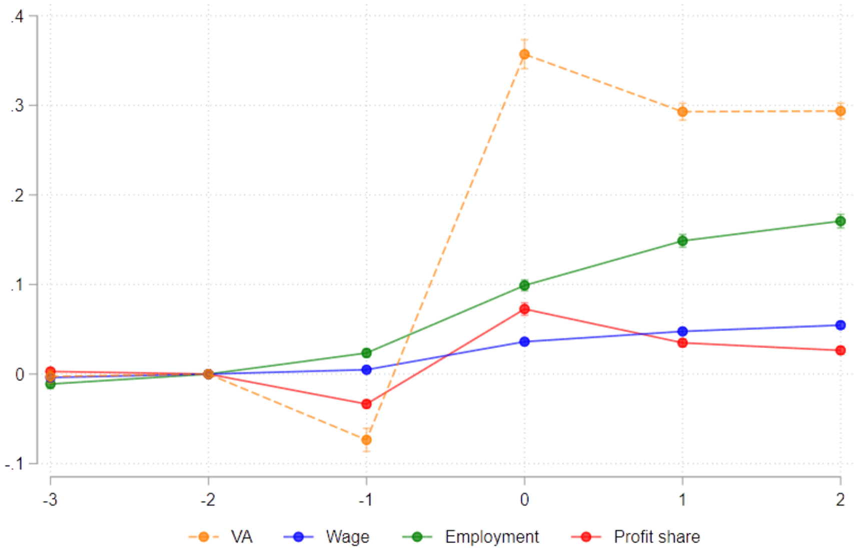
Table A4: Internal IV reduced form results (semi-elasticities)

	Internal IV (large VA change)		
	Constrained	Unconstrained	Difference
<u>**Panel (a)**</u>			
Wage	**0.044*** **	**0.053*** **	**0.009** **
	(0.0038)	(0.0016)	(0.0042)
Employment	**0.148*** **	**0.166*** **	**0.018** **
	(0.0079)	(0.0039)	(0.0089)
Profit-share	**0.076*** **	**0.023*** **	**-0.053*** **
	(0.0099)	(0.0014)	(0.0100)
N firms	6167	34120	
Obs	81576	589404	
<u>**Panel (b)**</u>			
Hires	**0.134*** **	**0.198*** **	**0.064*** **
	(0.0194)	(0.0073)	(0.0208)
N firms	5986	33577	
Obs	63071	483531	

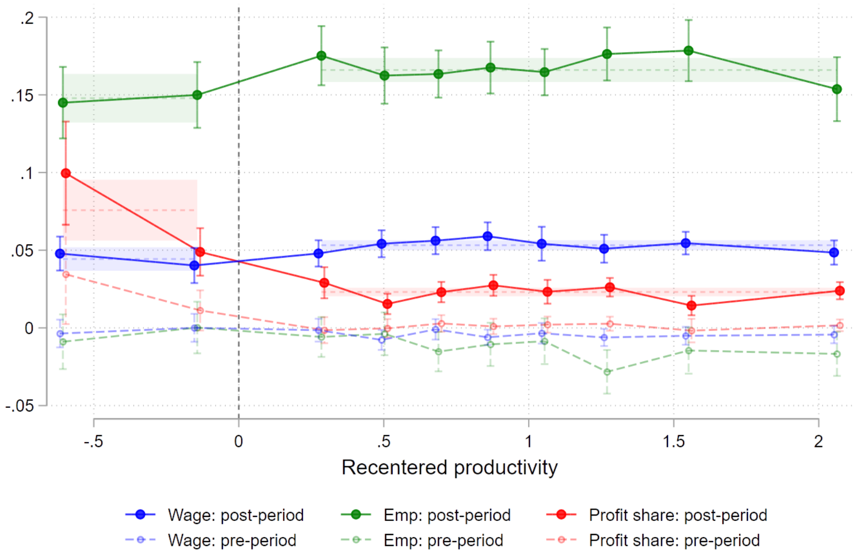
Notes: The table shows the reduced form results for the internal instrument event study, where treatment is above-median increase in firm value-added between periods -1 and 0. Panel (a) shows the results for the main balanced sample with effects on wages (of stayers), employment and profit share (all in logs). Panel (b) shows effects on hires. Estimates are normalised relative to period -2; the effects are average treatment effects across post-periods 1 and 2. See Section 5 for sample restrictions and specification. Estimates are shown separately for constrained and unconstrained firms, as well as the difference between estimates for constrained versus unconstrained. Standard errors are shown in parentheses and clustered at labor market by event; for differences these are calculated using the Delta method. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Figure A7: Internal IV reduced form results (semi-elasticities)

(a) Aggregate response



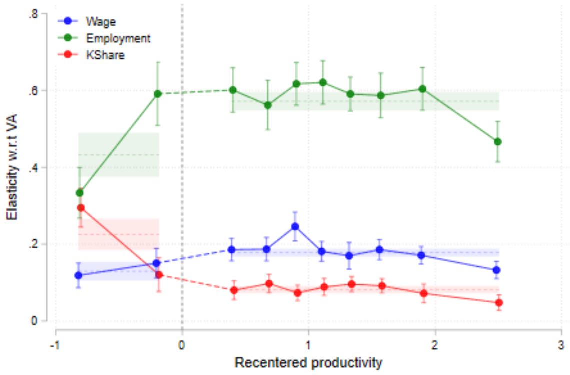
(b) Heterogeneous results



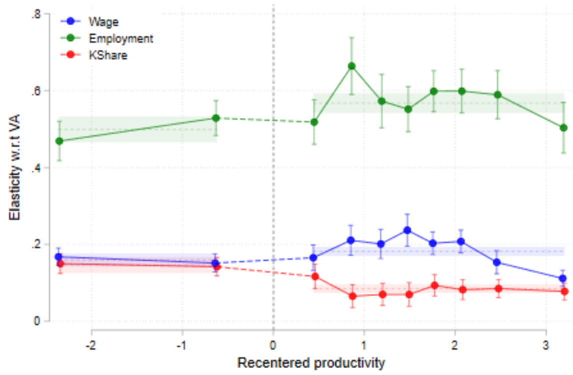
Notes: The figure shows the reduced form results for the internal instrument event study, where treatment is above-median increase in firm value-added between periods -1 and 0. Panel (a) shows the aggregate stacked event study for the full estimation sample. Estimates are normalised relative to period -2. Orange line shows response of log value-added, green log firm employment, blue the log of median wage of firm stayers (incumbents), and red log of profit share of value added. See section 5 for sample restrictions and specification. 95% confidence intervals are shown with vertical bars. Panel (b) shows estimates from event studies as in Panel (a), but estimated by productivity bin. The horizontal axis is firm productivity (estimated using the ACF method) recentered around the estimated productivity wage-kink $\hat{v}^*$ (see Section 4). Recentered productivity equals zero at the wage-kink, shown with a vertical dashed line, and this line divides minimum wage-constrained (to the left) and -unconstrained (to the right) firms. Ten approximately equally-sized productivity bins (deciles) are created. The solid lines and points show the average treatment effect across post-periods 1 and 2. The dashed lines and hollow points show effects estimated for pre-period -3. 95% confidence intervals are shown with vertical bars. The horizontal dashed lines with attendant shaded regions (95% confidence intervals) show applicable post-period treatment effects estimated across the productivity bins, pooled separately below and above the wage-kink value.

Figure A8: Internal IV results: Cobb-Douglas production function variations

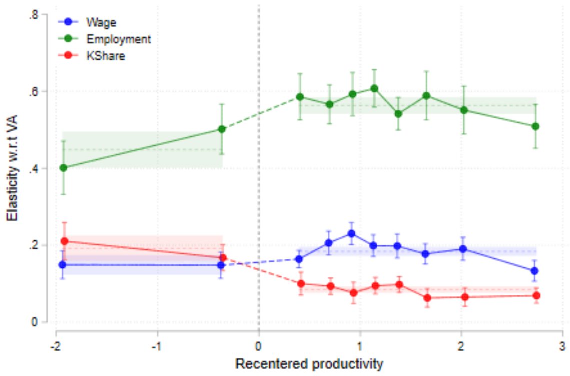
(a) Olley-Pakes



(b) Levinsohn-Petrin



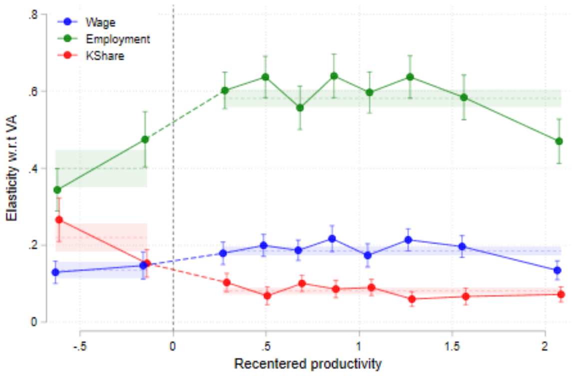
(c) ACF



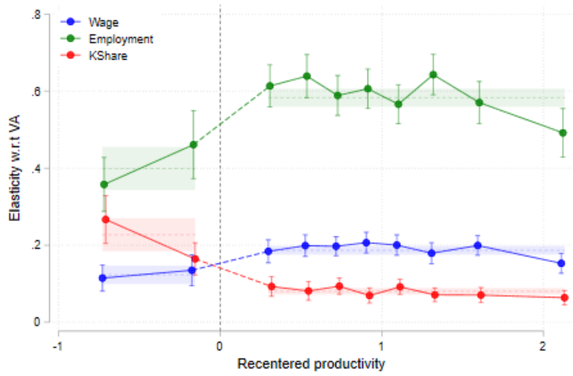
Notes: Figure shows robustness versions of main results in Figure 4. Each panel shows results when the production function underlying the recentered productivity term is estimated using the specified method for Cobb-Douglas production functions. The specification is otherwise the same as in Figure 4.

Figure A9: Internal IV results: Alternative production function estimation routines

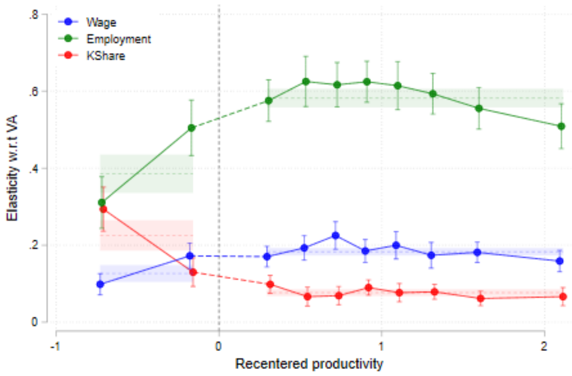
(a) Estimated by 1-digit industry



(b) Estimated by 2-digit industry



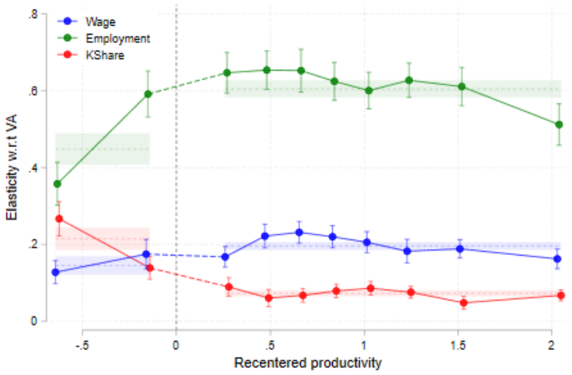
(c) ACF correction of Olley-Pakes



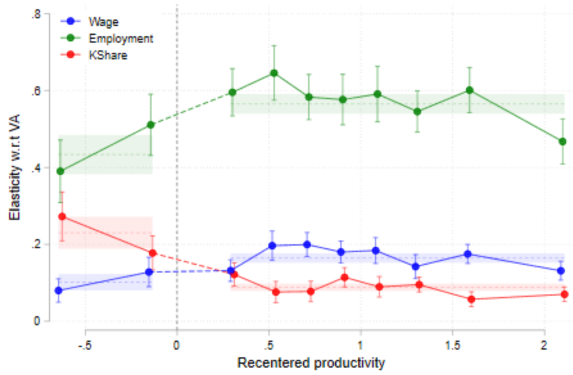
Notes: Figure shows robustness versions of main results in Figure 4. Panels (a) and (b) show results when production functions are estimated separately by 1- and 2-digit industries respectively. Panel (c) shows results when the ACF correction is applied to Olley-Pakes rather than Levinsohn-Petrin production function estimation. The specification is otherwise the same as in Figure 4 (including the specification of translog production functions).

Figure A10: Internal IV results: various “stayer” sample restrictions

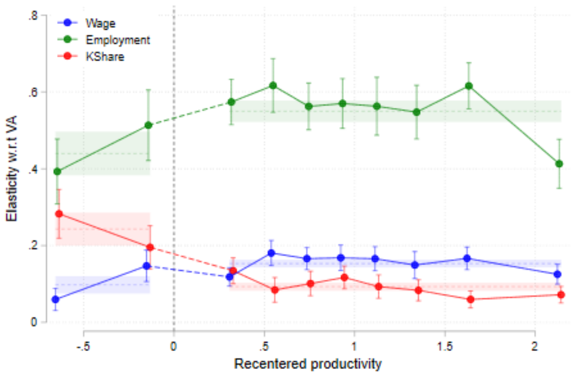
(a) ≥ 1 stayers



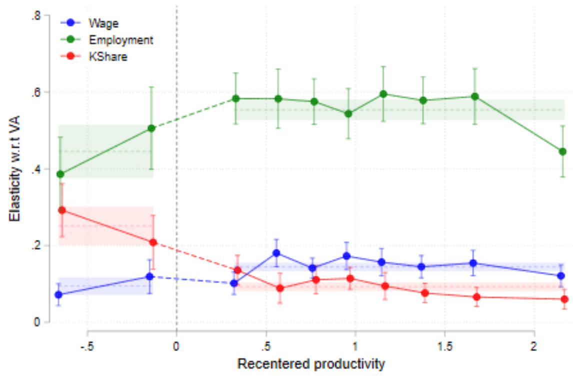
(b) ≥ 3 stayers



(c) ≥ 4 stayers



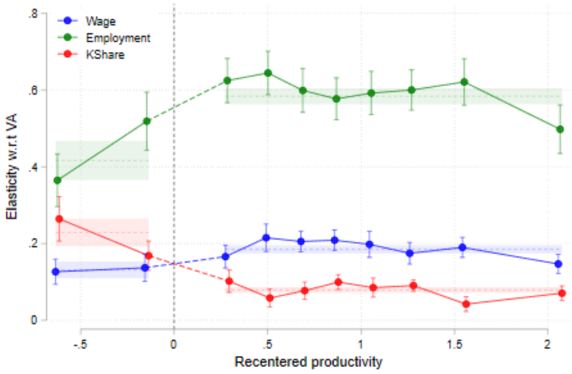
(d) ≥ 5 stayers



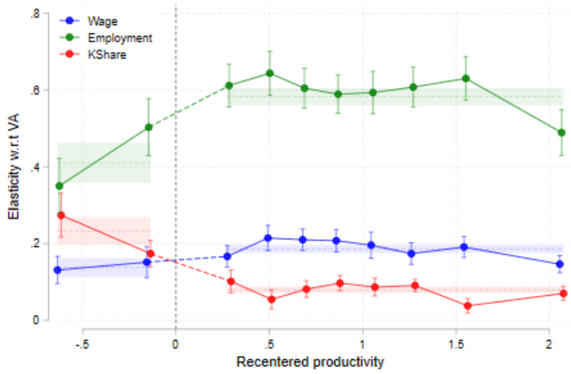
Notes: Figure shows robustness versions of main results in Figure 4. Each panel shows results when the sample is restricted to have at least 1, 3, 4 or 5 stayers over the event-study period. The specification is otherwise the same as in Figure 4.

Figure A11: Internal IV results: various labour market definitions

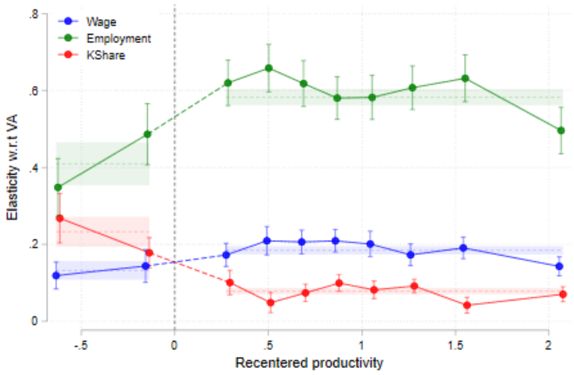
(a) Province × 2-digit industry



(b) District Council × 1-digit industry



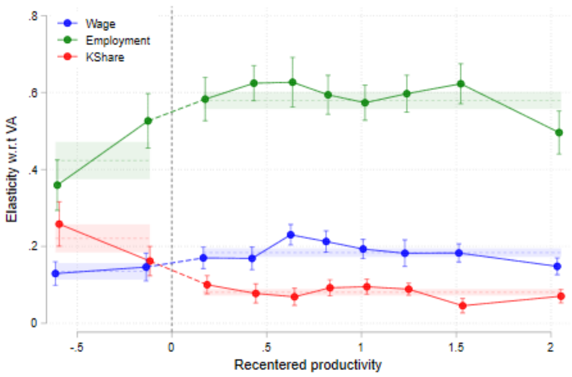
(c) District Council × 2-digit industry



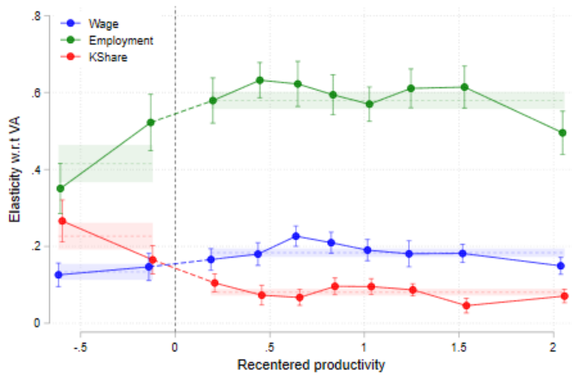
Notes: Figure shows robustness versions of main results in Figure 4. Each panel shows results when labour market fixed effects are defined by interactions of different geography and industry variables. There are 9 provinces and 52 district councils. The 1-digit industry level has 9 industry categories while the 2-digit level has 50 industry categories. The specification is otherwise the same as in Figure 4.

Figure A12: Internal IV results: trimming around the recentered productivity threshold

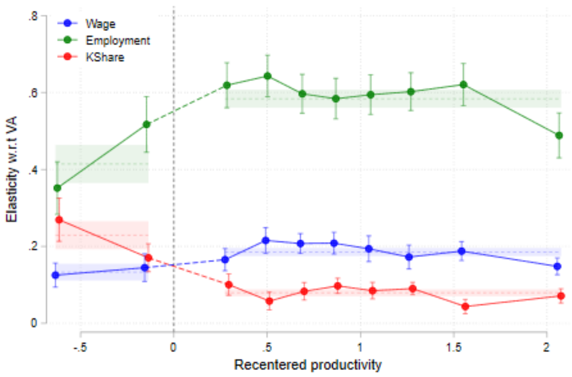
(a) No trimming



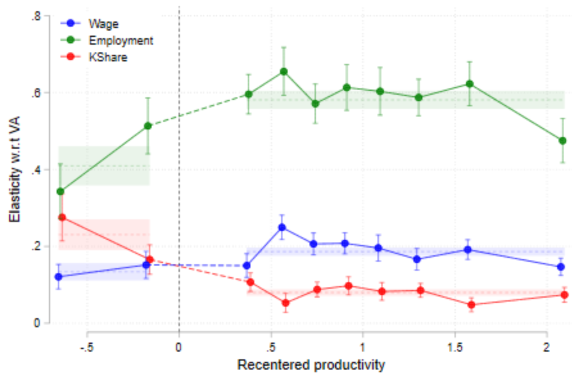
(b) Trimming at 1%



(c) Trimming at 5%



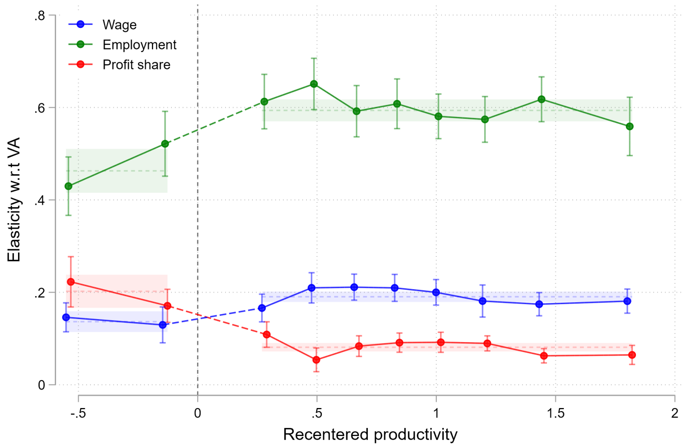
(d) Trimming at 10%



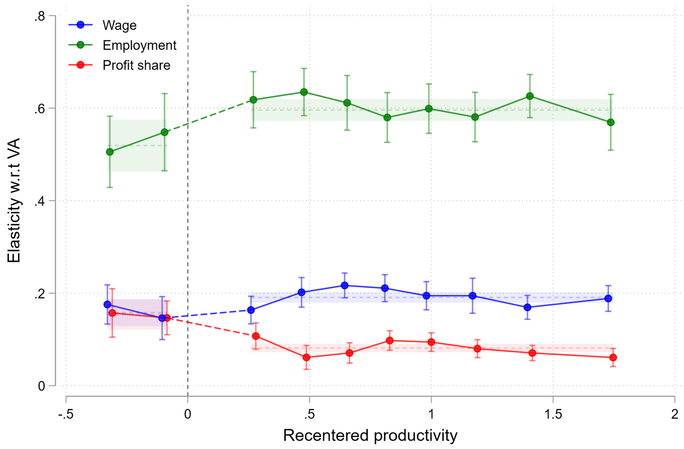
Notes: Figure shows robustness versions of main results in Figure 4. Each panel shows results for different choices of dropping firms close to the productivity threshold, where "Trimming at x%" means dropping the most productive x% of constrained firms and least productive x% of unconstrained firms. The specification is otherwise the same as in Figure 4.

Figure A13: Internal IV results: trimming the tails of the recentered productivity distribution

(a) Equal proportion trimming (5%)

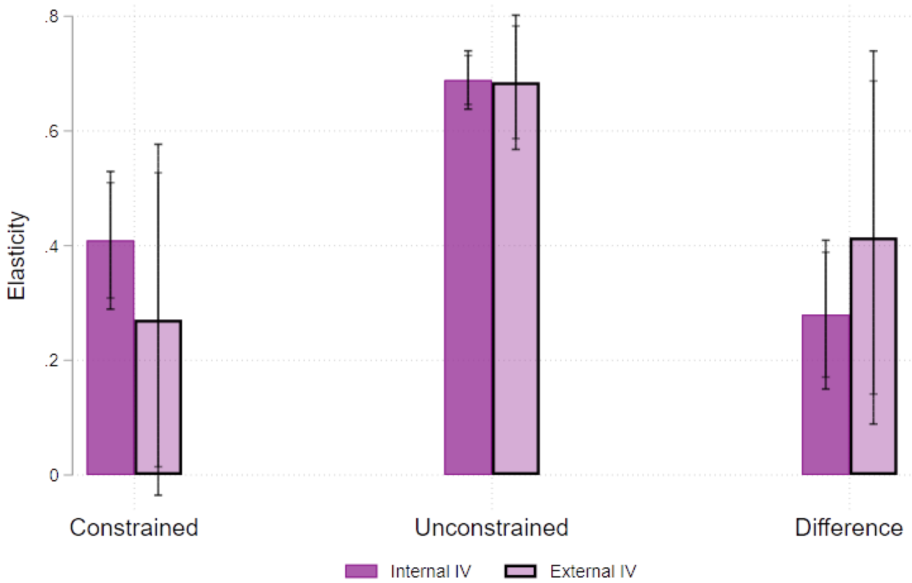


(b) Symmetric N trimming (5%)



Notes: Figure shows robustness versions of main results in Figure 4. Each panel shows results for different choices of dropping firms in the tails of the recentered productivity distribution. Panel (a) drops the 5% of the constrained firms with the lowest recentered productivity and 5% of the unconstrained firms with the highest productivity. Panel (b) similarly drops the lowest- and highest-productivity firms, but drops the bottom 5% of the combined firm distribution and the top 5% of the combined firm distribution (and therefore drops more firms, especially for constrained firms). The specification is otherwise the same as in Figure 4. Differences between the outcomes for constrained versus unconstrained firms are all significant at the 1% level in Panel (a) and at least at the 5% level in Panel (b).

Figure A14: Comparison of estimates from internal and external IV shocks: Hires



Notes: The figure plots the Hires estimates from Table 2 Panel (b), where the internal IV refers to above-median increases in firm VA between event periods -1 and 0 while the external IV refers to the shift-share trade shocks. See section 5 for sample and specification details. Vertical bars represent 90% and 95% confidence intervals, where standard errors are clustered at labor market by event and differences these are calculated using the Delta method.

Table A5: Rent sharing levels and pass-through estimates

	Internal IV (large VA change)		**External IV** (trade shock)	
	Constrained	Unconstrained	Constrained	Unconstrained
Rent sharing level	0.796	0.337	0.988	0.327
Rent sharing elasticity	0.132	0.186	0.129	0.188
Pass-through level	0.105	0.063	0.127	0.061

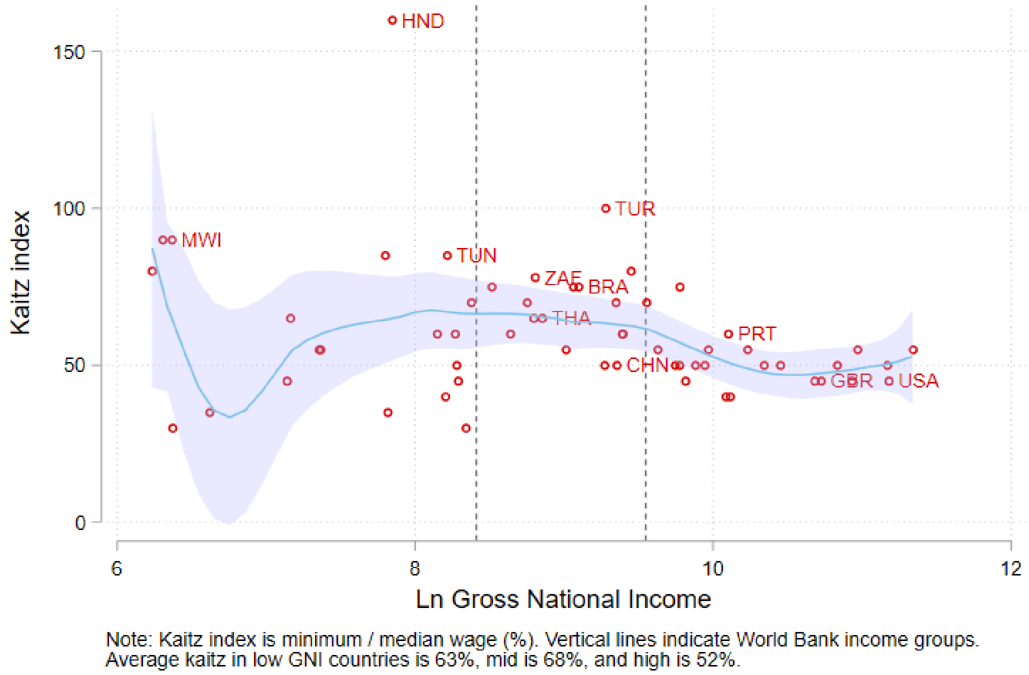
Notes: Table shows the rent sharing levels, rent sharing elasticities, and implied pass-through levels from the main specifications. The rent sharing elasticities come from Table 2. The rent sharing level is calculated as the firm total wage bill divided by firm value added, for the estimation sample upon which the rent sharing elasticity is calculated. The pass-through (the dollar increase in wages for a one dollar increase in value added) is the product of the elasticity and the level. Constrained and unconstrained firms are as defined in Table 2.

Table A6: External IV shock results using pre- and post-period restrictions

	External IV (trade shock)		
	Constrained	Unconstrained	Difference
Panel (a)			
Rent-sharing	**0.079**	**0.214*** **	**0.134**
	(0.1161)	(0.0449)	(0.1245)
Employment	**0.135**	**0.439*** **	**0.304**
	(0.2168)	(0.1533)	(0.2656)
Profit-share	**0.348** **	**0.040**	**-0.308***
	(0.1455)	(0.0667)	(0.1601)
F-stat	6.0	35.7	
N firms	956	12017	
Obs	6621	112122	
Panel (b)			
Hires	**0.296**	**1.380** **	**1.084**
	(0.5010)	(0.5508)	(0.7446)
F-stat	5.7	12.5	
N firms	918	11644	
Obs	5573	94257	

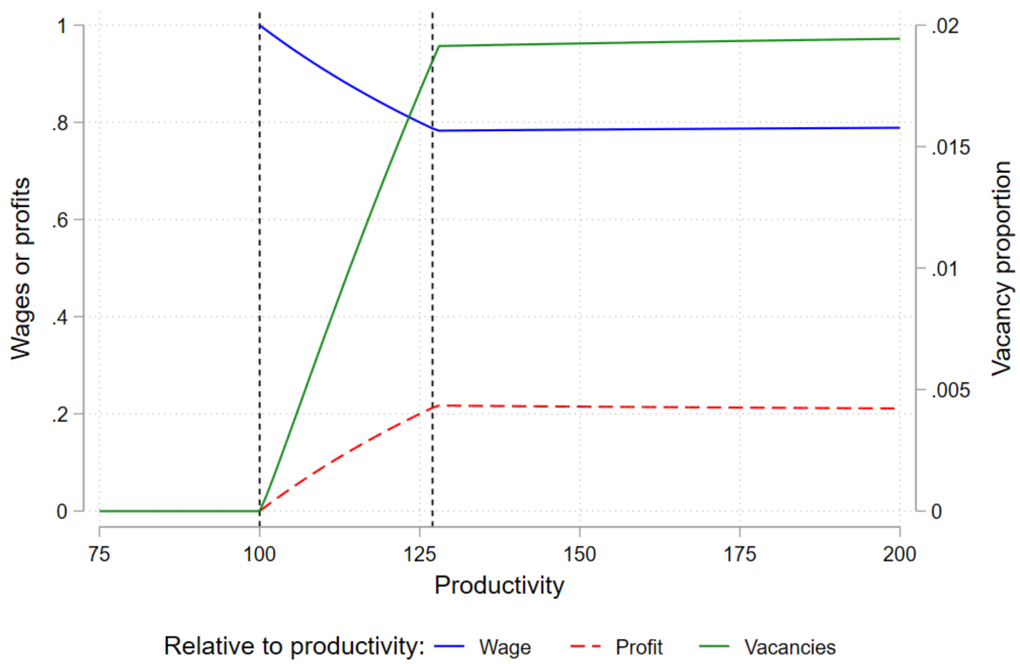
Notes: This table presents results analogous to the External IV results of Table 2 (the right-hand super column), but where the trade shock IV specification uses the pre- versus post-period structure of the data. Specifically, the export and import shares are estimated only over the pre-periods used in the internal instrument approach, and GDP shocks and effects on outcomes are only considered in the equivalent post-periods (see Equations 8 and 9). Standard errors are shown in parentheses and clustered at labor market by event; for differences these are calculated using the Delta method. * p < 0.1, ** p < 0.05, *** p < 0.01.

Figure A15: Kaitz Index by cross-country GNI



Notes: The plot shows the Kaitz index, defined as the ratio of the minimum wage to the median wage, by country. The x-axis is the log gross national income, using the Atlas method. The vertical lines indicate the World Bank classifications of lower (left) and middle (centre) income countries. Data are from ILO and World Development Indicators.

Figure A16: DMP, outcomes relative to productivity



Notes: The plot shows the DMP model with outcomes relative to productivity; see details of the model and simulation in Appendix B.2.

Appendix B: Models

B.1 Details of monopsony model

The full model in Manning (2003, pp. 338-345) is different from the simplified model in Section 2.1 mainly in that Manning (2003) incorporates the average market wage as a determinant of aggregate labour supply and a firm-specific supply-shifter b_i (e.g. disamenities), so that the firm-specific labour supply depends on the firm wage premium relative to the market wage and the firm-specific disamenity. The model below is essentially a stripped-down re-presentation of Manning (2003).

Specifically, retain Equation 2 for the demand for labour, but now model the share of total employment (N) supplied to firm i (N_i) as a function of its own wage (W_i) relative to an average market-level wage index (W) and the value of its disamenity (B_i): $\frac{N_i}{N} = \left(\frac{W_i}{B_i W}\right)^{1/\varepsilon}$. If one then models the labour supply to the whole market as $N = N_0 W^{\phi}$ and takes logs (again denoting logs of variables as lower case letters), the labour supply to the individual employer is

$$w_i = (1 - \varepsilon\phi)w + \varepsilon(n_i - n_0) + b_i \,,$$

or, subsuming n_0 into b_i and defining the coefficient on the average wage as θ,

$$w_i = \theta w + \varepsilon n_i + b_i \,. \tag{B1}$$

The marginal cost of labor in the absence of the minimum wage is then

$$\mathrm{mcl}_i = \ln(1+\varepsilon) + w_i = \ln(1+\varepsilon) + \varepsilon n_i + \theta w + b_i \,, \tag{B2}$$

which diverges from the simplified Equation 1 in its two additional terms reflecting the influence of the average wage and the firm-specific disamenity. Equating the expression for the MRPL in Equation 2 to the MCL above (Equation B2), and substituting in Equation B1, the firm's unconstrained wage is given by:

$$w_i^* = \frac{\eta\theta w - \varepsilon \ln(1+\varepsilon)}{\eta + \varepsilon} + v_i \tag{B3}$$

with

$$v_i = \frac{\varepsilon a_i + \eta b_i}{\eta + \varepsilon} \,, \tag{B4}$$

while the unconstrained employment level is

$$n_i^* = \frac{-\theta w - \ln(1+\varepsilon) + a_i - b_i}{\eta + \varepsilon} \,. \tag{B5}$$

With the introduction of a minimum wage w_{m}, the discussion of the simplified model in Section 2 explains how the value of a firm's "adjusted productivity" term v_i relative to the thresholds v^* and v_1^* determines which of the qualitative distinct demand-constrained, supply-constrained or unconstrained regions it falls into. Expressions for these threshold values can be derived by noting that v^* is the value of v_i where the unconstrained wage w_i^* is greater than or equal to the minimum wage w_{m}, so that, from Equation B3,

$$v^* = w_{\mathrm{m}} - \frac{\eta\theta w - \varepsilon \ln(1+\varepsilon)}{\eta+\varepsilon}\,. \tag{B6}$$

For firms which have $v_i < v^*$, for some it will be optimal to accept all workers forthcoming at the minimum wage w_{m}; these are supply-constrained firms. However for other firms with even lower v_i, it is not profitable to employ all the workers forthcoming at the minimum wage w_{m}; these are the demand-constrained firms. To find the threshold value of v_i which delineates these sets of firms, v_1^*, note that these firms set their wage at w_{m} but choose employment less than the potential supply at that wage so that $\mathrm{mrpl}_i = w_{\mathrm{m}}$. From Equations 2 and B1, we can resolve that

$$v_1^* = w_{\mathrm{m}} - \frac{\theta\eta w}{\eta+\varepsilon}\,. \tag{B7}$$

In order to find the equilibrium level of employment for supply-constrained firms, one can substitute $w_i = w_{\mathrm{m}}$ into the labour supply, Equation B1, which Manning (2003) shows can be expressed as:

$$n_i^{\mathrm{sc}} = n(w, a_i, b_i) + \frac{1}{\varepsilon}(v^* - v_i)\,, \tag{B8}$$

where $n(w, a_i, b_i)$ is the unconstrained employment level given in Equation B5. For our purposes it is useful to note that a_i does not enter Equation B1, and therefore does not enter the expression for n_i^{sc}, which reflects our main insight that equilibrium employment for supply-constrained firms is unaffected by local shifts in (revenue-) productivity, and in the special case where $b_i = 0$ all supply-constrained firms will have the same employment level, corresponding to the labour supplied at the minimum wage.[27]

To find the equilibrium employment for demand-constrained firms, again use that they will choose employment such that $\mathrm{mrpl}_i = w_{\mathrm{m}}$, and some rearranging leads to

$$n_i^{\mathrm{dc}} = n(w, a_i, b_i) + \frac{\ln(1+\varepsilon)}{\eta+\varepsilon} - \frac{1}{\eta}(v_1^* - v_i)\,. \tag{B9}$$

B.2 Baseline Diamond-Mortensen-Pissarides (DMP) model

We use a standard presentation of the baseline DMP model (e.g. Cahuc et al. 2014). A worker is matched with a firm vacancy through the matching function, which depends on labour market tightness $\theta = v/u$ (vacancies v over unemployment u). Workers and firms then split the surplus from the match, determined by worker bargaining power β, firm productivity p, and the cost of posting a vacancy c.

Setting reservation wage and discount rate to zero for simplicity, the standard wage curve is:

$$w = \beta(p + \theta c) \tag{B10}$$

[27] While v_i appears in Equation B8, expanding and simplifying necessarily causes the a_i term to drop out, and it resolves to $n_i^{\mathrm{sc}} = (w_{\mathrm{m}} - \theta w - b_i)/\varepsilon$.

And the job creation equations is as follows, where s is the exogenous job separation rate and η parametrizes the matching function (e.g. $m = u^{\eta} v^{1-\eta}$):

$$\theta = (\frac{p - w}{sc})^{1/\eta} \tag{B11}$$

Finally, the Beveridge curve pins down the ratio of vacancies to unemployment:

$$v = \theta \frac{s}{s + \theta^{1-\eta}} \tag{B12}$$

Since we are interested in models of firm heterogeneity, one can think of the relevant matching function as concerning the representative firm of a particular labour market (e.g. industry-region) with its own level of tightness.

To incorporate a minimum wage, we follow a similar procedure to above. B10 becomes a max function between the optimal wage in this equation and the mandated minimum wage. Then while firm productivity is below the minimum wage, there are no vacancies as firms post no matches. While wages are constrained by the minimum wage (i.e. the optimal wage is below the minimum wage), B11 applies with w as the minimum wage. The rest of the model is the same.

In the simulation, we use $\beta = 0.4$, $\eta = .9$, $c = 50$, $s = 0.25$, the minimum wage is 100, and firm productivity p varies from 75 to 200 in increments of 1. The model is qualitatively similar across a range of these parameter values.

B.3 Baseline union bargaining model

We focus on the insider-outsider union bargaining model since this seems most relevant to our setting of rent-sharing with productivity. Following Cahuc et al. (2014), a stock of "insider" L_0 workers at a firm are represented by a union, whose objective function is simply the utility of the wage premium (wage above the minimum wage $\bar{w}$, denoted b). The firm may hire "outsider" workers, and since these workers have no bargaining power the firm pays these workers the minimum wage. Firm profit from total workers L_u is therefore:

$$\pi = R(L_u) - \bar{w}L_u - bL_0 \tag{B13}$$

Given b, firms maximize L_u with respect to the minimum wage $\bar{w}$ with $R'(L_u) = \bar{w}$. Firms and unions maximize the following Nash with respect to b, where bargaining power is given by β, L_u is given as above, and we impose a simple linear utility function for insiders equal to the wage premium:

$$\max_{\{b\}} (R(L_u) - \bar{w}L_u - bL_0))^{1-\beta} (b)^{\beta} \tag{B14}$$

This gives the simple result that the wage premium is equal to the bargaining power parameter times by the quasi-rents $(R - \bar{w}L_u)$, shared across insiders:

$$b = \beta(R - \bar{w}L_u)/L_0 \tag{B15}$$

Profits are just the remainder portion of these rents:

$$\pi = R - \bar{w} - bL_0 = (1-\beta)(R - \bar{w}L_u) \tag{B16}$$

For simplicity, we use a Cobb-Douglas revenue function $R = AL^{\alpha}$. In the simulation, we use $\beta = 0.4$, $\alpha = .7$, a minimum wage of 1.5, and iterate the productivity shifter A between 1 and 1,000. We also take a simple view of the generating process of insider workers L_0: we take a maximum to the size of the union (in simulations, 30 workers). One can of course imagine that L_0 grows over time with L_u, perhaps for workers who have been at the firm for a few years, and so one can view the simulated process as a local rent-sharing dynamic conditional on time.[28] As above, the model is qualitatively similar across a range of these parameter values.

While the above focuses on the insider-outsider model, well-known alternative union bargaining models include the right-to-manage model and weakly efficient bargaining over wages and employment. However, these baseline models do not always generate rent-sharing as they require a non-homogenous production function such as the CES function (see also Manning 1993). Cahuc et al. (2014) notes regarding the right-to-manage model, and we confirm this through simulations and derivations, "If the revenue function of the firm is homogeneous of degree alpha (0,1), then [...] shocks to productivity or the firm's selling price do not affect the wage and lead only to employment adjustments." A similar condition holds for weakly efficient bargaining, for example with the revenue function $R = AL^{\alpha}$. The intuition for why this fails to deliver rent-sharing is that productivity shifts A do increase L, as well as w if *conditional on* L; however, marginal productivity of labour decreases with greater L and it turns out the increase in w due to A is exactly offset by the decrease in marginal productivity through L.

[28] If L_0 is equal to $L_{u,t-1}$, then b is negligible as union welfare comes through expansions to the number of workers rather than the premium. This is very similar to the lack of rent-sharing in the right-to-manage model where the union bargains only over the wage which applies to all workers. In fact this is the case in this model when L_0 is less than its maximum, see figure 6.

Appendix C: Data appendix

This data appendix is created as per UNU-WIDER requirements for users of the National Treasury Secure Data Facility (NT-SDF). It reports on data directly used for the results presented in this paper, and does not include other variables and programs used in our ongoing research on this topic.

Data Access

The data used for this research was accessed from the NT-SDF. Access was provided under a non-disclosure agreement, and our output was checked so that the anonymity of no firm or individual would be compromised. Our results do not represent any official statistics (NT or SARS). Similarly, the views expressed in our research are not necessarily the views of the NT or SARS.

Data used: CIT-IPR5 panel (`citirp5_v5_0`); year-by-year IRP5 job-level data (`v5`), and year-by-year transaction-level Customs data (beta version `5_0`). Date of first access for this project: 6 January 2023. Last accessed: 10 March 2025.

Software

Our analysis was conducted using Stata 17 and 18. User-written programs and schemes used include `reghdfe` (Correia 2014), `gtools` (Bravo 2018), `prodest` (Rovigatti and Mollisi 2016), `loghockey` (Lunt nd), `ivreg2` (Baum et al. 2002), `ivreghdfe` (Correia 2018), `binsreg` (Cattaneo et al. 2024), `binscatter` (Stepner 2013) and `plotplain` (Bischof 2017).

Variables

Variables used from the raw IRP5 data include: `taxyear taxrefno payereferenceno dateofbirth idno passportno province_geo busprov_geo districtmunicip_geo busdistmuni_geo periodemployedfr periodemployedto totalperiodsinyearofassessment totalperiodsworked`.

Employment income was created from the following IRP5 amount codes: `amt3601 amt3605 amt3606 amt3607 amt3615 amt3616`. A record of employment-related allowances was created from the following IRP5 amount codes: `amt3701 amt3704 amt3710 amt3711 amt3712 amt3713 amt3715`.

IRP5 employment records were identified by records which had non-zero income or allowances; those with zero or missing income and allowances data are dropped from the analysis.

Variables used from the CIT-IRP5 data include: `taxyear finyear FYE taxrefno g_sales g_cos g_grossprofit g_grossloss k_ppe k_faother comp_prof_sic5_1d comp_prof_sic5_2d`.

Value-added was calculated by subtracting cost of sales from gross sales.

The "composite profit code" industry variables we use were created by Budlender and Ebrahim (2020). We merge in Bargaining Council variables created by Bassier (2022).

Variables used from Customs data include: `taxrefno cust_refno calendaryear yearmonth countryoforigin countryofexport countryofdestination customsvalue purposecode purposecodedes flag`.

Exports were identified with purpose codes `EX1`, `ELG` or `ZE`. Imports were identified with purpose codes `WE`, `IM5`, `XRW` or `IM6`.

Cleaning & sample notes

Our analysis is conducted at the CIT level; PAYE entities without CIT tax reference numbers are excluded from the sample. CIT entities not matched to PAYE entities are also excluded, as this is primarily an employment analysis. When matching firm ITR14/IT14 balance sheet information to the IRP5 data, we match on the basis of firm financial years which best overlap with the tax years used in the IRP5 data. Full-time equivalent employment is calculated using the "periods worked" variables. We use the unbalanced firm-level panel, which is only balanced after creating stacked events. When industry or location data is missing for a particular firm, we iteratively forwards and backwards impute these variables. These notes represent some particularly noteworthy data cleaning and sample construction decisions, but for full details users are referred to our do-files which are available at the NT-SDF.

Appendix D: Estimation sample

To implement the LMS internal instruments approach of Section 5 we must create a dataset of "stacked" events (or cohorts) from the panel of firms. This division of time periods into pre- and post-periods for specific firms is also useful for our cross-sectional kink design of Section 4 (the results of which are used in the within-firm shock approaches of Section 5), and so we use this basic data structure throughout the empirical analysis.

D.1 Cross-sectional kink design

We start by dividing firms in the 2010-2019 period into four events (or cohorts), with the potential treatment date starting in 2014, 2015, 2016 and 2017 respectively. A firm may be in multiple events/cohorts, if it is observed in the panel for more than one of these treatment date starts, and is observed both prior to and after that treatment date. We call years prior to treatment start the pre-period, and including and after treatment start the post-period.

Separately for each cohort, we estimate productivity for each firm per Section 4 using only pre-period years.

Separately for each cohort (and each BC/SD), we then estimate the wage-kinks per the procedure in Section 4 using only post-period years. From this, we have a recentered productivity measure for each firm in each event. For our aggregate Tables and Figures in Section 4 we pool (or "stack") across events/cohorts (accounting for the stacked design when clustering our standard errors).

Using this stacked approach for the cross-section kink design as well as the within-firm IV analysis is useful because 1) it makes it straightforward to define event-specific recentered productivity values (and therefore constrained versus unconstrained firms) which we need for our internal IV analysis and 2) it provides a natural way to estimate productivity in a pre-period, so it is a fixed heterogeneity category for the cross-sectional and shock analyses in a post-period, without simply bifurcating the original panel and losing many observations 3) it allows a firm to have *some* time-varying productivity, so that a firm could be low productivity and constrained in earlier events and higher-productivity and unconstrained in later events. We do show in Appendix Figure A6 that the results of Section 4 do not depend on this pooling across events.

D.2 Internal instrument approach

The sample restrictions and key variable restrictions are an important feature of the LMS strategy, and we follow them closely, constructing our sample as follows:

1. Identify "stayers" in the worker-level data who remain employed at the same firm for 8 consecutive years, separately defining stayers for the tax year event periods 2010-2017, 2011-2018, 2012-2019, and 2013-2020. These are the cohorts/events mentioned immediately above, with 2010-2020 covering the usable period of the employment data. Drop stayers' records in the first and last years of this tenure (when they may have entered or

separated), and only keep workers who are full-time employed over this 6 year period at their firm. Count the number of stayers in each firm for each event period and create year-specific firm-level statistics for stayers' wages (specifically median wage).

2. For each event, only keep firms which have at least 2 stayers. LMS use a 10-stayers minimum as their baseline, but this is overly restrictive when it comes to the South African firm-size distribution and a labour market context defined by high churn (Kerr 2018). In our baseline specification we use a 2-stayer minimum, to mitigate measurement error in one-stayer firms where the one stayer may be an owner or otherwise unrepresentative of employer/employee dynamics. In Appendix Figure A10 we show that our results are not sensitive to the number of stayers.

3. Over the 6-year period for each event, we treat the first three periods as the pre-period and the latter three as the post. Treatment is defined as an above-median increase in firm value-added between periods -1 and 0 for each event, where the median increase is weighted by firm size. Events are stacked (Cengiz et al. 2019). Period -2 is used as the omitted reference period to allow for some mean reversion dynamics in period -1, as in LMS. For the same reason, periods 1 and 2 are considered the post periods of interest, rather than period 0 (results are essentially unchanged if we use only period 2 instead). Period -3 is used to assess pre-period parallel trend violations.

When estimating Equation 7 by productivity bin, and for constrained versus unconstrained firms (e.g. Table 2 and Figures 4, 5 and A7) we make a few small data construction/visualization decisions for the internal IV specification:

1. As discussed in Section 5 above, there is measurement and estimation error in the $\hat{v}^*$ threshold, and so we drop firms very close to the threshold before constructing bins. In our baseline specification we drop the 5% of constrained firms with the highest recentered productivity values, and the 5% of unconstrained firms with the lowest such values. Our results are not sensitive to this trimming procedure; see Appendix Figure A12.

2. We divide firms into 10 approximately equally-sized bins to get a sense of the shape of response along the distribution. We require that there be at least 2 bins on either side of $\hat{v}^*$, as the value of the bins is in seeing the shape of the marginal response against productivity, which means in practice that the bins in the constrained region are smaller than those in the unconstrained region; 15% of firms (or 12% of event-specific observations) fall in the constrained region in this baseline specification.

3. In Figure 4 and Panel (b) of A7, treatment effects are the average response across post-periods 1 and 2; results are essentially unchanged if we only use post-period 2.

D.3 External instrument approach

Due to the power issues discussed in Section 5.1.2, for our external instruments approach we do not use the event structure of the dataset for identification in our main results, but simply pool across firms in the internal IV estimation sample. This means we do impose restrictions such as requiring a balanced panel with a two-stayer minimum, but do not separate into pre- and post-periods, except for the supplementary exercise mentioned in footnote 20 and shown in Appendix Table A6.

Appendix E: Kaitz Index heterogeneity

To calculate the Kaitz Index we use worker-level data. In all cases the minimum wage for a particular worker is the minimum wage associated with their BC/SD. In the BC/SD specification, we calculate the median wage over the BC/SD, as the median annualized wage in the BC/SD. For the firm-specific Kaitz, we calculate the median wage as the median annualized wage in the detailed geography (DC) by 2-digit industry intersection, to better capture the relevant labour market. Therefore, for each firm, its Kaitz Index is the ratio of the BC/SD minimum wage to either the BC/SD median wage (which we refer to as the BC/SD Kaitz) or the industry-region median wage (which we refer to as the firm-specific Kaitz). Overall, the Kaitz indices we calculate are consistent with external data, considering that the median wage in our data is higher because we only observe formal firms: 0.68 for the BC/SD specification and 0.69 for the firm-specific specification, compared to 0.78 for South Africa in the ILO and World Development data shown in Figure A15.

As would be predicted by theory, firms with higher Kaitz Indices are more likely to be found in the constrained region, which is consistent with the estimated wage kink identifying the threshold below which firms are bound by the minimum wage. Figure E1 shows clear positive relationships between both Kaitz measures and the probability of a firm being found in the constrained region.

We test the statistical significance of these relationships by regressing each firm's constrained status (a dummy variable equal to 1 for constrained firms) on its Kaitz Index, for both the underlying continuous Kaitz Index and a binary transformation which equals 1 if the Kaitz Index is above its median value (Table E1).

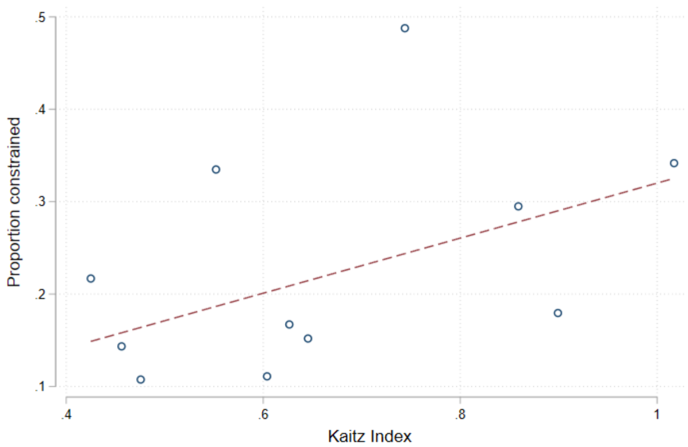
Table E1: Constrained probability by Kaitz Index

	BC/SD		Firm-specific	
Kaitz coefficient	.164	.298	.123	.229
	(.0696)	(.1534)	(.0419)	(.0815)
Constant	.165	.022	.154	.06
	(.0151)	(.0866)	(.0129)	(.0462)
N	324608	324608	324608	324608
Binary	Y		Y	
Continuous		Y		Y

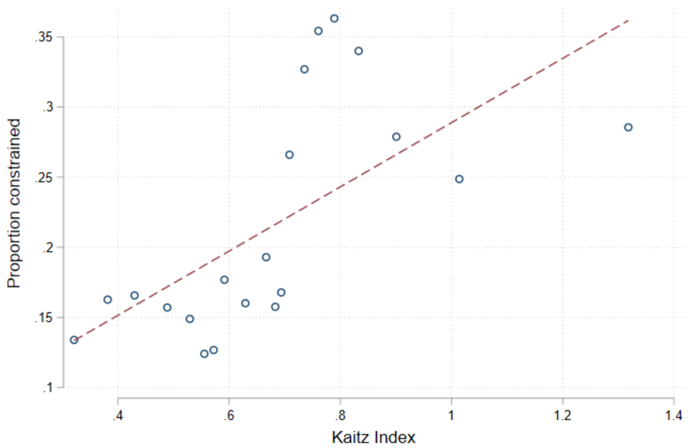
Notes: Table shows results of regressions of each firm's constrained status (a dummy variable equal to 1 for constrained firms) on its Kaitz Index, for both the underlying continuous Kaitz Index and a binary transformation which equals 1 if the Kaitz Index is above its median value. Standard errors are shown in parentheses.

Figure E1: Probability of being in constrained region, by Kaitz Index

(a) BC/SD Kaitz



(b) Firm-specific (DC × 2-digit industry) Kaitz



Notes: Figure shows bin scatter plots of firm Kaitz Index measures (x-axis) against firm constrained status (y-axis; a dummy variable equal to 1 for constrained firms).