Returns to Scale, Productivity Measurement, and Trends in U.S. Manufacturing Misallocation*

Sui-Jade Ho
Bank Negara Malaysia

Dimitrije Ruzic†
INSEAD

April 10, 2021

Abstract

Aggregate productivity suffers when workers and machines are not matched with their most productive uses. This paper builds a model that features industry-specific markups, industry-specific returns to scale, and establishment-specific distortions, and uses it to measure the extent of this misallocation in the economy. Applying the model to restricted U.S. census microdata on the manufacturing sector suggests that misallocation declined by 13% between 1982 and 2007. The finding of declining misallocation starkly contrasts with the 29% increase implied by the widely-used assumptions that all establishments charge the same markup and have constant returns to scale.

Keywords: returns to scale, productivity, misallocation, manufacturing

JEL Codes: D24, E23, E25, O47

*We would like to thank Andrei Levchenko, Matthew Shapiro, Stefan Nagel, and Joshua Hausman for invaluable guidance, suggestions, and encouragement. We also thank John Fernald, Kyle Handley, Bart Hobijn, Peter Klenow, and Sebastian Sotelo for helpful discussions. All remaining errors are our own. Any opinions and conclusions expressed herein are those of the authors and do not necessarily represent the views of the U.S. Census Bureau or of the Bank Negara Malaysia. All results have been reviewed to ensure that no confidential information is disclosed.

†Email: dimitrije.ruzic@insead.edu (corresponding author)

1 Introduction

Aggregate productivity retreats from its frontier when workers and machines mismatch with their most productive uses. Formalized elegantly by Restuccia and Rogerson (2008), this notion of misallocation has the potential to explain why countries differ in their incomes, or why aggregate productivity changes over time. Yet, quantifying the extent of this misallocation is challenging, in part because we do not observe productivity directly. Most commonly, we must infer an establishment's total factor productivity from its revenue. This inference is a two-step process: first, we must deduce how establishments set prices, so we can map revenue to output; then, we must deduce how they produce, so we can map output to productivity.

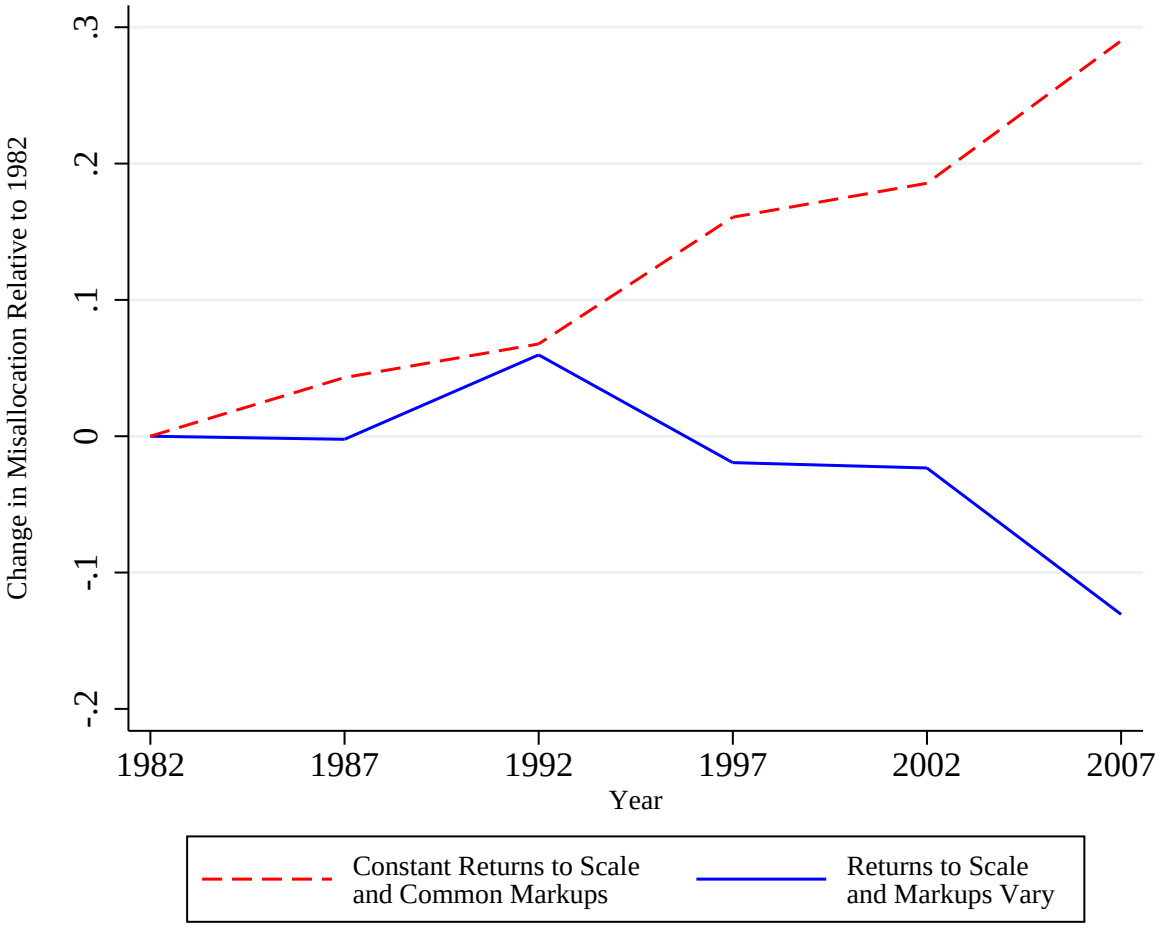
To infer productivity and measure misallocation, this paper builds a quantitative model in which returns to scale and markups of price over marginal cost differ across industries and time. We implement the model on restricted U.S. Census microdata covering the U.S. manufacturing sector from 1977 through 2007. In the process, we jointly estimate markups and returns to scale for individual industries over time. Our estimates show that industries differ meaningfully in both markups and returns to scale, with standard deviations across industries of about one-third the average level of the respective parameters. Moreover, while the average markup remained relatively constant over this period, the average returns to scale fell, starting off as increasing and ending as nearly constant. We use these parameters to infer productivity, and find that misallocation in U.S. manufacturing declined 13% between 1982 and 2007.

Allowing for heterogeneous markups and returns to scale is crucial when estimating productivity and misallocation. The widely-used Hsieh and Klenow (2009) model is a special case of our framework in which all industries have a common markup and constant returns to scale. Figure 1 contrasts the downward trend in misallocation under our estimated parameters with the upward trend implied by the Hsieh-Klenow assumptions. Both measures of misallocation answer the question: how much more productive would the U.S. manufacturing sector be if it were as misallocated today as it was in 1982? If misallocation by this measure has increased, productivity today would be higher at 1982 levels of misallocation. Indeed, as the dashed red line shows, the assumptions of a common markup and constant returns to scale suggest a 29% increase in misallocation over the last 25 years. By contrast, the solid blue line traces out the declining trend in misallocation from our model.

We arrive at the declining trend in misallocation by estimating markups and returns to scale using a control-function approach rooted in Olley and Pakes (1996) and Levinsohn

Figure 1: Misallocation in U.S. Manufacturing

Change in U.S. Manufacturing TFP at 1982 Levels of Misallocation



Note: Misallocation is the distance between aggregate productivity and a frontier where marginal revenue products are equalized across establishments in each industry. Positive (negative) values indicate an increase (decrease) in misallocation relative to 1982.

and Petrin (2003). Our estimating procedure infers markups and returns to scale even in datasets, like the U.S. Census microdata, where we observe revenues, but not output or prices. For this procedure, we derive a model-based estimating equation that relates establishment revenue to its inputs and to industry size, as in De Loecker (2011). We map the reduced-form revenue elasticities to the markup and returns-to-scale parameters using model equations. In line with prior empirical work [e.g. Hall (1990), Basu and Fernald (1997), Basu et al. (2006), Broda and Weinstein (2006)], we find that both markups and returns to scale indeed vary across industries. Moreover, the average markup for U.S. manufacturing has remained relatively constant over time, while returns to scale have declined, starting off as increasing in 1982 and ending as nearly constant by 2007.

The reduction in returns to scale—from increasing to nearly constant—can be interpreted as a narrowing gap between average and marginal cost. We argue that this reduc-

tion in returns to scale is intuitive. A common motivation for increasing returns is the presence of fixed costs (to build an establishment, to add an assembly line, to acquire more land, etc.,). Viewed through that lens, our estimates suggest that fixed costs are being spread over more units of output, bringing average cost closer to marginal cost. Indeed, the U.S. manufacturing sector more than doubled in size between 1977 and 2007, with value added increasing 132% in real terms. For returns to scale *not* to have declined, fixed costs would have had to grow at least as much as output. This growth of fixed costs could have taken place either through an extensive margin with a growing number of establishments, or through an intensive margin with growing fixed costs within each establishment. We argue that neither margin seem positioned to overturn the reduction in returns to scale. First, the number of manufacturing establishments grew only 2% over this period, making the extensive margin an unlikely source of growth for fixed costs. Second, while within-establishment fixed costs are notoriously difficult to measure directly, we use external data on automotive plants as a case study: we show that many first-order sources of fixed costs—number of manufacturing platforms, number of vehicle models per platform, plant surface area—have not kept pace with output, suggesting the estimated reduction in returns to scale as a sensible feature of the data.

We show that the decline in returns to scale is the key to rationalizing the different trends in misallocation between our model and the Hsieh-Klenow model. In short, ignoring the variation in markups and returns to scale leads to measures of productivity that conflate productivity and distortion. These conflated measures of productivity lead to incorrect inferences about the extent to which the most productive establishments bear the largest distortions, and hence lead to incorrect measures of misallocation. Our estimates suggest that the Hsieh-Klenow model understates misallocation on average. Over time, as the assumption of constant returns better fits the data for the U.S. manufacturing sector, the Hsieh-Klenow model understates misallocation less and less. This better fit drives the apparent upward trend in misallocation under a common markup and constant returns.

Outside their relevance for measuring productivity and misallocation, the patterns we document for markups and returns to scale also fit with the recent literature on the decline of the labor share, and, more broadly, the changing division of value added. For instance, a large literature documents a thirty-year decline in labor's share of value added both for the United States and for other economies [e.g., Elsby et al. (2013), Karabarbounis and Neiman (2014), Barkai (2020)]; we find this decline to be even larger for the U.S. manufacturing sector. Within that literature, using different approaches, both Karabarbounis and Neiman (2014) and Barkai (2020) suggest that the decline in labor's share of value added might not have been offset by an equivalent increase in the capital share. The result-

ing implication is that the share of profits in value added increased over time. Indeed, De Loecker et al. (2020) find evidence of rising profit rates both among U.S. publicly traded firms and in the national income accounts.

In contrast to a recent literature emphasizing changes in markups (see Basu (2019) for a survey), we find that the rising profit share for the U.S. manufacturing sector has been driven primarily by the reduction in the returns to scale. Most work in this literature shares a common idea: changes in factor shares can be understood as changes in either markups or in returns to scale. We jointly estimate industry-level parameters underlying demand and production from data on revenue. By estimating the parameters at the same level of aggregation, we can readily compare the relative contributions of returns to scale and of markups to changing factor shares. By contrast, approaches in the spirit of De Loecker et al. (2020) differ along two broad dimensions. First, rather than jointly estimating production and demand from data on revenue, these approaches treat revenue as a proxy of output. This approach would not identify the markup in our model, consistently with Bond et al. (Forthcoming) who show in a more general setting that using a revenue elasticity in place of an output elasticity provides no information about markups. Second, these approaches estimate industry-level production parameters and infer firm-level markups. By allocating many more degrees of freedom to markups, these approaches provide markups with more explanatory power. Although we emphasize our parameter estimates, we also show in a later robustness check that the divergent trends in misallocation we document are robust to attributing all changes in profits to rising markups.

In light of recent evidence from Autor et al. (2020) and Kehrig and Vincent (2021) that larger firms have systematically lower labor shares, we also generalize the baseline model to introduce markups—and therefore factor shares—that vary across establishments in an industry. We find that this generalization to firm-specific markups—modeled in the spirit of Atkeson and Burstein (2008)—implies lower levels of misallocation, yet still features divergent trends in misallocation between our model and the Hsieh-Klenow model. The generalization of the baseline model supplements the work of a growing literature that continues to refine the measurement of distortions (see Hopenhayn (2014) for a review). We show that, conditional on an industry-specific demand elasticity, the additional variation in markups changes the marginal revenue products of establishments and hence their measured distortions. In changing the level of measured misallocation, the generalization to variable markups within an industry is similar to Edmond et al. (2018) and other work where richer depictions of establishment behavior reduce the level of measured misallocation [e.g. Bartelsman et al. (2013), Asker et al. (2014), Gopinath et al. (2017)], and those who emphasize measurement issues that make inferring misallocation challenging

[e.g. White et al. (2018), Bils et al. (2017), Haltiwanger et al. (2018)].

While the core of this paper reflects a model where firms combine capital and labor to produce value added, we also provide evidence that the same patterns of misallocation hold for the production of gross output. Since a recent literature has highlighted the limitations of standard control-function methods for estimating returns to scale in gross output [e.g., Ackerberg et al. (2015), Gandhi et al. (2020)], we present three complementary approaches that yield parameter estimates for a gross-output version of our model. Although none of the three approaches can simultaneously overcome all the measurement challenges highlighted in the literature, all estimates overturn the sharp rise in misallocation from the Hsieh-Klenow setting. Moreover, two sets of estimates show a decline in returns to scale, while in the third we impose constant returns for identification purposes.

Within the recent literature on misallocation, our paper's closest counterparts are two works that emphasize the importance of measurement within the Hsieh-Klenow model: Bils et al. (2017) and Haltiwanger et al. (2018). The former explains the upward trend in U.S. manufacturing misallocation as an artefact of measurement error that increased over time. While we think measurement error is an important topic to address in the microdata, we show in Appendix F that the Bils et al. (2017) procedure risks conflating measurement error with model misspecification if returns to scale are not constant: ignoring a decline in returns to scale, like the one we document, could lead an econometrician to infer an increase in measurement error. The latter paper, Haltiwanger et al. (2018), uses eleven manufacturing products to show that deviations from production and demand assumptions in the Hsieh-Klenow model lead to estimates of establishment-level distortions that behave differently than the distortions in the baseline model. We share their emphasis on deviations from standard Hsieh-Klenow assumption and view the works as complementary.

The remainder of the paper proceeds as follows. In section 2 we derive a measure of misallocation in a model that allows for variation in markups and returns to scale; we then develop a toolkit to understand the discrepancies in measured productivity and misallocation that arise from ignoring the variation in these parameters. We map the model to the data, detail the estimation procedure, and present the estimates of markups and returns to scale in section 3. Section 4 presents our measure of misallocation and uses the toolkit to explain why our measure deviates from the Hsieh-Klenow measure that assumes a common markup and constant returns. Section 5 highlights the robust difference between the trends in misallocation in the two models across a number changes in model structure and estimation. Section 6 concludes.

2 Model

We build a model that features industry-specific markups, industry-specific returns to scale, and establishment-specific distortions. We then show how ignoring the variation in markups and returns to scale leads to measures of productivity that conflate productivity and distortions, and leads to incorrect measures of misallocation.

2.1 Deriving a Measure of Misallocation

In this section, we derive a measure of misallocation for the aggregate economy, accounting for industry variation in markups and returns to scale. We measure misallocation as the distance between aggregate productivity and a frontier where inputs are reallocated so that marginal revenue products are equal across establishments in each industry. We proceed in three steps. First, we show the aggregation in the model, allowing us to map from the distortions that establishments face to aggregate misallocation. Second, we show how establishments optimally respond to the distortions they face; these expressions allow us to characterize establishment behavior when we reallocate resources and change the distortions that they face. Third, we derive a measure of misallocation by comparing aggregate productivity before and after resources are reallocated. Since much of this derivation is standard in the literature, here we highlight the structure of the model and the key inputs into the measure of misallocation. We refer interested readers to appendix A for more details.

Aggregation

A representative firm aggregates the output Y_i of I different industries using a Cobb-Douglas production technology, and sells the aggregate output Y in a perfectly-competitive market, as in (1):

$$\text{Aggregate} \qquad Y = \prod_{i=1}^{I} Y_i^{\theta_i} \text{ with } \sum_{i=1}^{I} \theta_i = 1 \qquad\qquad P = \prod_{i=1}^{I} (P_i/\theta_i)^{\theta_i} = 1. \tag{1}$$

Cost minimization by this aggregating firm implies that the elasticities θ_i from the production function correspond to the share of each industry's value added ($P_i Y_i$) in aggregate value added (PY). This insight allows us to define the aggregate price index P, which we choose as the numeraire.

Within each industry, an aggregating firm combines the output Y_{ie} of N_i differentiated

establishments using a constant-elasticity-of-substitution (CES) technology, as in (2):

$$\text{Industry} \qquad Y_i = \left(\sum_{e=1}^{N_i} Y_{ie}^{\frac{\sigma_i - 1}{\sigma_i}} \right)^{\frac{\sigma_i}{\sigma_i - 1}} \qquad P_i = \left[\sum_{e=1}^{N_i} \left(\frac{1}{P_{ie}} \right)^{\sigma_i - 1} \right]^{\frac{-1}{\sigma_i - 1}}. \tag{2}$$

The CES aggregator implies that each establishment in the industry faces a downward-sloping demand curve for its output. Cost minimization by the industry aggregating firm leads to the standard CES price index P_i. Note that that the elasticity σ_i can potentially vary across industries.

Each establishment in the industry produces value-added output Y_{ie} by combining its total factor productivity A_{ie}, capital K_{ie}, and labor L_{ie} using the Cobb-Douglas production function in equation (3):

$$\text{Establishment} \qquad Y_{ie} = A_{ie} K_{ie}^{\alpha_{K_i}} L_{ie}^{\alpha_{L_i}}, \quad \alpha_i = \alpha_{K_i} + \alpha_{L_i} \, . \tag{3}$$

The returns to scale in production are α_i, the sum of output elasticities α_{K_i} and α_{L_i}; when returns to scale differ from unity, we have non-constant returns to scale. Moreover, returns to scale can differ across industries. We discuss the generalization of this model to gross-output production in section 5.

Optimization

Each establishment maximizes profits π_{ie} by choosing how much capital and labor to hire:

$$\pi_{ie} = P_{ie} Y_{ie} - (1 + \tau_{L_{ie}}) w_i L_{ie} - (1 + \tau_{K_{ie}}) R_i K_{ie}. \tag{4}$$

The establishment takes as given the input prices R_i and w_i from perfectly competitive input markets; however, the effective cost of an input varies across establishments, with $\tau_{K_{ie}}$ and $\tau_{L_{ie}}$ capturing the input-specific distortions for capital and labor. Consider, for instance, regulations that mandate the benefits that establishments have to provide to workers. These regulations change the effective cost of hiring labor. If two establishments are subject to different regulations, then these establishments also differ in their $\tau_{L_{ie}}$.

Establishments that face large distortions have high marginal revenue products. The first-order conditions from profit maximization, shown in equation (5) for capital and equation (6) for labor,

$$MRPK_{ie} = \frac{\alpha_{K_i}}{\frac{\sigma_i}{\sigma_i - 1}} \frac{P_{ie} Y_{ie}}{K_{ie}} = (1 + \tau_{K_{ie}}) R_i \tag{5}$$

$$MRPL_{ie} = \frac{\alpha_{L_i}}{\frac{\sigma_i}{\sigma_i - 1}} \frac{P_{ie} Y_{ie}}{L_{ie}} = (1 + \tau_{L_{ie}}) w_i, \tag{6}$$

show that establishments trade off the marginal contribution to revenue of a given input ($MRPK_{ie}$ or $MRPL_{ie}$) against the effective cost of hiring it. For instance, an establishment facing a cost-increasing labor regulation has a large $\tau_{L_{ie}}$; this establishment will hire labor until the contribution to revenue of the last unit hired, $MRPL_{ie}$, exactly offsets the effective cost of hiring labor $(1 + \tau_{L_{ie}})w_i$. In short, faced with larger distortion, the establishment requires larger marginal revenue products to justify hiring inputs. Moreover, in the absence of distortions, marginal revenue products are equalized in an industry. This notion will help define a productivity frontier and subsequently misallocation.

Optimal responses to larger distortions lead establishments to charge higher prices. The establishment price in equation (7) is a markup over marginal cost:

$$P_{ie} = \underbrace{\frac{\sigma_i}{\sigma_i - 1}}_{Markup} \underbrace{\left[\left(\frac{R_i}{\alpha_{K_i}} \right)^{\alpha_{K_i}} \left(\frac{w_i}{\alpha_{L_i}} \right)^{\alpha_{L_i}} \right]^{\frac{1}{\alpha_i}} \left(Y_{ie} \right)^{\frac{1-\alpha_i}{\alpha_i}} \left[\frac{(1 + \tau_{K_{ie}})^{\alpha_{K_i}} (1 + \tau_{L_{ie}})^{\alpha_{L_i}}}{A_{ie}} \right]^{\frac{1}{\alpha_i}}}_{Marginal\ Cost}. \tag{7}$$

The model allows the markup $\sigma_i/(\sigma_i - 1)$ in equation (7) to be industry specific. Furthermore, the introduction of potentially non-constant returns to scale allows the marginal cost to change with the establishment's scale of production. Under the standard assumption of constant returns to scale ($\alpha_i = 1$), marginal cost is constant and independent of output Y_{ie}. However, if returns to scale deviate from unity ($\alpha_i \neq 1$), then marginal cost is increasing in output for decreasing returns to scale, and vice versa. Lastly, larger distortions increase the marginal cost of production and thus force the establishment to charge a higher price.[1]

An establishment responds to large distortions by choosing a smaller input bundle and shrinking in size. Since much of this paper is about the allocation of resources across establishments in an industry, the relevant measure of size captures the establishment's value added relative to the value added of the industry, s_{ie} in equation (8):

$$s_{ie} = \frac{P_{ie} Y_{ie}}{P_i Y_i} = \frac{\left[A_{ie} \left(\frac{\overline{1 + \tau_{K,i}}}{1 + \tau_{K_{ie}}} \right)^{\alpha_{K_i}} \left(\frac{\overline{1 + \tau_{L,i}}}{1 + \tau_{L_{ie}}} \right)^{\alpha_{L_i}} \right]^{\frac{1}{\frac{\sigma_i}{\sigma_i - 1} - \alpha_i}}}{\sum_{e=1}^{N_i} \left[A_{ie} \left(\frac{\overline{1 + \tau_{K,i}}}{1 + \tau_{K_{ie}}} \right)^{\alpha_{K_i}} \left(\frac{\overline{1 + \tau_{L,i}}}{1 + \tau_{L_{ie}}} \right)^{\alpha_{L_i}} \right]^{\frac{1}{\frac{\sigma_i}{\sigma_i - 1} - \alpha_i}}}. \tag{8}$$

[1] Formally seen by rewriting (7) to eliminate output term Y_{ie}.

For instance, if the labor distortion faced by the establishment $(1+\tau_{L_{ie}})$ increases relative to the average labor distortion in the industry $\overline{(1+\tau_{L_i})}$, the establishment declines in size.

We can also see from equation (8) that the size of the establishment after we reallocate resources will depend solely on its productivity A_{ie}. From the earlier first-order conditions, we know that equalizing marginal products is akin to equalizing distortions. The reallocation of resources would then eliminate the relative distortions in equation (8), and the counterfactual size of the $s_{ie}|_{\tau=\bar{\tau}}$ would be strictly increasing in productivity A_{ie}.

Misallocation

By combining the model aggregation with the establishment responses to distortions, we follow the literature and measure misallocation as the distance between aggregate productivity and its frontier. At this frontier, all establishments in the industry have the same marginal revenue products. The more that actual productivity lags from its frontier, the larger is the measure of misallocation. Formally, industry misallocation Φ_i in equation (9):

$$\Phi_i = \frac{TFP_i\big|_{\tau=\bar{\tau}}}{TFP_i} = \frac{\left[\sum_{e=1}^{N_i}\left(A_{ie}\times\Omega_{TFP,\tau=\bar{\tau},ie}\right)^{\sigma_i-1}\right]^{\frac{1}{\sigma_i-1}}}{\left[\sum_{e=1}^{N_i}\left(A_{ie}\times\Omega_{TFP,ie}\right)^{\sigma_i-1}\right]^{\frac{1}{\sigma_i-1}}}, \tag{9}$$

captures the distance between actual industry total factor productivity TFP_i and its frontier where distortions, and hence marginal revenue products, are equalized across establishments $TFP_i|_{\tau=\bar{\tau}}$. Since industry output is produced using a CES technology, as per equation (2), the industry total factor productivity TFP_i is also a CES aggregate of establishment productivity A_{ie}. The scaling factor $\Omega_{TFP,ie}$—which we discuss below in more detail—captures the extent to which each establishment shapes industry productivity. When we reallocate resources to equalize marginal revenue products, each establishment's scaling factor changes from $\Omega_{TFP,ie}$ to $\Omega_{TFP,\tau=\bar{\tau},ie}$. We now provide some intuition about this change in scaling parameters and then define them in terms of model objects.

Since a highly distorted establishment becomes more integral to industry productivity when its distortions are removed, the extent of misallocation depends on which establishments bear the greatest distortions. If the most productive establishments also bear the largest distortions, we measure more misallocation than if less productive establishments bear the same distortions. In short, the correlation between productivity and distortion shapes the extent of misallocation, a notion first emphasized by Restuccia and Rogerson

(2008).[2] In our model, this notion relies on the claim that the scaling factor $\Omega_{TFP,\tau=\bar{\tau},ie}$ increases more relative to the scaling factor $\Omega_{TFP,ie}$ when an establishment is highly distorted. We substantiate this claim below after relating the scaling factors to model objects.

The scaling factors are based on establishments' *revenue* productivity $TFPR_{ie}$, which summarizes the impact of distortions on the establishments. As in Foster et al. (2008), $TFPR_{ie}$ measures an establishment's ability to generate revenue per input bundle:

$$TFPR_{ie} = \frac{P_{ie}Y_{ie}}{K_{ie}^{\alpha_{K_i}} L_{ie}^{\alpha_{L_i}}} = P_{ie}A_{ie}. \tag{10}$$

Equation (10) highlights the implication that, when comparing two establishments with the same physical productivity A_{ie}, a higher revenue productivity $TFPR_{ie}$ reflects a higher price. As we showed earlier, a higher price reflects larger distortions.

As the model focuses on the allocation of resources across establishments, the scaling factors compare the average revenue productivity of the industry, $\overline{TFPR_i}$, with the revenue productivity of an establishment, $TFPR_{ie}$. Equation (11) shows that this relative revenue productivity depends on the size of the establishment and the relative distortions that it faces. In a comparison of two equally productive establishments, the more distorted establishment would have a smaller $\overline{TFPR_i}/TFPR_{ie}$ ratio. Equation (12) shows that the relative revenue productivity after equalizing marginal products is a function of the post-reallocation size of the establishment.

$$\Omega_{TFP,ie} = \frac{\overline{TFPR_i}}{TFPR_{ie}} = \left(\frac{P_{ie}Y_{ie}}{P_iY_i}\right)^{\alpha_i - 1} \left(\frac{\overline{1+\tau_{K,i}}}{1+\tau_{K_{ie}}}\right)^{\alpha_{K_i}} \left(\frac{\overline{1+\tau_{L,i}}}{1+\tau_{L_{ie}}}\right)^{\alpha_{L_i}} \tag{11}$$

$$\Omega_{TFP,\tau=\bar{\tau},ie} = \frac{\overline{TFPR_i}}{TFPR_{ie}}\bigg|_{\tau=\bar{\tau}} = \left(\frac{P_{ie}Y_{ie}}{P_iY_i}\bigg|_{\tau=\bar{\tau}}\right)^{\alpha_i - 1} = \left(\frac{\left[A_{ie}\right]^{\frac{1}{\frac{\sigma_i}{\sigma_i-1}-\alpha_i}}}{\sum_{e=1}^{N_i}\left[A_{ie}\right]^{\frac{1}{\frac{\sigma_i}{\sigma_i-1}-\alpha_i}}}\right)^{\alpha_i - 1} \tag{12}$$

Before formally characterizing how the scaling factors in equations (11) and (12) differ from each other, we want to emphasize how they are shaped by variations in markups and returns to scale. First, deviations from constant returns to scale (i.e. $\alpha_i \neq 1$) imply that the size of the establishment affects its revenue productivity. By contrast, in the Hsieh-

[2] Hopenhayn (2014) makes clear that a discussion of correlations in this setting requires the comparison of the same proportional distortion. In his summary and re-framing of the literature, correlations matter because the same proportional distortion $\tau_{L_{ie}}$ would displace more labor at a more productive establishment.

Klenow model, returns to scale are constant and the size term drops out of the scaling factors; for instance, the counterfactual TFPR ratio in equation (12) is then unity for all establishments, regardless of industry. Second, the difference between the markup $\sigma_i/(\sigma_i - 1)$ and the returns to scale α_i shapes the counterfactual size of the establishment in (12). In our model, two industries could be populated by equally productive establishments, and yet different wedges between markups and returns to scale would lead the industries to differ in their counterfactual size distributions. Under the Hsieh-Klenow assumptions, the counterfactual size distribution would be the same in both industries. We examine the impact of these types of differences on misallocation in greater detail in section 2.2.

Returning now to the measure of misallocation, we show that, when rid of its distortions, a more distorted establishment becomes more integral to industry productivity. In equation (13) we isolate the establishment-specific components of the relative scaling factors:

$$\frac{\Omega_{TFP,\tau=\bar{\tau},ie}}{\Omega_{TFP,ie}} \propto \left[\left(\frac{1+\tau_{K_{ie}}}{1+\tau_{K,i}} \right)^{\alpha_{K_i}} \left(\frac{1+\tau_{L_{ie}}}{1+\tau_{L,i}} \right)^{\alpha_{L_i}} \right]^{\frac{\frac{\sigma_i}{\sigma_i-1}-1}{\frac{\sigma_i}{\sigma_i-1}-\alpha_i}}. \tag{13}$$

Since establishment productivity A_{ie} enters both scaling factors in the same manner, the only establishment-specific difference between the two comes from distortions. Note that the exponent on the distortions in (13) is positive, so that the derivative of $\Omega_{TFP,\tau=\bar{\tau},ie}/\Omega_{TFP,ie}$ with respect to the distortions is positive. In other words, the relative increase in scaling factor $\Omega_{TFP,\tau=\bar{\tau},ie}$ is greater for a more distorted establishment.

Having defined all elements of industry-level misallocation, we use the model structure to express the economy-wide misallocation Φ as the geometric average of the industry measures Φ_i, as per equation (14):

$$\Phi = \prod_{i \in I} \Phi_i^{\theta_i}. \tag{14}$$

Misallocation here captures the aggregate productivity loss from distortions faced by establishments within industries.

While this measure is standard within the literature, its construction implicitly relies on some additional assumptions. For instance, by focusing on equalizing distortions within industries, we leave average distortions unchanged across industries. This assumption overlooks the potential productivity improvement from reallocating resources across industries. Moreover, this measure of misallocation assumes no changes in entry and exit of establishments when we alter distortions. Another potential concern might be the absence

of taste (i.e., demand) shocks from the benchmark model. For that particular case, we show in appendix D that the measure of misallocation is unchanged for a simple extension where we allow establishment-specific taste parameters. In short, Φ is a counterfactual that holds all non-distortion parameters—including tastes—fixed; the measure of misallocation above would correctly capture productivity losses even in that extended model.

2.2 Ignoring the Variation in Returns to Scale and Markups

In this section, we show that inappropriately imposing constant returns to scale and a common markup leads to incorrect measures of productivity and misallocation. Imposing constant returns to scale when returns to scale are decreasing, or understating the markup of price over marginal cost, leads us to measure more distorted establishments as more productive. This spurious positive correlation between productivity and distortion leads us to overstate misallocation. We use the expressions we derive in this section to help us explain in section 4 why and when the divergent trends in misallocation arise.

The discrepancies we highlight arise from inappropriate mappings from the observable establishment revenue to the unobservable establishment productivity. As we emphasized in the introduction, mapping from revenue to productivity is a two-step process: first we map revenue to output with the help of a pricing model, and then we map output to productivity with the help of a production function. We begin to formalize this notion by combining the demand for establishment output with the establishment production function, and derive the expression for establishment productivity A_{ie} in equation (15):

$$\ln A_{ie} = \frac{\sigma_i}{\sigma_i - 1} \ln \left(\frac{P_{ie} Y_{ie}}{P_i Y_i} \right) - \alpha_i \ln \left[K_{ie}^{\frac{\alpha_{K_i}}{\alpha_i}} L_{ie}^{\frac{\alpha_{L_i}}{\alpha_i}} \right] + \ln Y_i. \tag{15}$$

This expression clarifies the first mapping by showing the markup $\sigma_i/(\sigma_i - 1)$ as the elasticity of productivity with respect to the revenue-based measure of size $P_{ie}Y_{ie}/(P_iY_i)$. Furthermore, returns to scale in production α_i highlight the second mapping, as α_i is the elasticity of productivity with respect to the input bundle under the assumption of constant returns to scale $K_{ie}^{\alpha_{K_i}/\alpha_i} L_{ie}^{\alpha_{L_i}/\alpha_i}$. We now explore the discrepancies in measures of productivity and misallocation from imposing constant returns to scale and a common markup.

Discrepancy from Imposing Constant Returns to Scale

To measure total factor productivity A_{ie}, we need to impose a production function on the data; as suggested by equation (15), if we mismeasure the returns to scale in production, we incorrectly measure productivity. We formalize this notion in equation (16) by compar-

ing the productivity $\widehat{A_{ie}}$ measured under constant returns to scale to the productivity A_{ie} measured under returns to scale α_i:

$$\left(\frac{\widehat{A_{ie}}}{A_{ie}}\right)_{CRTS\ Discrepancy} = \left(K_{ie}^{\frac{\alpha_{K_i}}{\alpha_i}} L_{ie}^{\frac{\alpha_{L_i}}{\alpha_i}}\right)^{\alpha_i - 1}. \tag{16}$$

For example, if we impose constant returns to scale on an industry where returns to scale are decreasing, then the exponent on the input bundle in equation (16) is negative. As a result, if we compare two equally productive establishments in this decreasing returns to scale industry, then the more distorted establishment with the smaller input bundle would be perceived as more productive. The discrepancy works in the opposite direction when returns to scale are increasing: more distorted establishments with smaller input bundles appear less productive than they are.

These discrepancies in measured productivity lead us to discrepancies in measured misallocation. In equation (17) we compare the misallocation $\widehat{\Phi}_i$ derived under the imposition of constant returns to scale with the misallocation Φ_i derived under the returns to scale α_i:

$$\left(\frac{\widehat{\Phi}_i}{\Phi_i}\right)_{CRTS\ Discrepancy} = \frac{\left[\sum_{e=1}^{N_i}\left(A_{ie}\frac{\overline{TFPR}_i}{TFPR_{ie}}\Big|_{\tau=\bar{\tau}}\ \Xi_{crts,ie}^{\,1-\alpha_i}\right)^{\sigma_i-1}\right]^{\frac{1}{\sigma_i-1}}}{\left[\sum_{e=1}^{N_i}\left(A_{ie}\frac{\overline{TFPR}_i}{TFPR_{ie}}\Big|_{\tau=\bar{\tau}}\right)^{\sigma_i-1}\right]^{\frac{1}{\sigma_i-1}}}, \tag{17}$$

$$\textit{where } \Xi_{crts,ie} = \left\{\left[\frac{1+\tau_{K_{ie}}}{1+\tau_{K,i}}\right]^{\frac{\alpha_{K_i}}{\alpha_i}}\left[\frac{1+\tau_{L_{ie}}}{1+\tau_{L,i}}\right]^{\frac{\alpha_{L_i}}{\alpha_i}}\frac{s_{ie}|_{\tau=\bar{\tau}}}{s_{ie}}\right\}.$$

When returns to scale are constant so that $\alpha_i = 1$, then the exponent on the establishment-specific scaling factor $\Xi_{crts,ie}$ is 0, and the ratio in (17) collapses to 1: the two measures of misallocation are identical. However, deviations from constant returns to scale lead to incorrect measures of misallocation.

The size of the discrepancy in misallocation depends on the extent to which returns to scale are not constant, and on the correlation between productivity and distortion. Note that the scaling factor $\Xi_{crts,ie}$ takes values above 1 for heavily distorted establishments; each of the three ratios defining the scaling factor exceeds 1 for a heavily distorted establishment. Under decreasing returns to scale, the positive exponent on $\Xi_{crts,ie}$ puts larger weights on the distorted establishments. If productivity and distortions are positively correlated, then the numerator in (17) exceeds the denominator, and we overstate misalloca-

tion. For the same positive correlation of productivity and distortion, an industry in which returns to scale are increasing would induce a negative exponent on $\Xi_{crts,ie}$ and lead us to understate misallocation if we inappropriately impose constant returns. After estimating returns to scale, we use these expressions to understand how imposing constant returns leads the Hsieh-Klenow measure of misallocation to deviate from our measure.

Discrepancy from Imposing a Common Markup

We also need the markup so as to map establishment revenue to establishment productivity; as hinted by equation (15), an incorrect markup leads to incorrect measures of output and productivity. We formalize this notion in equation (18) where we compare the productivity $\widehat{A_{ie}}$, measured under the markup generated by $\widehat{\sigma}_i$, with the productivity A_{ie}, measured under the true markup σ_i:

$$\left(\frac{\widehat{A_{ie}}}{A_{ie}}\right)_{Markup\ Discrepancy} = \left(\frac{P_{ie}Y_{ie}}{P_iY_i}\right)^{\frac{\widehat{\sigma}_i}{\widehat{\sigma}_i-1}-\frac{\sigma_i}{\sigma_i-1}}. \tag{18}$$

In short, imposing an incorrect markup leads to a measure of productivity that is a function of the establishment size $P_{ie}Y_{ie}/(P_iY_i)$. For instance, when the imposed markup overstates the true markup, then the exponent on establishment size is positive. Consequently, if we compare two equally productive establishments, then the more distorted establishment will be smaller in size, and would be incorrectly perceived as less productive. In this respect, overstating the markup induces similar discrepancies in measuring productivity as does understating the returns to scale in equation (16).

The imposition of an incorrect markup results in an incorrect measure of misallocation. To anticipate our subsequent decomposition, we formalize this notion under the assumption of constant returns to scale. In equation (19), we compare the misallocation $\widehat{\Phi}_i$ measured under the incorrect markup to the misallocation Φ_i measured under the true markup:

$$\left(\frac{\widehat{\Phi}_i}{\Phi_i}\right)_{Markup\ Discrepancy} = \left[\sum_{e=1}^{N_i} s_{ie}\big|_{\tau=\bar{\tau}}\ \Xi_{markup,ie}^{\ \frac{\widehat{\sigma}_i-\sigma_i}{\sigma_i-1}}\right]^{\frac{1}{\widehat{\sigma}_i-1}}, \tag{19}$$

$$\text{where } \Xi_{markup,ie} = \frac{s_{ie}\big|_{\tau=\bar{\tau}}}{s_{ie}}$$

If the markup is measured correctly, so that $\widehat{\sigma}_i = \sigma_i$, then the establishment-specific scaling factor $\Xi_{markup,ie}$ disappears; and, since the relative establishment sizes $s_{ie}|_{\tau=\bar{\tau}}$ sum to 1 by definition, there is no error in measuring misallocation. However, deviations from the

correct markup lead to discrepancies in measured misallocation.

The magnitude of the discrepancy in measured misallocation depends on the direction in which we mismeasure the markup, and the correlation of productivity and distortion. We note that the scaling factor $\Xi_{markup,ie}$ takes values greater than 1 for distorted establishments since distorted establishments grow larger in size when the distortions are removed. Consider a setting in which productivity and distortion are positively correlated. If we understate the markup, the scaling factor puts more weight on the large, productive establishments, and puts less weight on the small, unproductive establishments. This rescaling of establishment size makes the expression in equation (19) exceed 1, leading to a measure of misallocation that is too large. By contrast, overstating the markup makes the exponent on the scaling factor negative, reversing the impact of the scaling on the relative establishment sizes, and leading us to understate misallocation. Below we use these expressions to understand the forces that differentiate our measure of misallocation from the Hsieh-Klenow measure that imposes a common markup and constant returns to scale.

3 Mapping the Model to Data

In this section, we show how to map the available U.S. Census microdata to measure distortions and productivity in U.S. manufacturing. With data only on establishment revenue—not output or prices—we emphasize the need for an estimating equation that jointly estimates returns to scale and price markups. We show that the reduced-form elasticities from this estimating equation inform us about profit shares, and that the model can be used to translate these reduced-form elasticities into returns to scale and markups. We then provide estimates of returns to scale and markups that are consistent with the estimated profit shares.

3.1 Data

Our analysis relies on two core data sets from the U.S. Census Bureau: the Census of Manufactures (CMF) and the Annual Survey of Manufactures (ASM). The Census data sets provide us with the establishment-level variables from which we infer productivity and distortions. The CMF is conducted every five years (for years ending in 2 and 7) and contains information about all manufacturing establishments in the United States. The ASM is conducted in all non-Census years and covers establishments with at least 250 employees, as well as a randomly sampled panel of smaller establishments. On average, the ASM surveys 50,000–65,000 establishments selected from the approximately 350,000 establish-

ments in the CMF. From these datasets, we obtain measures of value added, hours worked, materials expenditures, capital stock, and the relevant price deflators. The industry price deflators come from the NBER-CES manufactuing database, and the capital stocks are constructed following Foster et al. (2016a). Our sample period spans 1977 through 2007. We exclude establishments whose information is imputed from administrative records, as well as those with missing information.

As industry classification in the U.S. changed during the sample period, we build off the concordance made by Fort and Klimek (2015) that assigns establishments a time-consistent NAICS (North American Industrial Classification System) 2002 code. For a small number of the 400+ 6-digit NAICS industries, we identify discontinuities in industry employment and establishment counts around the years where industry classification changed.[3] If the NAICS dictionaries suggest that the industries in question are cross-listed, we attempt to merge them into a single industry. When the merging eliminates discontinuities, we use the merged industries; otherwise, we exclude the industries from analysis. We also exclude industries that contain fewer than five establishments in any given year.

To construct more comprehensive industry measures of expenditures on labor, we supplement the Census data on salaries and wages with BLS measures of benefit payments. While the ASM and the CMF exhaustively cover many aspects of manufacturing establishments, the U.S. Census microdata on total labor compensation is much sparser; only direct payments to labor for services in production (i.e., salaries and wages) are widely documented. By contrast, for a smaller sample of establishments, the BLS-run National Compensation Survey collects data on wages, paid leave, insurance, retirement contributions, legally required benefits, and supplemental pay. From these data, the BLS constructed for us unpublished estimates of the hourly wage and the hourly total benefit cost. Using these data, we construct a BLS Adjustment with which we can adjust the Census industry labor payment to reflect payments to labor:

$$\text{BLS Adjustment}_{i,t} = \frac{\text{BLS hourly wage}_{i,t} + \text{BLS hourly benefits}_{i,t}}{\text{BLS hourly wage}_{i,t}}.$$

Given the survey size, to pass BLS disclosure review, our BLS Adjustment factors are constructed at the NAICS 3-digit level for five-year intervals spanning 1983–2007.[4]

[3] We construct mid-point growth rates, and flag growth rates of establishment counts or hours worked that exceed 0.5 in absolute value.

[4] We apply BLS Adjustment factors from 1983–1987 to the Census data in both 1987 and 1982.

3.2 Step 1: Measuring Distortions

To measure misallocation, we need to know the distortions faced by an establishment relative to the average distortions in the industry. We derive relative distortions by rearranging the first-order conditions from equations (5) and (6) and dividing by their weighted averages over all establishments in the industry. The resulting expressions, in equations (20) and (21), are independent of the returns to scale and markup parameters, which are common to all establishments in the industry, and map transparently to Census data:

$$\frac{1+\tau_{K_{ie}}}{\overline{1+\tau_{K,i}}} = \frac{\frac{P_{ie}Y_{ie}}{K_{ie}}}{\left[\sum_{e=1}^{N_i} \frac{P_{ie}Y_{ie}}{P_iY_i}\left(\frac{P_{ie}Y_{ie}}{K_{ie}}\right)^{-1}\right]^{-1}} = \frac{\frac{\text{Value Added}_{ie}}{\text{Capital Stock}_{ie}}}{\left[\sum_{e=1}^{N_i} \frac{\text{Value Added}_{ie}}{\text{Value Added}_i}\left(\frac{\text{Value Added}_{ie}}{\text{Capital Stock}_{ie}}\right)^{-1}\right]^{-1}} \tag{20}$$

$$\frac{1+\tau_{L_{ie}}}{\overline{1+\tau_{L,i}}} = \frac{\frac{P_{ie}Y_{ie}}{L_{ie}}}{\left[\sum_{e=1}^{N_i} \frac{P_{ie}Y_{ie}}{P_iY_i}\left(\frac{P_{ie}Y_{ie}}{L_{ie}}\right)^{-1}\right]^{-1}} = \frac{\frac{\text{Value Added}_{ie}}{\text{Labor Hours}_{ie}}}{\left[\sum_{e=1}^{N_i} \frac{\text{Value Added}_{ie}}{\text{Value Added}_i}\left(\frac{\text{Value Added}_{ie}}{\text{Labor Hours}_{ie}}\right)\right]^{-1}}. \tag{21}$$

The model interprets high revenue productivity in inputs as an indicator for the presence of distortions. In a world without distortions, this model suggests that all establishments hire inputs so as to equalize their average capital $P_{ie}Y_{ie}/K_{ie}$ and average labor $P_{ie}Y_{ie}/L_{ie}$ revenue productivities. If an establishment has a high revenue productivity in a certain input, it would maximize profits by continuing to hire that input until this measure of revenue productivity declined and equaled that of the other establishments in the industry. If an establishment in the data has a high average revenue productivity in a given input, it must have been prevented from hiring more of the input; hence, the model assigns this establishment a high distortion.

These strong assumptions identify distortions and reflect the model's attempt to describe a steady-state economy. In a dynamic setting, we can think of frictions that might prevent an establishment from hiring the steady-state profit-maximizing quantity of an input. Asker et al. (2014), for instance, focus on adjustment costs in the hiring of capital as one reason that an establishment's choice might deviate from these steady-state predictions. Nonetheless, for the purpose of measuring misallocation across longer periods of time, we think these assumptions are a reasonable starting point. To match this view of the model's purpose, our estimates of model parameters and misallocation are based on five-year periods; we also document the robustness of the main results in section 5 by extending this estimating window to ten years.

3.3 Step 2: Measuring Productivity

With data on establishment revenue, not output and prices separately, we cannot directly estimate the returns to scale and markup we need to infer productivity. Instead, the revenue elasticities from our estimating equation inform us about the division of value added among labor, capital, and profits. Nonetheless, using model equations we can indirectly map these reduced-form revenue elasticities into returns to scale and markups, and then infer establishment productivity.

A common approach to measuring returns to scale in data sets with establishment revenue entails creating a proxy for output by dividing revenue $P_{ie}Y_{ie}$ with an industry price index P_i; this common practice leads to a downward bias in estimated returns to scale that was first pointed out by Marschak and Andrews (1944) and later made particularly salient by Klette and Griliches (1996). Intuitively, this bias arises because we expect the most productive establishments to hire the largest input bundles, to produce the most output, and—when output markets are imperfectly competitive—to charge the lowest prices. If the most productive establishments charge the lowest prices, then the proxy for output is likely to understate output most for these productive establishments. A cross-sectional estimator using this output proxy would understate the increase in output from having the large input bundles, and hence underestimate returns to scale.[5]

The derivation of our estimating equation highlights this downward bias in returns-to-scale estimates. Specifically, we follow De Loecker (2011) and combine two model equations: the establishment's production function and the demand for its output. Rearranging this combined expression to solve for the ratio of revenue $P_{ie}Y_{ie}$ and the price index P_i, and taking logs, we derive the estimating equation (22):

$$\ln\left(\frac{P_{ie}Y_{ie}}{P_i}\right) = \beta_{K_i}\ln(K_{ie}) + \beta_{L_i}\ln(L_{ie}) + \beta_{Y_i}\ln(Y_i) + \beta_{A_i}\ln(A_{ie}), \tag{22}$$

$$\text{where } \beta_{K_i} = \frac{\alpha_{K_i}}{\frac{\sigma_i}{\sigma_i-1}},\ \beta_{L_i} = \frac{\alpha_{L_i}}{\frac{\sigma_i}{\sigma_i-1}},\ \beta_{Y_i} = \frac{1}{\sigma_i},\ \text{and } \beta_{A_i} = \frac{\sigma_i - 1}{\sigma_i}$$

$$\text{and } P_{ie}Y_{ie} = \text{Value Added}_{ie},\ K_{ie} = \text{Capital Stock}_{ie},\ L_{ie} = \text{Labor Hours},^{6}$$

$$P_i = \text{NBER-CES Industry Price Index}_i,\ P_iY_i = \sum_{e=1}^{N_i}\text{Value Added}_{ie},\ Y_i = \frac{P_iY_i}{P_i}.$$

The revenue elasticities $\beta_{i,L}$ and $\beta_{i,K}$ are quotients of the returns-to-scale parameters and

[5] That revenue elasticities are not synonymous with production-function parameters has also been prominently emphasized in work by Cooper and Haltiwanger (2006) and Foster et al. (2016b), among others.

[6] We compute total labor hours as the sum of the reported production-worker hours and the calculated non-production-worker hours following Kehrig (2011).

the markup of price over marginal cost. Since we expect establishments to price at or above marginal cost, the gross markup exceeds 1. As a result, even when correctly estimated, the revenue elasticities understate the returns-to-scale parameters.[7]

Although they do not directly estimate returns to scale, the revenue elasticities β_{K_i} and β_{L_i} are useful descriptors of differences across industries: they correspond to capital's and labor's share of value added and together imply an industry's profit share. Rearranging the first-order conditions from equations (5) and (6), and summing across establishments within an industry, we show in (23) that β_{K_i} and β_{L_i} are the distortion-inclusive expenditures on inputs as a share of value added:

$$\beta_{K_i} = \frac{\sum_{e=1}^{N_i}(1+\tau_{K_{ie}})R_iK_{ie}}{P_iY_i} \text{ and } \beta_{L_i} = \frac{\sum_{e=1}^{N_i}(1+\tau_{L_{ie}})w_iL_{ie}}{P_iY_i}. \tag{23}$$

In addition, we show in equation (24) that industry profits are the residual share of value added (i.e., the difference between one and the sum of the revenue elasticities):

$$\frac{\Pi_i}{P_iY_i} = 1 - (\beta_{K_i} + \beta_{L_i}). \tag{24}$$

Since we expect establishments to earn weakly positive profits, the expression in (24) emphasizes that the sum of revenue elasticities is bounded from above by 1 in this model. This is yet another way to see the bias emphasized by Klette and Griliches (1996): if this model correctly characterizes the world, and if we lived in a world with returns to scale α_i in excess of 1, the standard estimating equation would still produce revenue elasticities that sum to less than 1.

The third revenue elasticity β_{Y_i}, the elasticity of establishment revenue with respect to industry output, is key to identifying the returns to scale and markup parameters from the revenue elasticities β_{K_i} and β_{L_i}. Specifically, the inverse of β_{Y_i} is the elasticity of substitution σ_i, from which we can construct the markups $\sigma_i/(\sigma_i - 1)$. With the estimated markup we can then back out the returns to scale parameters α_{K_i} and α_{L_i} as the products of the markup and the respective revenue elasticities. With the parameters for the markup and the returns to scale in hand, we can infer productivity.

We estimate $\widehat{\beta_{L_i}}$, the first of the three key elasticities, using the rearranged first-order condition for labor in expression (23). We map this expression to the data by multiplying the sum of salaries and wages reported in the U.S. Census microdata by the BLS Adjustment

[7]This estimating equation can also be derived from a gross-output production function that is Leontief in an intermediate input whose price is proportional to the price of output, as in the Monte-Carlo experiments of Ackerberg et al. (2015).

factors we detailed in section 3.1. In this way, our measure of industry labor expenditures attempts to capture not only the wage payments to labor, but also the benefits and indirect payments, from insurance to retirement contributions, that are not widely reported to the Census. To estimate $\widehat{\beta_{L_i}}$, we divide this measure of labor costs by the industry value added:

$$\widehat{\beta_{L_i}} = \frac{\left[\sum_{e=1}^{N_i} \text{Salaries and Wages}_{ie}\right] \times \text{BLS Adjustment}_i}{\sum_{e=1}^{N_i} \text{Value Added}_{ie}}. \tag{25}$$

This $\widehat{\beta_{L_i}}$ estimate implicitly assumes that the labor distortions faced by establishments are priced into the labor costs reported by establishment while distortions that are not priced into reported labor costs net to zero within an industry. More formally, let us label by $\tau_{L,P,ie}$ distortions are be priced into reported labor costs and by $\tau_{L,U,ie}$ distortions that are not priced. An expanded version of equation (23) would then read as

$$\beta_{L_i} = \frac{\sum_{e=1}(1+\tau_{L,P,ie}+\tau_{L,U,ie})w_i L_{ie}}{P_i Y_i} = \underbrace{\frac{\sum_{e=1}(1+\tau_{L,P,ie})w_i L_{ie}}{P_i Y_i}}_{\text{Data}} + \frac{w_i L_i}{P_i Y_i}\sum_{e=1}(\tau_{L,U,ie})\frac{w_i L_{ie}}{w_i L_i},$$

where the data on the (BLS-adjusted) labor share represents the distortions that are priced in, and where we are additionally assume that appropriately-weighted unpriced distortions net out to zero.[8] As a robustness check in section 5, we also estimate this elasticity from the variation in labor usage across establishments. Even under these different assumptions required to thus estimate the elasticity, we find the path of U.S. manufacturing misallocation to look very different under the assumptions of our model and those of the Hsieh-Klenow model.

We estimate the remaining two elasticities β_{K_i} and β_{Y_i} using a two-step Generalized Methods of Moments (GMM) procedure based on the control-function approach in Levinsohn and Petrin (2003). This approach addresses the issue that productivity is unobserved in estimating equation (26) by substituting out the unobserved productivity with a function of observable variables. The choice to estimate the labor elasticity β_{L_i} in an earlier step is driven by the Ackerberg et al. (2015) critique that highlights the inability of the control-function procedure to identify the labor elasticity under standard assumptions. We

[8] Hsieh and Klenow (2009) assume that distortions are unpriced, which implies that reported labor expenditures represent undistorted $w_i L_{ie}$. This assumption would be inconsistent with using the reported labor share from the data to proxy for the model's labor share, which should be inclusive of distortions.

show later that our findings are robust to imposing additional assumptions and estimating all three elasticities jointly using the Ackerberg et al. (2015) estimator.

The control function we use is the choice of intermediate inputs, assumed to increase in establishment productivity: $ln(M_{ie}) = m(\ln K_{ie}, \ln Y_i, \ln A_{ie})$. If we can invert the expression characterizing this choice to express productivity as a function of the intermediate inputs, then we can substitute the unobservable A_{ie} in equation (22) with observables K_{ie}, M_{ie}, and Y_i as follows:

$$\underbrace{\ln\left(\frac{P_{ie}Y_{ie}}{P_i}\right) - \widehat{\beta_{L_i}}\ln(L_{ie})}_{pynet_{ie}} = \beta_{K_i}\ln(K_{ie}) + \beta_{Y_i}\ln(Y_i) + \beta_{A_i}m^{-1}(\ln K_{ie}, \ln Y_i, \ln M_{ie}) + u_{ie}, \quad (26)$$

where u_{ie} represents idiosyncratic shocks to production. For this substitution to be feasible and useful, we need to assume that the choice of intermediate inputs is invertible, and that productivity is the only unobservable component in the choice of intermediate inputs.[9] The first step of the procedure regresses the left-hand-side term of equation (26) $pynet_{ie}$ on a flexible polynomial of the observables to construct the predicted $\widehat{pynet_{ie}}$.

The second step of the procedure uses the assumption that log productivity $\ln A_{ie}$ evolves following a general first-order Markov process to construct moment conditions with which to estimate the elasticities β_{K_i} and β_{Y_i}. Specifically, we let $\varepsilon_{ie,t}$ correspond to the mean-zero innovations in productivity realized at time t. For a given guess $(\widehat{\beta_{K_i}}, \widehat{\beta_{Y_i}})$ of the elasticities, we construct an implied measure of log productivity by differencing $\widehat{pynet_{ie}}$ and $\widehat{\beta_{K_i}}\ln(K_{ie}) + \widehat{\beta_{Y_i}}\ln(Y_i)$. Regressing the implied productivity on a polynomial of its past value gives us the implied innovation to productivity $\varepsilon_{ie,t}(\widehat{\beta_{K_i}}, \widehat{\beta_{Y_i}})$, and the following moment conditions with which to estimate the two key elasticites:

$$\frac{1}{N}\frac{1}{T}\sum_{e\in N_i}\sum_{t\in T}\begin{pmatrix}\widehat{\varepsilon_{ie,t}}(\widehat{\beta_{K_i}}, \widehat{\beta_{Y_i}})\ln K_{ie} \\ \widehat{\varepsilon_{ie,t}}(\widehat{\beta_{K_i}}, \widehat{\beta_{Y_i}})\ln Y_i\end{pmatrix} = 0. \quad (27)$$

To estimate the elasticities in a model-consistent way, we constrain the parameter space to meet three criteria. First, to ensure that industry profits are weakly positive and less than 1 as a share of value added, we impose that $\widehat{\beta_{K_i}}$ and $\widehat{\beta_{L_i}}$ sum to a value between 0 and 1. Second, to estimate labor and capital shares of value that are strictly positive, we require that $\widehat{\beta_{K_i}}$ and $\widehat{\beta_{L_i}}$ are strictly positive. Third, to back out gross markups with values between 1 and 2, we impose that $\widehat{\beta_{Y_i}}$ be strictly positive and less than 0.5.[10]

[9] While common, these assumptions are strong and not directly testable. E.g., the second assumption eliminates the possibility that distortions in intermediate input markets are correlated with productivity.

[10] We think this is a reasonable parameter range as common choices for the elasticity σ_i range between 3 and 11, and imply markups between 1.1 and 1.5.

3.4 Division of Value Added in U.S. Manufacturing

By our estimates in panel A of table 1, labor's share of value added in U.S. manufacturing declined from 64% in 1982 to 39% in 2007; over the same period, the capital share increased from 20% to 25%. Together, these changes in the labor and capital shares imply that the profit share increased 20 percentage points, rising from 16% in 1982 to 36% in 2007. While, to our knowledge, this is the first paper to document these dynamics of industry profits for U.S. manufacturing, the findings are broadly consistent with other recent work. The decline of the labor share has been widely documented for the U.S. and for the global economy [e.g. Karabarbounis and Neiman (2014), Elsby et al. (2013), Barkai (2020)]. Moreover, using data on the U.S. non-financial corporate sector, Barkai (2020)

Table 1: U.S. Manufacturing – Division of Value Added

Panel A	Weighted Average across Industries		
	Capital Share β_{K_i}	Labor Share β_{L_i}	Profit Share $1-(\beta_{L_i}+\beta_{K_i})$
1982	0.20	0.64	0.16
1987	0.21	0.61	0.18
1992	0.27	0.55	0.18
1997	0.25	0.49	0.26
2002	0.31	0.46	0.23
2007	0.25	0.39	0.36
Panel B	Standard Deviation across Industries		
	Capital Share β_{K_i}	Labor Share β_{L_i}	Profit Share $1-(\beta_{L_i}+\beta_{K_i})$
1982	0.19	0.20	0.17
1987	0.20	0.19	0.20
1992	0.25	0.20	0.20
1997	0.23	0.18	0.24
2002	0.25	0.19	0.21
2007	0.26	0.19	0.30

Note: Reported values in panel A are weighted averages of industry-level coefficients, with the weights comprising industry value added. The underlying coefficients are estimated using five-year panels. Data for the estimation comes from the Annual Survey of Manufactures from the U.S. Census and the National Compensation Survey from the U.S. Bureau of Labor Statistics.

finds that both the labor and capital shares declined, leading to an increase in profits over the last 30 years. Complementary exercises in Karabarbounis and Neiman (2014) also suggest that the capital share increased insufficiently to offset the decline in the labor share, implying that profits increased.

In addition to documenting the evolution of these shares across time, we document in panel B large variations in capital, labor, and profit shares across industries. At all points in time, the standard deviation of profit shares across industries is roughly as large as the average level of the profit shares. The standard deviations of capital and labor shares are of quantitatively similar magnitudes. These large standard deviations imply that the U.S. manufacturing sector is populated both by industries where profit margins are slim, as well as by industries in which establishments earn large profits as shares of value added.[11]

3.5 Returns to Scale and Markups in U.S. Manufacturing

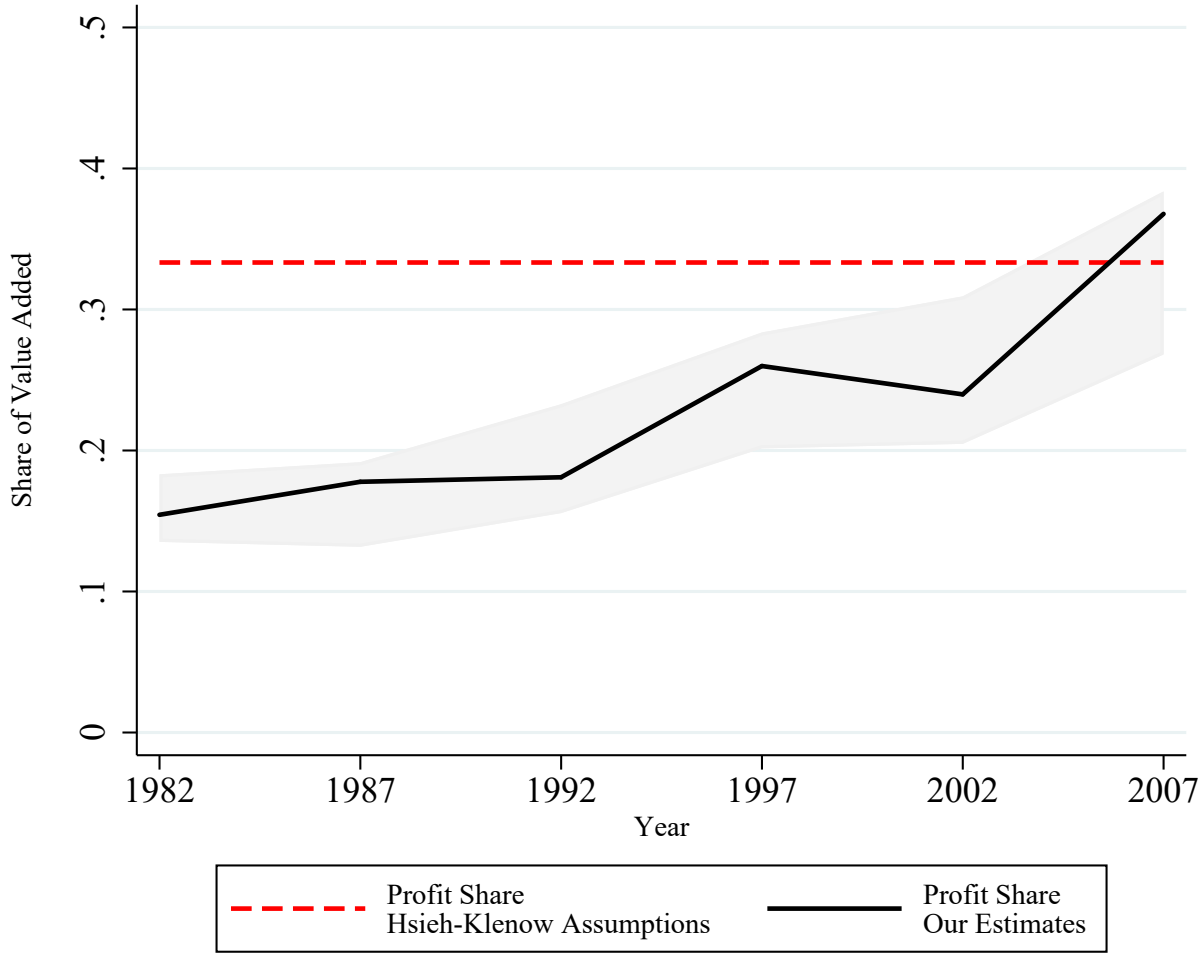
Accommodating these estimates of the profit shares requires deviating from the standard assumptions of constant returns to scale and a common markup. From Basu and Fernald (1997) we know that profits drive a wedge between markups and returns to scale under very general assumptions on the functional forms for production and demand. In our model, this relationship takes the following form:

$$1 - \frac{\Pi_i}{P_i Y_i} = \frac{\alpha_i}{\frac{\sigma_i}{\sigma_i - 1}}, \tag{28}$$

where the industry profit shares $\Pi_i/(P_i Y_i)$ act as a wedge between the returns to scale α_i and markup $\sigma_i/(\sigma_i - 1)$. By imposing constant returns to scale and a markup of 1.5 in every industry, the Hsieh-Klenow model implies that all establishments in all industries earn a third of their value added as profits. We emphasize this point in figure 2. The solid black line plots our estimated share of profits in value added. This rising measure of profits contrasts with the invariance of profit shares in the Hsieh-Klenow model, plotted as the dashed red line. The average profit share in 2007 of 0.36 roughly matches the Hsieh-Klenow assumptions. However, the variation across industries and the smaller profit shares throughout the 1980s and 1990s fit these assumptions less well.

[11] The literature on the decline of the labor share has also been complemented with evidence of increasing concentration of output, e.g., Autor et al. (2020). In Appendix C we show that in a whole class of monopolistic-competition models increases in profits—either from markups or from returns to scale–put downward pressure on variance and concentration of market shares. In this class of models, rationalizing increasing concentration requires a greater dispersion of productivity in addition to a change in profits.

Figure 2: Profits as a Share of Value Added in U.S. Manufacturing



Note: Shaded area encloses 5-95% bootstrapped confidence interval.

To understand why markups and returns to scale can rationalize these variations in profit shares, we focus on the fact that an establishment earns profits when its price exceeds the average cost of production:

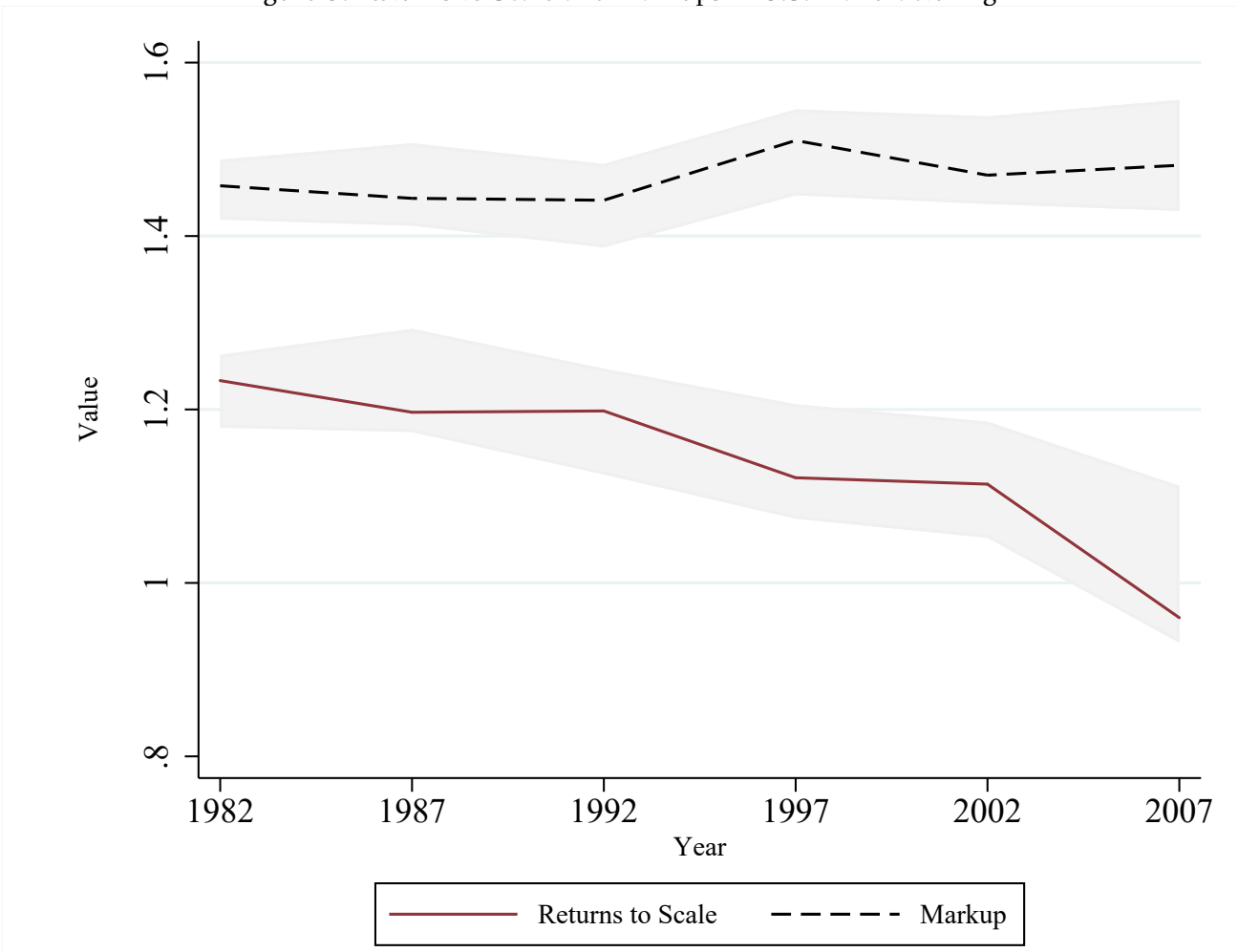
$$\frac{\pi_{ie}}{Y_{ie}} = \underbrace{\frac{\sigma_i}{\sigma_i - 1}\text{Marginal Cost}_{ie}}_{\text{Price}_{ie}} - \text{Average Cost}_{ie}. \tag{29}$$

The profits per unit sold, as per equation (29), can increase either if the markup increases or if the returns to scale decline. First, an establishment could increase its profit margin by charging a higher markup over marginal cost. Second, an establishment could increase its profit margin if average cost falls relative to marginal cost. A reduction in returns to scale drives such a shift in costs. For example, constant returns imply a constant marginal cost, while decreasing returns imply a marginal cost that increases with each unit produced. As

a result, if returns to scale decline from constant to decreasing, the marginal cost for the last unit would exceed the average cost of all units produced, increasing the profit margin. Some combination of an increase in markups and a reduction in returns to scale drives the increase in profit shares in the data.

In figure 3 (and in table 2), we show that, while markups increased from 1.46 to 1.48, the decline in returns to scale from 1.23 to 0.96 is the primary driver of rising profit shares between 1982 and 2007. In short, the U.S. manufacturing sector exhibited meaningfully increasing returns to scale in the early 1980s.[12] Since then, returns to scale have declined, driving up marginal cost relative to the average cost of production. By increasing the profit margin on each unit sold, this decline in returns to scale led to the rise in profit shares for U.S. manufacturing.

Figure 3: Returns to Scale and Markups in U.S. Manufacturing



Note: Shaded area encloses 5-95% bootstrapped confidence intervals. The reported values are weighted averages across industries.

[12]Using industry data for 1959 to 1980, Basu and Fernald (1997) also find evidence in their Table 2 of increasing returns to scale (γ^V) for manufacturing, and in particular for durable goods.

Table 2: U.S. Manufacturing – Returns to Scale and Markups

Panel A	Average Level across Industries	
	Returns to Scale α_i	Markups $\frac{\sigma_i}{\sigma_i - 1}$
1982	1.23	1.46
1987	1.20	1.44
1992	1.20	1.44
1997	1.12	1.51
2002	1.11	1.47
2007	0.96	1.48

Panel B	Standard Deviation across Industries	
	Returns to Scale α_i	Markups $\frac{\sigma_i}{\sigma_i - 1}$
1982	0.42	0.41
1987	0.46	0.40
1992	0.48	0.41
1997	0.49	0.43
2002	0.44	0.42
2007	0.58	0.42

Note: Reported values in Panel A are weighted averages of industry-level coefficients, with the weights comprising industry value added. Data for the estimation comes from the Annual Survey of Manufactures from the U.S. Census, and the National Compensation Survey from the U.S. Bureau of Labor Statistics.

Much like profit shares, both returns to scale and markups vary widely across industries. The standard deviations of both measures range between one third and one half the average values of their respective variables. For returns to scale, this variation suggests that, even as returns to scale have declined on average, the U.S. manufacturing sector is still comprised of both increasing and decreasing returns-to-scale industries. Similarly, while the average markup may be large, there are many industries with markups low enough to approximate perfect competition, as well as many industries where the degree of imperfect competition, and hence the markup, is large.

While table 1 showed how changing capital and labor shares drive the evolution of profits, here we show how the same evolution can be understood in terms of changing

returns to scale and markups:

$$\Delta \frac{\Pi_i}{P_i Y_i} = \underbrace{-\Delta\alpha_{K_i} \frac{1}{\left(\frac{\sigma_i}{\sigma_i - 1}\big|_{2007}\right)} - \Delta\alpha_{L_i} \frac{1}{\left(\frac{\sigma_i}{\sigma_i - 1}\big|_{2007}\right)}}_{\text{Contribution of Returns to Scale}} + \underbrace{\left(\Delta \frac{\sigma_i}{\sigma_i - 1}\right) \frac{1 - \left(\frac{\Pi_i}{P_i Y_i}\big|_{1982}\right)}{\left(\frac{\sigma_i}{\sigma_i - 1}\big|_{2007}\right)}}_{\text{Contribution of Markup}}. \tag{30}$$

An increase in an industry's profit share between 2007 and 1982, $\Delta\Pi_i/(P_i Y_i)$, is driven either by a decline in returns to scale $-\Delta\alpha_i$ or an increase in the markup $\Delta\sigma_i/(\sigma_i - 1)$, as per equation (30). Applying this decomposition to the manufacturing-sector data in tables 1 and 2, we show that of the 20-percentage-point increase in the manufacturing profit share, 18 percentage points come from the decline in returns to scale and 1–2 percentage points from the rise in the markup.[13] We can further decompose the 18 percentage points to emphasize separately the contributions of the capital and labor elasticities, α_{K_i} and α_{L_i}. The increase in the capital elasticity, reflected principally by the rising capital share, put downward pressure on the profit share of about –6 percentage points. Meanwhile, the sharp decline in the labor elasticity, reflected in the falling labor share, contributed 24 percentage points to the increase in the manufacturing profit share.

This finding that a change in returns to scale is an important driver of changing profits stands in contrast to recent work summarized by Basu (2019) that emphasizes sharp increases in markups. Both our approach and the markup-emphasizing approach epitomized by De Loecker et al. (2020) share the same common idea: changes in factor shares can be understood as changes in either markups or in returns to scale. Our studies differ on a number of data driven dimensions, from focusing on establishments versus focusing on firms, to using U.S. Census versus accounting measures of inputs. Yet, the largest difference is methodological and focused on the estimation of output elasticities and markups.

We leverage the idea that revenue elasticities contain information on both output elasticities and markups, while the approach of De Loecker et al. (2020) uses revenue elasticities as proxies of output elasticities. Specifically, when looking at publicly listed firms in COMPUSTAT, their approach infers a markup as the residual of a firm's factor share that is not explained by an estimated revenue elasticity; this revenue elasticity is used in place of the theoretically-desired output elasticity. This approach would not identify the markup in our model. We show in equation (23) that our revenue elasticities are equal to our factor

[13] A Jensen's inequality term leads to the small discrepancy. The manufacturing profit share in table 1 is the weighted average of industry profit shares, which is equal to $1 - \sum_{i \in I} \theta_i \frac{\alpha_i}{\frac{\sigma_i}{\sigma_i - 1}}$. Meanwhile, the average returns to scale and markup reported in table 2 are also weighted averages and do not imply exactly the same manufacturing profit share $1 - \left(\sum_{i \in I} \theta_i \alpha_i\right) / \left(\sum_{i \in I} \theta_i \frac{\sigma_i}{\sigma_i - 1}\right)$.

shares. Comparing a revenue elasticity and a factor share reveals no information about the markup. Bond et al. (Forthcoming) emphasize this idea in a less parametric setting: they show that if a revenue elasticity is used in place of an output elasticity, then the inferred residual from a factor share contains no information about the markup.

Even when De Loecker et al. (2020) use the same U.S. Census data on the manufacturing sector as we do here, our approaches differ in how we estimate output elasticities. When using Census data, they estimate their output elasticities as cost-shares. Since cost shares sum to one, this approach is tantamount to imposing constant returns to scale at every point in time. Imposing constant returns to scale over time will overstate the importance of markups: when returns to scale are assumed to be constant and time invariant, the estimated markups will mechanically reflect all variation in profits over time.

3.6 Returns to Scale in Context

In this section we provide evidence of changes within U.S. manufacturing that are consistent with a reduction in returns to scale. As we emphasized earlier, a given set of profits can be rationalized by either markups or returns to scale. The difficulty in disentangling markups from returns to scale, especially in data where we observe only revenue and not price and quantity separately, motivates our earlier emphasis on estimating these paratemeters jointly. We now discuss a particularly common—although not exclusive—interpretation of non-constant returns and, through it, we rationalize the estimated reduction in returns to scale.

A common motivation for increasing returns is the presence of fixed costs (of building an establishment, of building an assembly line, etc.,). Viewed through this lens, our estimates are consistent with output growing faster than the fixed costs establishments face. Namely, a decline in returns to scale of the sort we estimated—going from increasing to nearly constant—reflects a narrowing gap between average and marginal cost. Output growing faster than fixed costs implies precisely such a narrowing: fixed costs are spread over more units of output, bringing average cost closer to marginal cost.

Between 1977 and 2007 the U.S. manufacturing sector more than doubled in size: value added increased 132% and gross output increased 104%, both in real terms. For returns to scale *not* to have declined, fixed costs would have had to grow at least as much as output. Fixed costs could have grown in two ways. First, the number of establishments could have increased proportionally with output. In this case, even if fixed costs within establishments remained unchanged, the multiplication of establishments would have driven up fixed costs. Second, fixed costs within the average establishment could

have grown, perhaps as establishments physically grew in size, installed more assembly lines, or incurred other fixed costs of operating.

Although the construction of new establishments generates perhaps the largest fixed cost of production, this margin did not contribute meaningfully to the rise of fixed costs; while the total output of the manufacturing sector more than doubled between 1977 and 2007, the number of manufacturing establishments grew only 2% over this period.[14] With the near-constant number of establishments, the growth in output becomes an upper bound on the growth of internal fixed costs.

As an accounting of within-establishment fixed costs is infeasible with administrative Census data, we present a case study of automotive plants to argue that it is unlikely that fixed costs kept pace with the growth of output over this period. Much like the broader manufacturing sector, the U.S. automotive sector grew in terms of output but not in terms of the number of assembly plants. In real terms, value added for NAICS code 336111, automotive manufacturing, increased 50% between 1985 and 2007. Over the same period, data from WardsIntelligence shows that the number of automotive plants did not keep pace with the rise in output; the number of assembly plants actually declined from 76 to 68.[15]

Automotive plants did not change meaningfully in terms of first-order sources of within-establishment fixed costs: the number of production platforms, the number of vehicle series produced per platform, or in terms of the land area covered by the plants. The largest cost for an automotive plant is in setting up a platform, which is a common design and engineering base from which to produce vehicles with potentially different exteriors (e.g., Ford uses the same platform to produce the F-series trucks and the Expedition sports utility vehicle). The average plant has and continues to specialize in one vehicle platform: in 1985 the average plant had 1.24 platforms while in 2007 the average was 1.41. Another important fixed cost is the modification of platforms to produce multiple vehicle series. The number of different vehicle series per platform did not increase; it actually declined from 1.83 to 1.74. To look at the land area covered by plants we use *The Harbour Report*, published in 1995 and 2007, which focuses on the Big Three automakers (Chrysler, Ford, and General Motors). According to the report, the average assembly-plant floor increased roughly 13% over the period in question, substantially less than the increase in output.

While a look inside a single industry is not dispositive about the entire manufacturing sector, the case study illustrates the challenges of identifying fixed costs of production that could have overturned our estimated shift from increasing to near-constant returns to

[14]Furthermore, we find that industries that experienced larger declines in returns to scale over a five-year period also experienced slower growth in the number of establishments over the same period.

[15]Wards is one of the premier automotive industry publications. In addition to receiving sales data from all auto manufacturers in the United States, Wards maintains detailed data on the automotive plants themselves.

scale. When looking for evidence on fixed costs outside the case study, the type of capital employed by the manufacturing sector could be indicative of what is happening to fixed costs associated with production. To that effect, the measurement of physical capital is often split into equipment and structures, where structures are more likely to represent investments that are more intensive in terms of fixed costs. Yet, the share of structures in capital for the U.S. manufacturing sector has declined from 46% in 1977 to 31% in 2007, suggesting that finding a sharp rise in fixed costs could be challenge.

The challenge of identifying sharp increases in fixed costs is also reinforced by other findings in the literature that point to firms increasingly spreading overhead costs across establishments. In that spirit, Aghion et al. (2019) argue that improvements in Information Technology during the 1990s likely lowered the overhead costs of managing multiple product lines, while Fort et al. (2018) show that many manufacturing firms have grown primarily by acquiring non-manufacturing establishments; both arguments are consistent with the spreading of overhead costs across more establishments within a firm. In light of the rather sedate growth of these enumerated first-order fixed costs, the general pattern we identify could be overturned only if the unenumerated fixed costs (e.g., production overhead) increased at rates far greater than the rate of output growth.[16] We next turn to misallocation and emphasize the importance of incorporating the documented variation in markups and returns to scale.

4 Misallocation

In this section, we present our measure of misallocation and contrast it with the Hsieh-Klenow measure that ignores variation in markups and returns to scale. We then decompose the discrepancy in measurement and show that the divergent trends in misallocation are driven by the decline in returns to scale over time. Lastly, we relate changes in misallocation to changes in business dynamism.

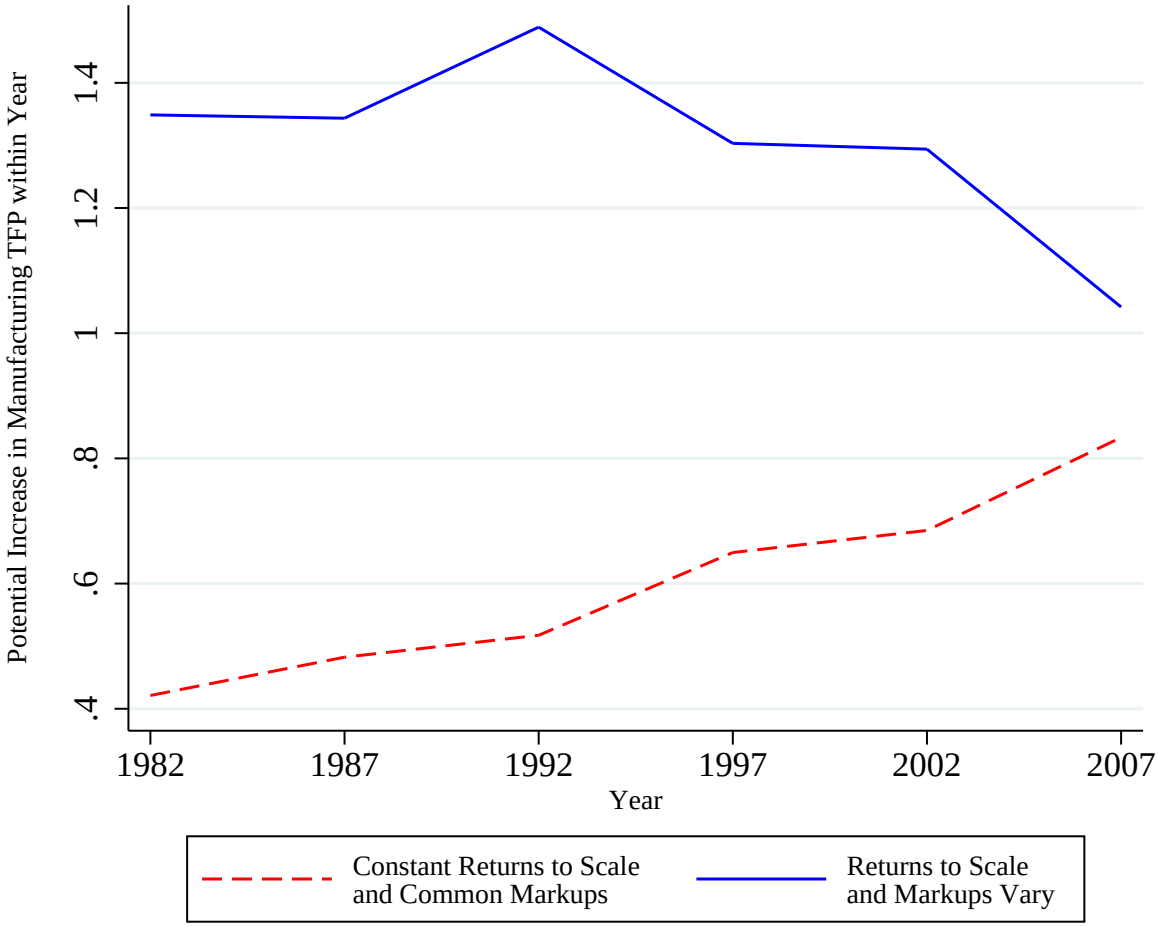
4.1 Misallocation Has Not Been Increasing

Our estimates suggest that misallocation in U.S. manufacturing decreased over the last 30 years. Figure 4 quantifies misallocation as the potential increases in U.S. manufacturing

[16]Production overheads are also challenging to identify in the data. For instance, data on Selling, General and Administrative Expenses (SGA) from publicly-listed firms in COMPUSTAT is occasionally taken as informative about overhead costs. The correlation of log firm size and log SGA is 0.9 for the period 1977-2007. This almost-log-linear relationship is challenging to reconcile with standard framings of overhead costs where they are assumed to be the same across firms (e.g., Bartelsman et al. (2013)). Moreover, even if we were to overcome conceptual questions about whether SGA measures production overhead in an accurate manner, these costs would have had to grow much faster than output to compensate for the slower growth of the previously-discussed fixed costs.

Figure 4: Misallocation in U.S. Manufacturing

Change in U.S. Manufacturing TFP from Equalizing Within-Industry Distortions



TFP from equalizing the distortions establishments face within an industry, as per equation (9). The solid blue line depicts our model while the dashed red line depicts the Hsieh-Klenow model. By our estimates, the level of misallocation declined from 135% in 1982 to 104% in 2007. Meanwhile, misallocation increased under the Hsieh-Klenow assumptions, so that in 2007 the U.S. manufacturing sector could have been 83% more productive, nearly twice the potential increase of 42% in 1982. Figure 1 presented the same results expressed as changes relative to 1982.

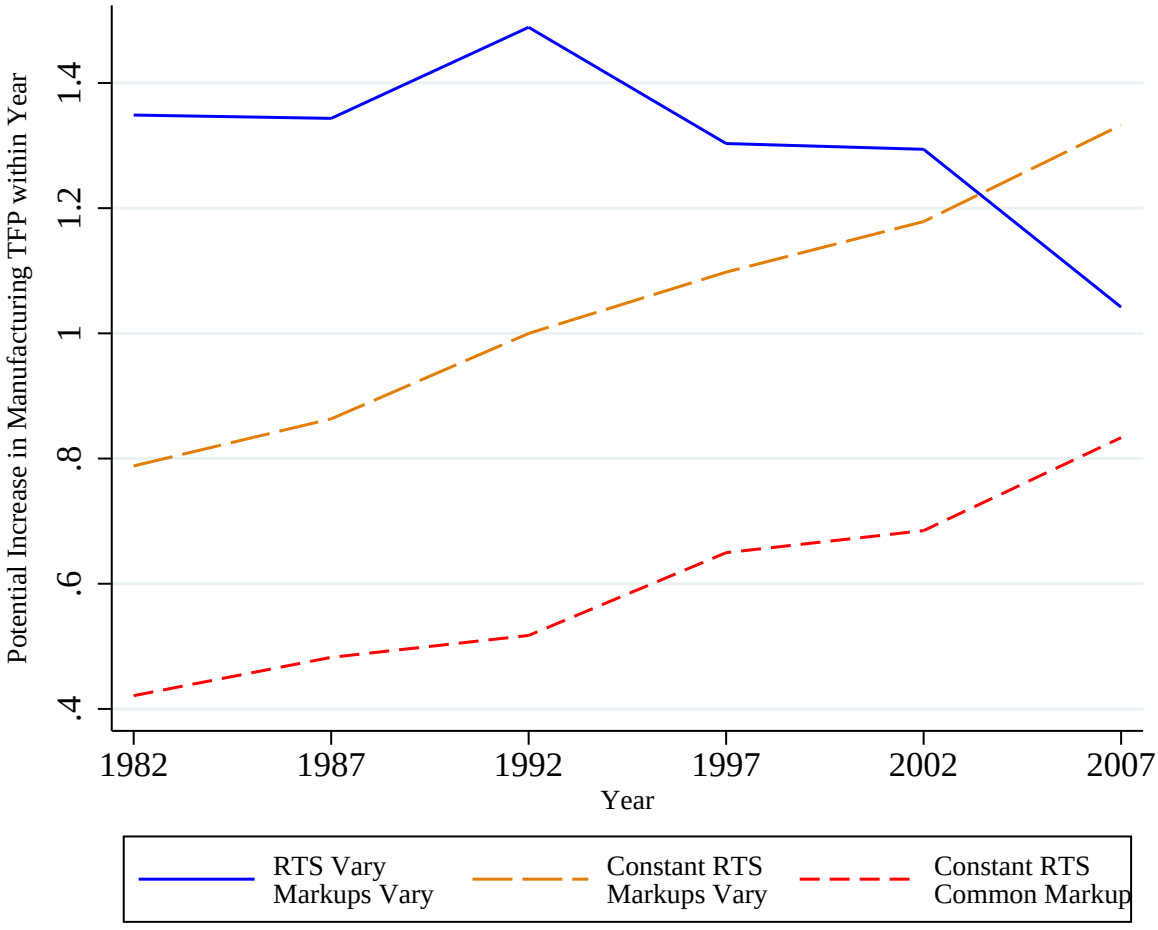
We focus on trends in misallocation, rather than levels, because the model is static and consequently imposes the long-run steady state at each point in time. As we described in section 3.2, the model infers distortions by assuming that, in a world without misallocation, establishments hire inputs until their average revenue products are equalized across establishments. Short-run considerations can change that inference: for instance, adjustment costs or the time required to build productive capital could lead non-distorted estab-

lishments to differ in their average revenue products at a point in time. Despite these costs, we follow the literature and impose the steady-state assumption for two reasons. First, by using a static model we can transparently document the role that industry-varying markups and returns to scale play in changing the measure of misallocation. Second, while these short-run considerations may lead us to misstate the level of misallocation, they likely have a smaller impact on trends across long periods of time.[17]

To understand the source of the divergent trends in misallocation, we next decompose the discrepancy in measured misallocation into a component from imposing a common markup across industries and a component from imposing constant returns to scale. In figure 5 we preview the formal decomposition by plotting an intermediate measure of misallocation in which we include only one source of industry variation. In the long-dashed

Figure 5: Misallocation in U.S. Manufacturing

Change in U.S. Manufacturing TFP from Equalizing Within-Industry Distortions



[17]Also, White et al. (2018) use special imputation flags available in the 2002 and 2007 Census of Manufacturing to show that imputation procedures tend to compress the measured distribution of TFPR. The tendency to impute the mean would likely lower the level of measured misallocation. However, if the tendency to impute remained relatively constant over time, then trends in misallocation could be better measured.

orange line we impose constant returns to scale, but maintain the estimated markups that vary across industries. The discrepancy in measured misallocation between our model and the Hsieh-Klenow model can now be split into two parts. The discrepancy from imposing the common markup is the distance from the intermediate model's long-dashed line and the Hsieh-Klenow model's short-dashed line. The discrepancy from imposing constant returns to scale is the distance between the our model's solid line and the new intermediate model's long-dashed one.

As we formally show over the next two sections, the divergent trends in misallocation are driven by the reduction in returns to scale between 1982 and 2007. As the U.S. manufacturing sector began to better approximate the assumed constant returns in the Hsieh-Klenow model, the discrepancy from imposing constant returns declined, leading to a perceived rise in misallocation. Figure 5 shows that most of the discrepancy in 1982 came from imposing constant returns to scale. By 2007, the discrepancy from imposing constant returns was less than half its initial value in absolute terms, while the discrepancy from imposing a common markup remained relatively unchanged. This reversal is reflected in the changing distances between the three lines. The shrinking distance between our solid blue line and the intermediate model's orange long-dashed line reflects the declining discrepancy in misallocation from imposing constant returns to scale. By contrast, the relatively stable distance between the Hsieh-Klenow model's and the intermediate model's lines suggests a more stable discrepancy over time from imposing a common markup.

4.2 Aggregate Decomposition

Having shown in section 2 that incorrect markups and returns to scale lead to spurious correlations between productivity and distortion, we now show how those spurious correlations lead to discrepancies between our measure of misallocation and the Hsieh-Klenow measure. We emphasize that these discrepancies are positive when we overstate the correlation of productivity and distortion, and that the discrepancies are negative when we understate the correlation of productivity and distortion.

In the following schematic, we present the theoretical decomposition where the second row splits the aggregate discrepancy into one component from imposing constant returns to scale and another component from imposing the common markup. The aggregate discrepancy measures the difference in misallocation between the Hsieh-Klenow model (constant returns to scale [CRTS] and a common markup of 1.5 everywhere [$\sigma = 3$]) and our own (returns to scale [VRTS] and markups [$\widehat{\sigma}$] can both vary). We first capture the component from imposing constant returns by comparing the CRTS and VRTS measures

of misallocation under the estimated markups $\widehat{\sigma}$. We then capture the component from the common markup by comparing the common markup ($\sigma = 3$) misallocation to the variable markup ($\widehat{\sigma}$) misallocation under CRTS. We can further decompose each component to understand the contribution of decreasing versus increasing returns to scale, as well as understating versus overstating the markup.

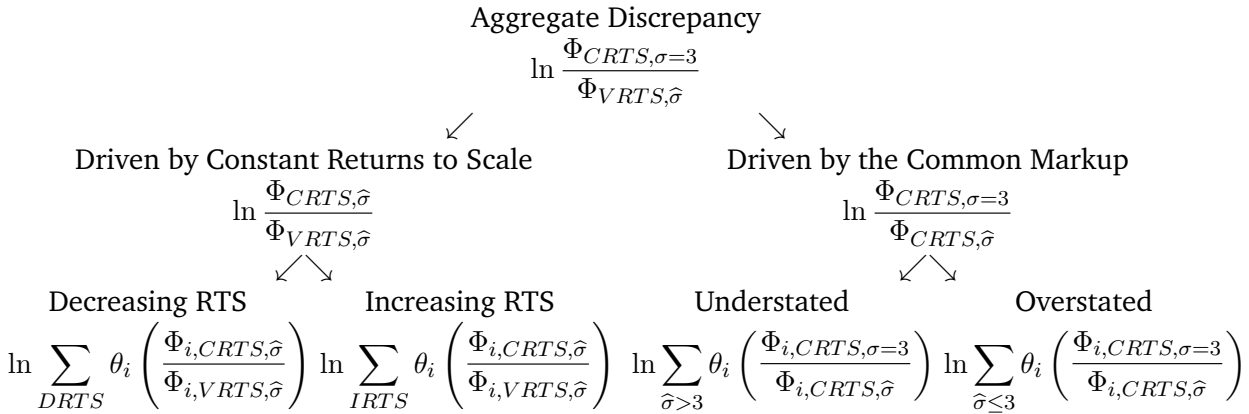


Table 3 decomposes the aggregate discrepancy in misallocation and shows that a decline in returns to scale explains why the discrepancy is smaller in 2007 than in 1982. The first two rows of panel A show that the 50% difference in misallocation between the Hsieh-Klenow model and our own in 1982 is split rather evenly between the imposition of constant returns to scale and the imposition of a common markup. By 2007, the aggregate discrepancy of 12% is split unevenly: the returns-to-scale component is half its previous value in absolute terms, while the markup component is essentially unchanged in size. These values quantify the visual decomposition from figure 5; the values in 1982 and 2007 capture the vertical distances among the three lines in the figure.

The third row of table 3 relates the discrepancy in misallocation to spurious correlations of productivity and distortion. In parentheses, the third row reports the difference in the correlation of productivity and distortion between the Hsieh-Klenow model and our own. For instance, the top of panel A indicates that imposing constant returns to scale on decreasing-returns industries in 1982 leads us to overstate the correlation of productivity and distortion by 0.13. By overstating this correlation, the constant-returns model also overstates misallocation, in this instance by 17%.[18] Across all deviations from the Hsieh-Klenow assumptions and across both years, inducing spurious positive correlation of productivity and distortion leads us to overstate misallocation, and inducing spurious negative correlations leads us to understate misallocation.

[18]The 17% is scaled by the size of industries with decreasing returns to scale. This scaling helps explain why understating the markup contributes only 3% to the overall discrepancy even though the correlation of productivity and distortion is overstated by a 0.28. In short, many fewer industries overstate the markup.

Table 3: Decomposing the Differences in Misallocation

Panel A: 1982

Aggregate Discrepancy			
-0.4968			
Driven by Constant Returns to Scale		Driven by the Common Markup	
-0.2589		-0.2378	
Decreasing RTS	Increasing RTS	Understated	Overstated
0.1761	-0.4470	0.0317	-0.2599
(0.1315)	(-0.2805)	(0.2847)	(-0.2218)

Panel B: 2007

Aggregate Discrepancy			
-0.1212			
Driven by Constant Returns to Scale		Driven by the Common Markup	
0.1249		-0.2461	
Decreasing RTS	Increasing RTS	Understated	Overstated
0.4676	-0.3349	0.0455	-0.2853
(0.1850)	(-0.2641)	(0.1290)	(-0.2891)

4.3 Misallocation and Business Dynamism

Having shown how the aggregate measure of misallocation has evolved over time, we now relate the measures of misallocation to measures of business dynamism; we show that industries with larger relative declines in misallocation experienced more job reallocation, as well as more establishment entry and exit. For this exercise, we draw on the publicly-available Business Dynamics Statics (BDS) from the U.S. Census. At the 4-digit NAICS level, the BDS measures each industry's job reallocation rate as the sum of the job creation and job destruction rates. Entry and exit rates in each industry are measured as counts of entering/existing establishments relative to a count of active establishments.

Table 4 shows that misallocation and business dynamism measures move in opposite directions: industries where misallocation increases over any five-year period are also likely to see less job reallocation and falling rates of establishment entry and exit. To match the level of aggregation of the BDS statistics, we proceed in two ways. Our baseline measure of misallocation are at the 6-digit NAICS level; we aggregate those estimates to the 4-digit level using a Cobb-Douglas aggregator as in equation (1). We also estimate the model directly at the 4-digit NAICS level (and discuss these misallocation estimates further in section 5). Misallocation measures tend to be inversely correlated with measures of

Table 4: Five-Year Changes in Misallocation and Business Dynamism

Dependent Variable	Job Reallocation Rate		Entry Rate		Exit Rate	
	(1)	(2)	(3)	(4)	(5)	(6)
Industry Misallocation (4-digit NAICS)	–0.3596		–0.1604		–0.0915	
	(0.1675)		(0.0806)		(0.0648)	
Industry Misallocation (6-digit NAICS)		–0.4616		–0.3057		0.0248
		(0.3171)		(0.1287)		(0.1054)
Observations	450	450	450	450	450	450
R-squared	0.2667	0.2619	0.1681	0.1713	0.3483	0.3445

Note: The dependent variables are drawn from the 2018 vintage of the Business Dynamics Statistics. The measures of misallocation are constructed using data from the Annual Survey of Manufactures from the U.S. Census, and from the National Compensation Survey from the U.S. Bureau of Labor Statistics. Each row shows the results of a different estimation: estimates in the first row correspond to a definition of the industry at the 4-digit NAICS level; estimates for the second row are constructed at the 6-digit NAICS level and then aggregated to the 4-digit level to map to the Business Dynamics Statistics.

business dynamism, with stronger correlations when both BDS statistics and misallocation are measured at the 4-digit NAICS level.

5 Robustness

In this section, we argue that different trends in misallocation persist even when we incorporate additional modifications to the model and the data. We restrict the model and explore the possibility that all changes in profits are driven by markups; we generalize the model to allow markups to vary across establishments in an industry; and, we explore changes to baseline samples, industry definitions, and estimation assumptions. We end with a discussion of value added and gross output measures of misallocation. We highlight the challenges in estimating the gross-output version of model and provide estimates from three complementary approaches. The resulting trends in misallocation are qualitatively similar to our baseline results across a variety of modeling assumptions and estimation approaches. Although different parameter estimates lead to different point estimates for the growth in misallocation, the stark qualitative differences between our model and the Hsieh-Klenow model remain throughout.

Model Parametrization

In our first robustness exercise, we emphasize the need for time-varying model parameters for capturing the evolution of the profit shares. While our estimates match the rising profit shares through a decline in returns to scale, we consider an alternative parametrization: we impose constant returns to scale, and calculate hypothetical markups that account for all the industry and time variation in profit shares. In panel A of table 5, we show that matching industry profits through markups alone also does away with the increasing trend in misallocation from the Hsieh-Klenow model. By this alternative calculation, misallocation between 1982 and 2007 is virtually unchanged, increasing by 3%. By contrast, the baseline misallocation from the Hsieh-Klenow model increased 29% over the same period. We view the elimination of this upward trend in misallocation as evidence that accounting for changing industry profits is of first-order importance for measuring misallocation.

In our second robustness exercise, we allow establishments to charge different markups *within* an industry. Formally, we follow Atkeson and Burstein (2008) in assuming that establishments sell their output in oligopolistically competitive markets instead of monopolistically competitive ones. In this setting, an establishment is aware that its choice of how much to produce affects both its own price *and* also the price level of the whole industry. Larger establishments exert a larger impact on the industry price level and this influence is reflected in larger markups. This establishment-specific markup depends on the elasticity of substitution σ_i, which is common to all industries in the Hsieh-Klenow model and varies across industries in our model. We present full details of the model in appendix B. One key challenge in this extension is to solve for the establishment-specific markup in the counterfactual where we eliminate distortions. This problem is akin to a contraction mapping, and we solve it by iterating on an initial guess. A second challenge is one of endogeneity: large firms do not take the industry price index as a given when choosing their price and output. To deal with this challenge we drop the 5% largest establishments by industry market share when estimating the parameters of equation (22); we then bring those establishments back into the sample when quantifying misallocation.[19]

Panel A of table 5 shows that the additional generalization to markups that vary within the industry leaves trends in misallocation essentially unchanged. Relative to the baseline 29% increase and the 13% decline, allowing markups to vary across establishments leads to a 28% increase and an 15% decline, respectively, in the Hsieh-Klenow model and in our own. While the trends in misallocation remain unchanged, the levels of misallocation

[19]The choice to drop the largest establishments in estimation is consistent with quantitative findings by di Giovanni and Levchenko (2012) and Gaubert and Itskhoki (2021) that only the very largest couple of firms set a markup meaningfully different from the mononopolistic-competition benchmark in these models.

Table 5: U.S. Manufacturing Misallocation in 2007 Relative to 1982, Robustness

Panel A:	Baseline Estimates	
	Hsieh-Klenow Model	Our Model
Baseline	0.29	-0.13
Model Change: impose constant returns to scale with implicit markups to match profit shares		0.03
Model Change: allow markups to vary across establishments in an industry	0.28	-0.15
Sample Change: use Census of Manufactures instead of Annual Survey of Manufactures	0.27	-0.09
Panel B:	Alternate Estimates	
	Hsieh-Klenow Model	Our Model
Estimation Change: estimate labor share of value added using Ackerberg et al (2015) instead of FOC	0.22	0.09
Estimation Change: define industries more broadly as NAICS 4-digit instead of NAICS 6-digit	0.26	-0.32
Estimation Change: use ten-year panels instead of five-year panels and compare 2007 to 1987	0.18	-0.02

decline with heterogeneous markups within the industry. The decline is more notable in our model, with misallocation some 10% lower per year (e.g., from 104% to 96% in 2007), while the level in the Hsieh-Klenow model declines about 3% (e.g., 83% to 81% in 2007).

In a third robustness exercise, also reported in panel A, we argue that the different patterns of misallocation are robust to accounting for sample selection in the Annual Survey of

Manufactures. The survey covers all large establishments and a random sample of smaller ones. Our baseline estimates of misallocation account for this sample selection by weighting establishments by their Census-provided sampling weights in calculating industry and aggregate misallocation. For this exercise, we construct the measure of misallocation using the full Census of Manufactures in 1982 and 2007, two of the years for which we have such data available. At a 27% increase and a 9% decline, the results of this extension replicate the baseline patterns.

Model Estimation

We next consider alternative ways, and sets of assumptions, for estimating markups and returns to scale, and argue that introducing industry and time variation in these parameters continues to remove the sharp increase in misallocation from the Hsieh-Klenow model. First, instead of calculating the labor share of value added β_{L_i} directly as the share of labor expenditures, we estimate β_{L_i} in a control-function procedure alongside the two other elasticities. Second, we estimate markups and returns to scale for more broadly defined industries. Third, we lengthen the time frame of the estimation, using ten-year panels instead of five-year panels of data to estimate markups and returns to scale.

While our baseline estimates directly measure the labor share as the ratio of labor costs to value added, at the top of panel B we instead estimate the labor share as a revenue elasticity using the Ackerberg et al. (2015) correction to the Levinsohn and Petrin (2003) control-function procedure. To estimate this labor elasticity, we need additional assumptions that justify the use of intermediate inputs as proxies for productivity. One possibility is that some unobserved component of productivity is realized after an establishment chooses its labor and before it chooses its intermediate inputs. Hence, we now have to assume that establishments choose the labor they hire before they choose their intermediate inputs, and that unobserved productivity is realized before the intermediate-input choice. Our estimates of this labor elasticity suggest an 11% decline in labor's share of value added, compared to our direct calculation of a 25% decline. With a smaller decline of the labor share, we also find a smaller reduction in returns to scale over time. Ultimately, this more modest change in returns to scale over time leads to a smaller departure from the Hsieh-Klenow model's trend in misallocation; these alternate estimates imply a 9% increase in misallocation, a bit less than half the increase in the Hsieh-Klenow model.

We next estimate markups and returns to scale for more broadly-defined industries, and find that the divergent patterns of misallocation are amplified. Specifically, we use the NAICS-4 industry code instead of the more detailed NAICS-6. For instance, an industry now corresponds to "Dairy Product" instead of "Ice Cream and Frozen Dessert." The sec-

ond entry in panel B shows that while misallocation in the Hsieh-Klenow model increases a bit over of 20%, misallocation in our model falls 32%, more than twice our baseline decline. This larger decline reflects an interaction of two forces. First, our measure of misallocation focuses on within-industry reallocation of resources. When we broaden the industry definition, we implicitly allow resources to be allocated across the NAICS-6 industries that comprise a NAICS-4. Second, returns to scale determine how large an establishment grows as a share of the industry when its distortions are removed. The larger are the returns to scale, the greater is the share of industry revenue generated by the most productive establishment. The interaction of larger industries and the reduction in returns to scale over time amplifies the decline in misallocation relative to our baseline results.

We then use ten-year instead of five-year panels to estimate the model parameters; this procedure attenuates the differences in parameter values across time and hence reduces the differences in misallocation trends between the two models. Under these parameter estimates, our model suggests that misallocation decreased 2% between 1987 and 2007 while the Hsieh-Klenow model implies an increase of 18%. We contextualize these estimates by reference to table 2, panel A, in which we document a continuous decline in returns to scale over the same period. By pooling the last decade of data in this exercise, our estimate of the decline in returns to scale is smaller than when we compare returns to scale only using the first five and the last five years of the sample. Nonetheless, even this smoothing of parameter estimates preserves the divergent trends in misallocation.

Gross Output Alternative

While we take as our baseline a model where establishments combine capital and labor to produce value added, we also provide evidence of divergent patterns of misallocation in models of gross output. The value-added baseline allows us to estimate returns to scale and misallocation in a model-consistent manner. The drawback to the value-added specification is that the implied measures of productivity for value added and for gross output are identical only under specific modeling assumptions.[20] To draw attention more broadly to the importance of returns to scale and markups in the measurement of misallocation, we therefore extend our analysis to gross-output production functions.

We face two key impediments to estimating a gross-output version of our model. For one, control-function approaches are unable to identify returns to scale in gross-output production functions, as highlighted by Ackerberg et al. (2015) and Gandhi et al. (2020).

[20]Namely, the core estimating equation in terms of value added (7) can also be derived from a gross-output production function that is Leontief in materials whose price is proportional to the price of output, as discussed, for instance, in Ackerberg et al. (2015).

Table 6: U.S. Manufacturing Misallocation in 2007 Relative to 1982
Gross Output versus Value Added

Panel A:	Misallocation in 2007 Relative to 1982	
	Hsieh-Klenow Model	Our Model
Value Added Baseline	0.29	-0.13
Gross Output Alternatives		
1. estimate labor and materials elasticities from FOCs and the rest using GMM	0.11	0.00
2. rescale value-added parameters following Basu & Fernald (2002)	0.12	-0.03
3. impose constant returns to scale on the estimation of all elasticities	0.10	0.05

Panel B:	Returns to Scale		Markups	
	1982	2007	1982	2007
Value Added Baseline	1.23	0.96	1.46	1.48
Gross Output Alternatives				
1. estimate labor and materials elasticities from FOCs and the rest using GMM	1.20	1.14	1.29	1.35
2. rescale value-added parameters following Basu & Fernald (2002)	1.04	0.94	1.15	1.17
3. impose constant returns to scale on the estimation of all elasticities	1.00	1.00	1.10	1.22

The challenge to estimation is that a freely-chosen input (e.g., materials) cannot simultaneously be used both to proxy for productivity through a control function and also to estimate the revenue elasticity with respect to itself. Moreover, this conceptual challenge is compounded by the data limitation that we observe only expenditures on materials and not the physical quantity chosen of materials. This data limitation creates the tension that

we have to use the same data object both to estimate the expenditure share on materials and also to apply as the physical measure of the input.

In view of these estimation challenges, we present three complementary approaches to estimating the gross-output version of the model; while none of the three can simultaneously overcome all the measurement challenges highlighted in the literature, each approach draws on a different source of identification. First, we extend the estimating equation (22) to include an additional revenue elasticity β_{M_i}. We estimate this elasticity using the expenditures on intermediate inputs as a share of gross output—an object we calculate directly in the data—and we then estimate the capital and output elasticities as before. Second, we re-scale the parameters from the value-added model to construct gross-output parameters following Basu and Fernald (2002). Third, we impose constant returns to scale when we estimate the production function and thus assign all variation in profits to markups across industries and time.

The results in table 6 reinforce our baseline findings: estimating returns to scale and markups undoes the sharp rise in misallocation from the baseline Hsieh-Klenow model. Across all specifications, the change in misallocation is smaller for the gross-output model than for the value-added model, as per panel A. The increase in misallocation for the Hsieh-Klenow model averages about 11%. Misallocation in our model is either unchanged or falls by 3% when we estimate both returns to scale and markups; when we attribute all changes to the markup and impose constant returns to scale, we find an increase of 5% in misallocation. Behind these estimates of changing misallocation are the estimates of reductions in returns to scale and rises in markups in panel B.

6 Conclusion

We argue in this paper that accounting for industry and time variation in markups and returns to scale leads to a measure of misallocation in U.S. manufacturing that is decreasing over time; this result stands in contrast to the increasing measure of misallocation under the widely-applied assumptions of a common markup and constant returns to scale, as in the Hsieh-Klenow model. To quantify these differences, we use five-year panels of restricted U.S. Census microdata to estimate markups and returns to scale across manufacturing industries. We find that industries differ meaningfully in these parameters at a given point in time, and that the average returns to scale in U.S. manufacturing declined between 1982 and 2007.

We decompose the differences in misallocation between the two models, and identify the decline in returns to scale as the primary driver of the divergent trends in misallocation. The Hsieh-Klenow measure on average understates our measure of misallocation. The

assumption of constant returns to scale is a better fit for the data in 2007 than it is for 1982. Consequently, as the U.S. manufacturing sector began to reflect more closely the assumption of constant returns, the discrepancy in measuring misallocation declined. As this discrepancy declined, the Hsieh-Klenow measure of misallocation asymptoted toward our measure from below and hence drove the upward trend in misallocation.

We formalize the source of these differences in misallocation and show that, by ignoring the variation in markups and returns to scale, the Hsieh-Klenow model measures productivity in a way that conflates productivity and distortions. These spurious correlations lead us to incorrectly infer the extent to which the most productive establishments bear the most burdensome distortions, and hence to an incorrect measure of misallocation.

We think the patterns we identify in markups and returns to scale, and the discrepancies we highlight in measuring productivity, could be of broader interest. Outside the literature on misallocation, the measurement of establishment-level productivity is a key input in other attempts to trace the impacts of policies and shocks from affected establishments to aggregate outcomes.

References

Ackerberg, Daniel A., Kevin Caves, and Garth Frazer, "Identification Properties of Recent Production Function Estimators," *Econometrica*, 2015, *83* (6), 2411–2451.

Aghion, Philippe, Antonin Bergeaud, Timo Boppart, Peter J. Klenow, and Huiyu Li, "Missing Growth from Creative Destruction," *American Economic Review*, August 2019, *109* (8), 2795–2822.

Asker, John, Allan Collard-Wexler, and Jan De Loecker, "Dynamic Inputs and Resource (Mis)Allocation," *Journal of Political Economy*, 2014, *122* (5), 1013–1063.

Atkeson, Andrew and Ariel Burstein, "Pricing-to-Market, Trade Costs, and International Relative Prices," *American Economic Review*, 2008, *98* (5), 1998–2031.

Autor, David, David Dorn, Lawrence F Katz, Christina Patterson, and John Van Reenen, "The Fall of the Labor Share and the Rise of Superstar Firms," *The Quarterly Journal of Economics*, 02 2020, *135* (2), 645–709.

Barkai, Simcha, "Declining Labor and Capital Shares," *The Journal of Finance*, 2020, *75* (5), 2421–2463.

Bartelsman, Eric, John Haltiwanger, and Stefano Scarpetta, "Cross-Country Differences in Productivity: The Role of Allocation and Selection," *American Economic Review*, 2013, *103* (1), 305–34.

Basu, Susanto, "Are Price-Cost Markups Rising in the United States? A Discussion of the Evidence," *Journal of Economic Perspectives*, August 2019, *33* (3), 3–22.

_ **and John G. Fernald**, "Returns to Scale in U.S. Production: Estimates and Implications," *Journal of Political Economy*, 1997, *105* (2), 249–283.

_ **and** _ , "Aggregate productivity and aggregate technology," *European Economic Review*, 2002, *46* (6), 963 – 991.

_ , _ **, and Miles S. Kimball**, "Are Technology Improvements Contractionary?," *The American Economic Review*, 2006, *96* (5), 1418–1448.

Bils, Mark, Peter J. Klenow, and Cian Ruane, "Misallocation or Mismeasurement?," *Working Paper*, 2017.

Bond, Stephen, Arshia Hashemi, Greg Kaplan, and Piotr Zoch, "Some Unpleasant Markup Arithmetic: Production Function Elasticities and their Estimation from Production Data," *Journal of Monetary Economics*, Forthcoming.

Broda, Christian and David E. Weinstein, "Globalization and the Gains From Variety," *The Quarterly Journal of Economics*, 2006, *121* (2), 541–585.

Cooper, Russell W and John C Haltiwanger, "On the nature of capital adjustment costs," *The Review of Economic Studies*, 2006, *73* (3), 611–633.

De Loecker, Jan, "Product Differentiation, Multiproduct Firms, and Estimating the Impact of Trade Liberalization on Productivity," *Econometrica*, 2011, *79* (5), 1407–1451.

_ **, Jan Eeckhout, and Gabriel Unger**, "The Rise of Market Power and the Macroeconomic Implications," *The Quarterly Journal of Economics*, 01 2020, *135* (2), 561–644.

di Giovanni, Julian and Andrei A. Levchenko, "Country Size, International Trade, and Aggregate Fluctuations in Granular Economies," *Journal of Political Economy*, 2012, *120* (6), 1083–1132.

Edmond, Chris, Virgiliu Midrigan, and Daniel Yi Xu, "How Costly Are Markups?," *NBER Working Paper*, 2018.

Elsby, Michael W. L., Bart Hobijn, and Aysegul Sahin, "The decline of the U.S. labor share," *Brookings Papers on Economic Activity*, 2013, pp. 1–42.

Fort, Teresa C. and Shawn D. Klimek, "The Effect of Industry Classification Changes on U.S. Employment Composition," *Working Paper*, 2015.

_ **, Justin R. Pierce, and Peter K. Schott**, "New Perspectives on the Decline of US Manufacturing Employment," *Journal of Economic Perspectives*, May 2018, *32* (2), 47–72.

Foster, Lucia, Cheryl Grim, and John Haltiwanger, "Reallocation in the Great Recession: Cleansing or Not?," *Journal of Labor Economics*, 2016, *34* (S1), S293–S331.

_ , _ , _ **, and Zoltan Wolf**, "Firm-Level Dispersion in Productivity: Is the Devil in the Details?," *American Economic Review*, May 2016, *106* (5), 95–98.

_ **, John Haltiwanger, and Chad Syverson**, "Reallocation, Firm Turnover, and Efficiency: Selection on Productivity or Profitability?," *American Economic Review*, 2008, *98* (1), 394–425.

Gandhi, Amit, Salvador Navarro, and David A. Rivers, "On the Identification of Gross Output Production Functions," *Journal of Political Economy*, 2020, *128* (8), 2973–3016.

Gaubert, Cecile and Oleg Itskhoki, "Granular Comparative Advantage," *Journal of Political Economy*, 2021, *129* (3), 871–939.

Gopinath, Gita, Sebnem Kalemli-Ozcan, Loukas Karabarbounis, and Carolina Villegas-Sanchez, "Capital Allocation and Productivity in South Europe," *The Quarterly Journal of Economics*, 2017, *132* (4), 1915–1967.

Hall, Robert E., "Invariance Properties of Solow's Productivity Residual," in Peter Diamond, ed., *Growth/ Productivity/ Unemployment: Essays to Celebrate Bob Solow's Birthday*, Cambridge, Mass.: MIT Press, 1990.

Haltiwanger, John, Robert Kulick, and Chad Syverson, "Misallocation Measures: The Distortion That Ate the Residual," *National Bureau of Economic Research Working Paper Series*, 2018, *No. 24199*.

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

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

Karabarbounis, Loukas and Brent Neiman, "The Global Decline of the Labor Share," *The Quarterly Journal of Economics*, 2014, *129* (1), 61–103.

Kehrig, Matthias, "The Cyclicality of Productivity Dispersion," *US Census Bureau Center for Economic Studies Working Paper*, 2011, *No. CES-WP-11-15*, 67.

_ **and Nicolas Vincent**, "The Micro-Level Anatomy of the Labor Share Decline," *The Quarterly Journal of Economics*, 03 2021, *136* (2), 1031–1087.

Klette, Tor Jakob and Zvi Griliches, "The Inconsistency of Common Scale Estimators When Output Prices are Unobserved and Endogenous," *Journal of Applied Econometrics*, 1996, *11* (4), 343–361.

Levinsohn, James and Amil Petrin, "Estimating Production Functions Using Inputs to Control for Unobservables," *The Review of Economic Studies*, 2003, *70* (2), 317–341.

Marschak, Jacob and William H. Andrews, "Random Simultaneous Equations and the Theory of Production," *Econometrica*, 1944, *12* (3/4), 143–205.

Olley, G. Steven and Ariel Pakes, "The Dynamics of Productivity in the Telecommunications Equipment Industry," *Econometrica*, 1996, *64* (6), 1263–1297.

Restuccia, Diego and Richard Rogerson, "Policy distortions and aggregate productivity with heterogeneous establishments," *Review of Economic Dynamics*, 2008, *11* (4), 707–720.

White, T. Kirk, Jerome P. Reiter, and Amil Petrin, "Imputation in U.S. Manufacturing Data and Its Implications for Productivity Dispersion," *The Review of Economics and Statistics*, 07 2018, *100* (3), 502–509.

Appendices

A Model Summary

Aggregation

We assume that the manufacturing sector is characterized by a representative establishment selling its output Y in a perfectly competitive market. This firm aggregates the output Y_i of I different industries using a Cobb-Douglas production technology with elasticities θ_i:

$$Y = \prod_{i=1}^{I} Y_i^{\theta_i}, \text{ with } \sum_{i=1}^{I} \theta_i = 1. \tag{A.1}$$

Cost minimization by this aggregating firm implies that θ_i is also each industry's share of aggregate expenditure

$$P_i Y_i = \theta_i P Y, \tag{A.2}$$

where P_i is the price of an industry composite good, and P is the price of the final good

$$P = \prod_{i=1}^{I} \left(\frac{P_i}{\theta_i} \right)^{\theta_i}. \tag{A.3}$$

An industry aggregating firm produces Y_i from the output of N_i differentiated establishments via a constant-elasticity-of-substitution (CES) technology with elasticity σ_i

$$Y_i = \left(\sum_{e=1}^{N_i} Y_{ie}^{\frac{\sigma_i - 1}{\sigma_i}} \right)^{\frac{\sigma_i}{\sigma_i - 1}}. \tag{A.4}$$

Cost minimization by the industry aggregating firm implies a standard CES price index P_i:

$$P_i = \left[\sum_{e=1}^{N_i} \left(\frac{1}{P_{ie}} \right)^{\sigma_i - 1} \right]^{\frac{-1}{\sigma_i - 1}}. \tag{A.5}$$

Establishment Optimization

Each establishment in the industry produces value-added output Y_{ie} by combining its TFP A_{ie}, capital K_{ie} and labor L_{ie} in a Cobb-Douglas production function

$$Y_{ie} = A_{ie} K_{ie}^{\alpha_{K_i}} L_{ie}^{\alpha_{L_i}}, \tag{A.6}$$

where the industry level returns to scale α_i are the sum of the output elasticities α_{K_i} and α_{L_i}. The establishment maximizes profits by taking as given the prices R and w from perfectly competitive input markets. However, the effective cost of an input varies across establishments, with the $\tau_{K_{ie}}$ and $\tau_{L_{ie}}$ capturing these input-specific distortions for capital and labor, respectively

$$\pi_{ie} = P_{ie} Y_{ie} - (1 + \tau_{L_{ie}}) w L_{ie} - (1 + \tau_{K_{ie}}) R K_{ie}. \tag{A.7}$$

By internalizing the demand for its variety, the establishment charges a price that is a constant markup over its marginal cost. Note that the marginal cost under variable RTS depends on the scale of production:

$$P_{ie} = \Omega_{P_i} \left[\frac{(1 + \tau_{K_{ie}})^{\alpha_{K_i}} (1 + \tau_{L_{ie}})^{\alpha_{L_i}}}{A_{ie}} \right]^{\frac{1}{\alpha_i + \sigma_i (1 - \alpha_i)}} \tag{A.8}$$

$$\text{where } \Omega_{P_i} = \left(P_i^{\sigma} Y_i \right)^{\frac{1 - \alpha_i}{\alpha_i + \sigma_i (1 - \alpha_i)}} \left[\left(\frac{\sigma_i}{\sigma_i - 1} \right)^{\alpha_i} \left(\frac{R}{\alpha_{K_i}} \right)^{\alpha_{K_i}} \left(\frac{w}{\alpha_{L_i}} \right)^{\alpha_{L_i}} \right]^{\frac{1}{\alpha_i + \sigma_i (1 - \alpha_i)}}$$

$$P_{ie} = \frac{\sigma_i}{\sigma_i - 1} \left[\left(\frac{R}{\alpha_{K_i}} \right)^{\alpha_{K_i}} \left(\frac{w}{\alpha_{L_i}} \right)^{\alpha_{L_i}} \right]^{\frac{1}{\alpha_i}} \left(Y_{ie} \right)^{\frac{1 - \alpha_i}{\alpha_i}} \left[\frac{(1 + \tau_{K_{ie}})^{\alpha_{K_i}} (1 + \tau_{L_{ie}})^{\alpha_{L_i}}}{A_{ie}} \right]^{\frac{1}{\alpha_i}}.$$

Within the confines of this model, there is a natural restriction on the returns to scale parameter. As in Basu and Fernald (1997), standard cost-minimization requires that the RTS parameter α_i is (weakly) less than the markup $\sigma_i/(\sigma_i - 1)$. The returns to scale and the markup shape the price elasticities of supply and demand, respectively. The price elasticity of supply is increasing in the RTS parameter α_i: when RTS are sufficiently large, the supply curve becomes downward sloping. The restriction that α_i is smaller than the markup guarantees that a downward-sloping supply curve is not steeper than a downward-sloping demand curve. This restriction ensures that the willingness-to-pay reflected in the demand curve exceeds the cost of production embodied by the supply curve when establishments are deciding whether to produce. A rearrangement of this inequality guarantees that the often-recurring term $[\alpha_i + \sigma_i(1 - \alpha_i)]$ is positive.

An establishment facing larger distortions uses less capital and labor.

$$K_{ie} \propto \left[\frac{A_{ie}^{\sigma_i - 1}}{(1+\tau_{K_{ie}})^{[\alpha_i + \sigma_i(1-\alpha_i)] + \alpha_{K_i}(\sigma_i - 1)} (1+\tau_{L_{ie}})^{\alpha_{L_i}(\sigma - 1)}} \right]^{\frac{1}{\alpha_i + \sigma(1-\alpha_i)}} \tag{A.9}$$

$$L_{ie} \propto \left[\frac{A_{ie}^{\sigma_i - 1}}{(1+\tau_{K_{ie}})^{\alpha_{K_i}(\sigma_i - 1)} (1+\tau_{L_{ie}})^{[\alpha_i + \sigma(1-\alpha_i)] + \alpha_{L_i}(\sigma_i - 1)}} \right]^{\frac{1}{\alpha_i + \sigma(1-\alpha_i)}}. \tag{A.10}$$

Moreover, measured in terms of either physical output or the establishment's revenue share in the industry, a more distorted establishment is also smaller in size.

$$\frac{P_{ie}Y_{ie}}{P_i Y_i} = \frac{\left[A_{ie} \left(\frac{1}{\frac{1+\tau_{K_{ie}}}{1+\tau_{K,i}}} \right)^{\alpha_{K_i}} \left(\frac{1}{\frac{1+\tau_{L_{ie}}}{1+\tau_{L,i}}} \right)^{\alpha_{L_i}} \right]^{\frac{1}{\frac{\sigma_i}{\sigma_i - 1} - \alpha_i}}}{\sum_{e=1}^{N_i} \left[A_{ie} \left(\frac{1}{\frac{1+\tau_{K_{ie}}}{1+\tau_{K,i}}} \right)^{\alpha_{K_i}} \left(\frac{1}{\frac{1+\tau_{L_{ie}}}{1+\tau_{L,i}}} \right)^{\alpha_{L_i}} \right]^{\frac{1}{\frac{\sigma_i}{\sigma_i - 1} - \alpha_i}}}. \tag{A.11}$$

Marginal Revenue Products and Market Clearing

Distortions affect establishment choices by changing the marginal revenue gained from an additional unit of an input (e.g. $MRPK_{ie}$ for capital K_{ie}). In equilibrium, the marginal revenue product of an additional hired input equals the effective cost to the establishment of hiring the input. If an establishment faces barriers that make acquiring capital more expensive, then $(1+\tau_{K_{ie}})$ is high, and the establishment will only hire an additional unit of capital if its $MRPK_{ie}$ exceeds the cost $(1+\tau_{K_{ie}})R$. The same reasoning holds for all variable inputs in production.

$$MRPK_{ie} \triangleq MPK_{ie} \times P_{ie} \times \frac{\sigma_i - 1}{\sigma_i} = \alpha_{K_i} \frac{Y_{ie}}{K_{ie}} P_{ie} \frac{\sigma_i - 1}{\sigma_i} = (1+\tau_{K_{ie}})R \tag{A.12}$$

$$MRPL_{ie} \triangleq MPL_{ie} \times P_{ie} \times \frac{\sigma_i - 1}{\sigma_i} = \alpha_{L_i} \frac{Y_{ie}}{L_{ie}} P_{ie} \frac{\sigma_i - 1}{\sigma_i} = (1+\tau_{L_{ie}})w. \tag{A.13}$$

To understand the impact of establishment-level distortions for the productivity of the industry as a whole, we need to aggregate the establishment choices. Combining input-market-clearing conditions with establishment input choices, we can show that each indus-

try uses capital and labor in proportion to the industry's share of the national economy θ_i, to the industry's input elasticity α_{X_i} for a given factor X, and in inverse proportion to that factor's average marginal revenue products across the industry's establishments $\overline{MRPX_i}$.

$$K_i = K \frac{\alpha_{K_i}\theta_i \frac{1}{\overline{MRPK_i}}}{\sum_{i'=1}^{I} \alpha_{K_{i'}}\theta_{i'} \frac{1}{\overline{MRPK_{i'}}}} \tag{A.14}$$

$$L_i = L \frac{\alpha_{L_i}\theta_i \frac{1}{\overline{MRPL_i}}}{\sum_{i'=1}^{I} \alpha_{L_{i'}}\theta_{i'} \frac{1}{\overline{MRPL_{i'}}}}. \tag{A.15}$$

The average marginal revenue products are weighted by establishment size. In the absence of distortions, or if all establishments faced the same distortion, $MRPX_{ie}$ would be equal across establishments and hence equal to the industry $\overline{MRPX_i}$. We revisit this point below when we define a counterfactual allocation of resources in which all establishments are equally distorted.

$$\frac{1}{\overline{MRPK_i}} = \sum_{f=1}^{N_i} \frac{1}{MRPK_{ie}} \frac{P_{ie}Y_{ie}}{P_iY_i} = \frac{1}{R} \sum_{f=1}^{N_i} \frac{1}{(1+\tau_{K_{ie}})} \frac{P_{ie}Y_{ie}}{P_iY_i} \tag{A.16}$$

$$\frac{1}{\overline{MRPL_i}} = \sum_{f=1}^{N_i} \frac{1}{MRPL_{ie}} \frac{P_{ie}Y_{ie}}{P_iY_i} = \frac{1}{w} \sum_{f=1}^{N_i} \frac{1}{(1+\tau_{L_{ie}})} \frac{P_{ie}Y_{ie}}{P_iY_i}. \tag{A.17}$$

Much like the above definitions of average $MRPX$ in the industry, we simplify the notation for the average distortion in an industry by defining

$$\overline{(1+\tau_{X_i})} = \left[\sum_{e=1}^{N_i} \frac{P_{ie}Y_{ie}}{P_iY_i} \frac{1}{1+\tau_{X_{ie}}} \right]^{-1} \text{ for } X \in \{K, L\}.$$

Toward a Measure of Industry Productivity

Industry output can now be expressed as

$$Y_i = A_i K_i^{\alpha_{K_i}} L_i^{\alpha_{L_i}}, \tag{A.18}$$

where A_i is the total factor productivity TFP_i of the industry. In thinking about how distortions affect industry productivity, we introduce notation based on Foster et al. (2008) and Hsieh and Klenow (2009) that distinguishes the productivity for producing a quantity

of physical goods, A_{ie}, from the productivity for generating revenue, $TFPR_{ie}$.

$$TFPR_{ie} \triangleq P_{ie}A_{ie} = \frac{P_{ie}Y_{ie}}{K_{ie}^{\alpha_{K_i}} L_{ie}^{\alpha_{L_i}}}. \tag{A.19}$$

This distinction is helpful since two establishments with the same physical productivity A_{ie} can have different revenue productivities $TFPR_{ie}$ if they face different distortions. In other words, $TFPR$ can help summarize the impact of distortions on an establishment:

$$TFPR_{ie} = \left(\frac{\sigma_i}{\sigma_i - 1}\right)^{\alpha_i} \left(P_{ie}Y_{ie}\right)^{1-\alpha_i} \left[\frac{MRPK_{ie}}{\alpha_{K_i}}\right]^{\alpha_{K_i}} \left[\frac{MRPL_{ie}}{\alpha_{L_i}}\right]^{\alpha_{L_i}} \tag{A.20}$$

$$TFPR_{ie} \propto \left[(1+\tau_{K_{ie}})^{\alpha_{K_i}} (1+\tau_{L_{ie}})^{\alpha_{L_i}} A_{ie}^{(\sigma_i-1)(1-\beta_i)}\right]^{\frac{1}{\alpha_i+\sigma(1-\alpha_i)}}. \tag{A.21}$$

Revenue productivity increases in the level of distortions, as the establishment's input bundle has to compensate for a large effective cost of hiring the inputs.

We can define an industry revenue productivity following the establishment definition:

$$\overline{TFPR}_i \triangleq P_i A_i = \left(\frac{\sigma}{\sigma - 1}\right)^{\beta_i} \left(P_i Y_i\right)^{1-\beta_i} \left[\frac{\overline{MRPK}_i}{\alpha_{K_i}}\right]^{\alpha_{K_i}} \left[\frac{\overline{MRPL}_i}{\alpha_{L_i}}\right]^{\alpha_{L_i}}. \tag{A.22}$$

This formulation of industry revenue productivity allows us to write industry TFP_i as the CES aggregate of establishment physical productivity A_{ie}, weighted by the difference between industry and establishment revenue productivity $\overline{TFPR_i}/TFPR_{ie}$.

$$TFP_i = P_i A_i \frac{1}{P_i} = \overline{TFPR}_i \frac{1}{P_i} = \left[\sum_{e=1}^{M_i} \left(A_{ie} \frac{\overline{TFPR}_i}{TFPR_{ie}}\right)^{\sigma-1}\right]^{\frac{1}{\sigma-1}}. \tag{A.23}$$

The weight captures the establishment's size as well as the deviations of its marginal revenue products from their respective industry averages:

$$\frac{\overline{TFPR}_i}{TFPR_{ie}} = \left(\frac{P_i Y_i}{P_{ie} Y_{ie}}\right)^{1-\alpha_i} \left[\frac{\overline{MRPK}_i}{MRPK_{ie}}\right]^{\alpha_{K_i}} \left[\frac{\overline{MRPL}_i}{MRPL_{ie}}\right]^{\alpha_{L_i}} \tag{A.24}$$

$$\frac{\overline{TFPR}_i}{TFPR_{ie}} = \left(s_{ie}\right)^{\alpha_i - 1} \left(\frac{\overline{1+\tau_{K,i}}}{1+\tau_{K_{ie}}}\right)^{\alpha_{K_i}} \left(\frac{\overline{1+\tau_{L,i}}}{1+\tau_{L_{ie}}}\right)^{\alpha_{L_i}}. \tag{A.25}$$

Misallocation

More distorted establishments have smaller weights in industry productivity. Consequently, the correlation of productivity and distortion is important for measuring gains from equal-

izing the distortions faced by different establishments within the industry. If more productive establishments are also more distorted, then equalizing distortions would give larger weights to the more productive establishments in the counterfactual. This tilting of weights toward more productive establishments would translate to large TFP gains from reallocating inputs.

More formally, if all establishments within an industry face the same distortions, so that $\tau = \overline{\tau}$, then the establishment weights for calculating industry TFP_i simplify in the following manner:

$$\left.\frac{\overline{TFPR}_i}{TFPR_{ie}}\right|_{\tau=\overline{\tau}} = \left(s_{ie}|_{\tau=\overline{\tau}}\right)^{\alpha_i-1} = \left(\frac{\left[A_{ie}\right]^{\frac{1}{\frac{\sigma_i}{\sigma_i-1}-\alpha_i}}}{\sum_{e=1}^{N_i}\left[A_{ie}\right]^{\frac{1}{\frac{\sigma_i}{\sigma_i-1}-\alpha_i}}}\right)^{\alpha_i-1}. \tag{A.26}$$

Note that under constant returns to scale ($\alpha_i = 1$) $TFPR_{ie}$ is identical across all establishments. This equality is at the center of Hsieh and Klenow (2009) intuition: "A key result we exploit is that *revenue* productivity... should be equated across firms in the absence of distortions. To the extent revenue productivity differs across firms, we can use it to recover a measure of firm-level distortions" (1404). Note, however, that if returns to scale in an industry are not constant, then revenue productivity can vary across undistorted establishments. As a result, there is not a direct mapping between the variance of TFPR and the misallocation within industry. To calculate the gains from eliminating distortions, the econometrician has to calculate the counterfactual weight for each establishment.

For every industry i, we then define misallocation as Φ_i, the net gain to industry TFP from equalizing distortions across establishments within the industry:

$$\Phi_i = \frac{TFP_i|_{\tau=\overline{\tau}}}{TFP_i} = \frac{\left[\sum_{e=1}^{N_i}\left(A_{ie}\left.\frac{\overline{TFPR}_i}{TFPR_{ie}}\right|_{\tau=\overline{\tau}}\right)^{\sigma_i-1}\right]^{\frac{1}{\sigma_i-1}}}{\left[\sum_{e=1}^{N_i}\left(A_{ie}\frac{\overline{TFPR}_i}{TFPR_{ie}}\right)^{\sigma_i-1}\right]^{\frac{1}{\sigma_i-1}}}. \tag{A.27}$$

The misallocation for all of US manufacturing in a given year is then

$$\Phi = \prod_{i\in I}\Phi_i^{\theta_i}, \tag{A.28}$$

where θ_i is industry i's revenue share in the manufacturing sector.

B Heterogeneous Markups within Industry

In this appendix, we generalize our model to allow markups to vary across establishments in an industry. We introduce these heterogeneous markups by replacing monopolistic competition in output markets with oligopolistic competition, in the style of Atkeson and Burstein (2008). In short, we allow establishments to internalize their impact on the industry demand, leading them to change their price-setting behavior, with larger establishments now charging higher markups.

Previously, establishments internalized their own downward-sloping demand curves:

$$P_{ie} = P_i Y_i Y_i^{\frac{1-\sigma_i}{\sigma_i}} Y_{ie}^{\frac{-1}{\sigma_i}}. \tag{B.1}$$

Now they also internalize the demand for the industry aggregate, so we can write an individual establishment's demand curve as

$$P_{ie} = \theta_i P Y Y_i^{\frac{1-\sigma_i}{\sigma_i}} Y_{ie}^{\frac{-1}{\sigma_i}}. \tag{B.2}$$

Profit maximization on the part of these oligopolistic establishments leads to an updated expression for the equilibrium price, which is still a markup over marginal cost:

$$P_{ie} = \frac{\varepsilon(s_{ie})}{\varepsilon(s_{ie})-1}\left[\left(\frac{R}{\alpha_{K_i}}\right)^{\alpha_{K_i}}\left(\frac{w}{\alpha_{L_i}}\right)^{\alpha_{L_i}}\right]^{\frac{1}{\alpha_i}}\left(Y_{ie}\right)^{\frac{1-\alpha_i}{\alpha_i}}\left[\frac{(1+\tau_{K_{ie}})^{\alpha_{K_i}}(1+\tau_{L_{ie}})^{\alpha_{L_i}}}{A_{ie}}\right]^{\frac{1}{\alpha_i}}. \tag{B.3}$$

The establishment-specific markup $\varepsilon(s_{ie})/(\varepsilon(s_{ie})-1)$ is now based on the elasticity $\varepsilon(s_{ie})$, whose inverse is defined as the weighted average of inverses of the industry CES elasticity of substitution σ_i and of the aggregate economy's Cobb-Douglas elasticity 1.

$$\frac{1}{\varepsilon(s_{ie})} = \frac{1}{\sigma_i}(1-s_{ie}) + s_{ie} \tag{B.4}$$

$$\frac{\varepsilon(s_{ie})}{\varepsilon(s_{ie})-1} = \frac{\sigma_i}{\sigma_i - 1}\frac{1}{1-s_{ie}}. \tag{B.5}$$

Larger establishments charge higher markups:

$$\frac{\partial \frac{\varepsilon(s_{ie})}{\varepsilon(s_{ie})-1}}{\partial s_{ie}} = \left[\frac{1}{\varepsilon(s_{ie})-1} - \frac{\varepsilon(s_{ie})}{(\varepsilon(s_{ie})-1)^2}\right]\frac{\partial \varepsilon(s_{ie})}{\partial s_{ie}} = \frac{\sigma_i - 1}{\sigma_i}\left[\frac{\varepsilon(s_{ie})}{\varepsilon(s_{ie})-1}\right]^2 > 0. \tag{B.6}$$

Working through the model, we show that the establishment size now depends on the

establishment markup:

$$s_{ie} = \frac{P_{ie}Y_{ie}}{P_iY_i} = \frac{\left[\left(\frac{\varepsilon(s_{ie})}{\varepsilon(s_{ie})-1}\right)^{-\alpha_i}\frac{A_{ie}}{(1+\tau_{K_{ie}})^{\alpha_{K_i}}(1+\tau_{L_{ie}})^{\alpha_{L_i}}}\right]^{\frac{1}{\frac{\sigma_i}{\sigma_i-1}-\alpha_i}}}{\sum_{i=1}^{N_i}\left[\left(\frac{\varepsilon(s_{ie})}{\varepsilon(s_{ie})-1}\right)^{-\alpha_i}\frac{A_{ie}}{(1+\tau_{K_{ie}})^{\alpha_{K_i}}(1+\tau_{L_{ie}})^{\alpha_{L_i}}}\right]^{\frac{1}{\frac{\sigma_i}{\sigma_i-1}-\alpha_i}}}. \tag{B.7}$$

To calculate misallocation in this generalized model, we derive the scaling factors with and without distortions. The scaling factors defined by the relative revenue productivity now depend on $\widetilde{MRPX}$, the average marginal revenue products that are scaled by the establishment-specific markups:

$$\frac{\overline{TFPR}_i}{TFPR_{ie}} = \left(\frac{P_iY_i}{P_{ie}Y_{ie}}\right)^{1-\beta}\left[\frac{\widetilde{MRPK}_i}{MRPK_{ie}\frac{\varepsilon(s_{ie})}{\varepsilon(s_{ie})-1}}\right]^{\alpha_{K_i}}\left[\frac{\widetilde{MRPL}_i}{MRPL_{ie}\frac{\varepsilon(s_{ie})}{\varepsilon(s_{ie})-1}}\right]^{\alpha_{L_i}} \tag{B.8}$$

$$\frac{\overline{TFPR}_i}{TFPR_{ie}} = \left(\frac{P_iY_i}{P_{ie}Y_{ie}}\right)^{1-\beta}\left[\frac{\frac{K_{ie}}{P_{ie}Y_{ie}}}{\sum_{e=1}^{N_i}\frac{P_{ie}Y_{ie}}{P_iY_i}\frac{K_{ie}}{P_{ie}Y_{ie}}}\right]^{\alpha_{K_i}}\left[\frac{\frac{L_{ie}}{P_{ie}Y_{ie}}}{\sum_{e=1}^{N_i}\frac{P_{ie}Y_{ie}}{P_iY_i}\frac{L_{ie}}{P_{ie}Y_{ie}}}\right]^{\alpha_{L_i}}, \tag{B.9}$$

where the last expression above is now entirely in terms of data, making it straightforward to implement. In the absence of distortions, we can write the scaling factor as a function solely of the relative size in the absence of distortions $s_{ie}|_{\tau=\bar{\tau}}$:

$$\left.\frac{\overline{TFPR}_i}{TFPR_{ie}}\right|_{\tau=\bar{\tau}} = \left(s_{ie}|_{\tau=\bar{\tau}}\right)^{\alpha_i-1}\left[\frac{\left(1-s_{ie}|_{\tau=\bar{\tau}}\right)}{\sum_{e=1}^{N_i}s_{ie}|_{\tau=\bar{\tau}}\left(1-s_{ie}|_{\tau=\bar{\tau}}\right)}\right]^{\alpha_i} \tag{B.10}$$

$$\text{where } s_{ie}|_{\tau=\bar{\tau}} = \frac{\left[(1-s_{ie}|_{\tau=\bar{\tau}})^{\alpha_i}A_{ie}\right]^{\frac{1}{\frac{\sigma_i}{\sigma_i-1}-\alpha_i}}}{\sum_{i=1}^{N_i}\left[(1-s_{ie}|_{\tau=\bar{\tau}})^{\alpha_i}A_{ie}\right]^{\frac{1}{\frac{\sigma_i}{\sigma_i-1}-\alpha_i}}}. \tag{B.11}$$

C Returns to Scale, Markups, and Concentration

In this appendix we show that—for this class of monopolistically-competitive models—declines in returns to scale and increases in markups have observationally-equivalent implications for the variance and concentration of market shares. This discussion is helpful to relate the modeling and measurement within this paper (and this class of models) to the findings in the literature that larger firms have been capturing larger market shares over time (e.g., Autor et al. (2020), De Loecker et al. (2020)).

For expositional purposes, we present here the analytical results for an undistorted economy. Within the model, we can express an establishment's revenue (market) share s_{ie} as a function of productivity A_{ie} and the difference between the industry markup μ_i and the industry returns to scale α_i:

$$s_{ie} = \frac{P_{ie}Y_{ie}}{P_iY_i} = \frac{(A_{ie})^{\frac{1}{\mu_i-\alpha_i}}}{\sum_{e'\in I}(A_{ie'})^{\frac{1}{\mu_i-\alpha_i}}} \iff \ln s_{ie} = \frac{1}{\mu_i-\alpha_i}\ln A_{ie} - \ln\left(\sum_{e'\in I}(A_{ie'})^{\frac{1}{\mu_i-\alpha_i}}\right).$$

The gap between the markup μ_i and the returns-to-scale parameter α_i determines amplification of an establishment's productivity into its market share. This gap between the markup and the returns to scale is directly informed by the industry's economic profits Π_i:

$$\frac{\Pi_i}{P_iY_i} = 1 - \frac{\alpha_i}{\mu_i}.$$

In this class of models, the extent of profitability translates productivity differences into market shares. The smaller the profit share (i.e., the closer are the markup and returns to scale to each other), the "more competitive" is the industry, in the sense that the most productive firms have a greater market share. A rise in profits would push in the opposite direction: as the gap between the markup μ and the returns to scale α increases, there are fewer competitive pressures and the same productivity advantage leads to a proportionally smaller market share.

A decline in returns to scale increases profits and lowers the variance of market shares:

$$\begin{aligned} \mathbb{V}\text{ar}(\ln s_{ie}) &= \left(\frac{1}{\mu_i-\alpha_i}\right)^2 \mathbb{V}\text{ar}(\ln A_{ie}) \\ \frac{\partial \mathbb{V}\text{ar}(\ln s_{ie})}{\partial(-\alpha_i)} &= \frac{\partial \mathbb{V}\text{ar}(\ln s_{ie})}{\partial \mu_i} < 0 \end{aligned} \tag{C.1}$$

Note, however, that the same implication holds for an increase in the markup. A rise in the markup would drive up industry profits and lower variance in exactly the same way.

These patterns are counterfactual for the aggregate economy since both variance and concentration—measured below by the Herfindahl-Hirschman Index (HHI)—have increased over time. Autor et al. (2020), for instance, is a great example of how *within-industry variation* in either returns to scale or in markups can be used to explain rising concentration. Their 2017 NBER working paper presents a model of monopolistic competition (where all firms charge the same markup, much like in our paper) so that differences in factor shares are driven by differences in returns to scale across firms. The published 2020 paper switches the emphasis: returns to scale are the same across firms while markups now vary across firms. Yet, even as they switched how they model the heterogeneity (without, to our knowledge, providing evidence for one model over the other), the reported patterns of concentration and factor shares remain unchanged.

Having highlighted the theoretical impact of returns to scale and of markups on the dispersion in establishment size in the model, we show nonetheless that these theoretical predictions hold in a relative sense across industries. In the table 7 we regress five-year changes in variance/concentration on the five-year change in the difference between the markup and returns to scale (columns (1) and (3)), and on the markup and the returns to scale separately (columns (2) and (4)).

As predicted by the model's variance expression above, these relative coefficients from columns (1) and (3) show that a larger gap between markups and returns to scale—which reflects rising economic profits—is negatively correlated with changes in the variance and concentration of market shares. Columns (2) and (4) shows that higher markups and lower returns to scale also individually depress the variance and concentration measures.

Table 7: Five-Year Changes in Industry Market Concentration

Dependent Variable	Var of Log Revenue Shares		HHI of Revenue Shares	
	(1)	(2)	(3)	(4)
Markup (μ_i) – Returns to Scale (α_i)	–0.0389		–0.0465	
	(0.0109)		(0.0064)	
Markup (μ_i)		–0.3337		–0.1220
		(0.0588)		(0.0292)
Returns to Scale (α_i)		0.2672		0.1358
		(0.0506)		(0.0242)
Observations	2600	2600	2600	2600
R-squared	0.6649	0.6721	0.9079	0.9071

D Demand Shocks and Misallocation

Our measure of misallocation is based on a counterfactual in which we change distortions but keep fundamentals (e.g., tastes/demand, productivity, etc.) unchanged. We show that our residual-based measure of establishment productivity $\widehat{A_{ie}}$ would conflate productivity A_{ie} and demand in an augmented model where we allow for establishment-specific taste parameters ψ_{ie}. We then show that we would correctly calculate misallocation even when we cannot separately measure productivity and tastes in the residual $\widehat{A_{ie}}$. In short, the measure of misallocation requires us to capture this combined object of productivity and demand; it does not require us to separate the two.

If we allowed for establishment-specific taste parameters, our residual $\widehat{A_{ie}}$ would be a product of the establishment productivity and the taste parameter. We show this by modifying the industry CES aggregator from equation (2) to include establishment-specific taste shifters ψ_{ie}:

$$Y_i = \left(\sum_{e=1}^{N_i} (\psi_{ie} Y_{ie})^{\frac{\sigma_i - 1}{\sigma_i}} \right)^{\frac{\sigma_i}{\sigma_i - 1}}. \tag{D.1}$$

In this augmented model, the demand for an establishment's revenue depends on the consumer's tastes for the variety in question:

$$P_{ie} Y_{ie} = P_i Y_i^{\frac{1}{\sigma_i}} (\psi_{ie} Y_{ie})^{\frac{\sigma_i - 1}{\sigma_i}} \tag{D.2}$$

Following the standard process for backing out the residual $\widehat{A_{ie}}$, we now back out a term that conflates productivity A_{ie} and the taste shifter ψ_{ie}:

$$\widehat{A_{ie}} = \frac{(P_{ie} Y_{ie})^{\frac{\sigma_i}{\sigma_i - 1}}}{K_{ie}^{\alpha_{K_i}} L_{ie}^{\alpha_{L_i}}} = \kappa_i \psi_{ie} A_{ie} \tag{D.3}$$

Since productivity and taste parameters always enter multiplicatively in the expression for misallocation, we would calculate misallocation correctly even though we could not separately measure productivity and demand shocks. We note first that the relative

revenue productivity is unchanged from its expression in the baseline model:

$$\frac{\overline{TFPR}_i}{TFPR_{ie}} = \left(\frac{P_i Y_i}{P_{ie} Y_{ie}}\right)^{1-\beta} \left[\frac{\left(\overline{\frac{K_{ie}}{P_{ie}Y_{ie}}}\right)}{\left(\frac{K_{ie}}{P_{ie}Y_{ie}}\right)}\right]^{-\alpha_{K_i}} \left[\frac{\left(\overline{\frac{L_{ie}}{P_{ie}Y_{ie}}}\right)}{\left(\frac{L_{ie}}{P_{ie}Y_{ie}}\right)}\right]^{-\alpha_{L_i}}. \tag{D.4}$$

When we reallocate inputs to equalize distortions across establishments, the relative revenue productivity now depends on the product $\psi_{it}A_{ie}$:

$$\left.\frac{\overline{TFPR}_i}{TFPR_{ie}}\right|_{\tau=\bar{\tau}} = \left[\frac{\sum_{e=1}^{N_i} (\psi_{ie}A_{ie})^{\frac{1}{\frac{\sigma_i}{\sigma_i-1}-\beta}}}{(\psi_{ie}A_{ie})^{\frac{1}{\frac{\sigma_i}{\sigma_i-1}-\beta}}}\right]^{1-\beta} \tag{D.5}$$

Putting these pieces together, we show that we can calculate misallocation using the residual $\widehat{A_{ie}}$ even when we cannot separately measure productivity and demand shocks:

$$\Phi_i = \frac{TFP_i|_{\tau=\bar{\tau}}}{TFP_i} = \frac{\left[\sum_{e=1}^{N_i} \left((\psi_{ie}A_{ie})\left.\frac{\overline{TFPR}_i}{TFPR_{ie}}\right|_{\tau=\bar{\tau}}\right)^{\sigma-1}\right]^{\frac{1}{\sigma-1}}}{\left[\sum_{e=1}^{N_i} \left((\psi_{ie}A_{ie})\frac{\overline{TFPR}_i}{TFPR_{ie}}\right)^{\sigma-1}\right]^{\frac{1}{\sigma-1}}}. \tag{D.6}$$

E Discrepancies in Establishment-Level Productivity

Incorrect measures of misallocation, both from imposing constant returns to scale and from imposing a common markup, are rooted in spurious correlations between productivity and the distortions that establishments face. As we did in figure 5, we document these spurious correlations in turn, focusing first on returns to scale and then on markups.

In panel A of table 8 we show that inappropriately imposing constant returns to scale leads to measures of productivity that conflate productivity and distortion. The regressions in panel A control for the productivity estimated when returns to scale vary, and compare the constant-returns productivity of establishments with different input bundles. This conditioning allows us to compare establishments that have the same productivity under our model, but that face different distortions, and hence have different input bundles. The key regression coefficients are conditional correlations of constant-returns productivity and input bundles, shown separately for industries with decreasing and increasing returns. As suggested by equation (16), these correlations should be opposite in sign.

Columns 2 and 3 support model predictions that imposing constant returns to scale on industries where returns to scale are not constant leads to predictable spurious correlations between productivity and distortions. Column 2 emphasizes that imposing constant returns in place of decreasing returns leads us to perceive more distorted establishments (i.e., those with smaller input bundles) as more productive. Specifically, a 1-standard-deviation decrease in the log input bundle leads to a measure of productivity that is 0.32 standard deviations larger under the constant returns to scale model. Column 3 emphasizes the opposite pattern for increasing-returns industries. Following a 1-standard-deviation decrease in the log input bundle, productivity is 0.45 standard deviations smaller under constant returns to scale. In this case, imposing constant returns on industries where returns to scale are increasing leads us to perceive more distorted establishments as less productive.

In panel B of table 8, we show that understating the markup in an industry leads us to perceive more distorted establishments as more produtive, while overstating the markup leads us to perceive more distorted establishments as less productive. We document this pattern through the predictions from equation (18) by linking the mismeasurement of productivity to establishment size. In a parallel with panel A, we control for the productivity measured under the estimated markup, and then compare the common-markup productivity of establishments that differ in distortions, and hence in their sizes.

Columns 2 and 3 partition the sample by estimated markup size and back the model predictions. In particular, column 2 suggests that, indeed, understating the markup leads us to a spurious positive correlation between productivity and distortions: a 1-standard-

Table 8: Productivity Mismeasurement at the Establishment Level

Panel A: Imposing Constant Returns to Scale

Dependent Variable	Normalized Log Productivity (A_{ie}) (Constant Returns to Scale)		
	(1)	(2)	(3)
Normalized Log Input Bundle	0.1460	-0.3238	0.4465
	(0.0149)	(0.0147)	(0.0106)
Normalized Log Productivity (A_{ie})	0.8241	1.0527	0.7528
(Variable Returns to Scale)	(0.0068)	(0.0067)	(0.0074)
Industry-Year Sample	All	Decreasing RTS	Increasing RTS
Industry×Year Fixed Effects	Yes	Yes	Yes
Observations	292000	126000	166000
R-squared	0.8130	0.9268	0.9338

Panel B: Imposing a Common Markup across Industries

Dependent Variable	Normalized Log Productivity (A_{ie}) (Common Markup)		
	(1)	(2)	(3)
Normalized Log Value Added	0.2514	-2.0286	0.5993
	(0.0243)	(0.0566)	(0.0137)
Normalized Log Productivity (A_{ie})	0.4839	2.3961	0.4678
(Heterogeneous Markups)	(0.0155)	(0.0436)	(0.0104)
Industry-Year Sample	All	Understated Markup	Overstated Markup
Industry×Year Fixed Effects	Yes	Yes	Yes
Observations	292000	116000	176000
R-squared	0.5630	0.7099	0.9143

Note: Unit of observation is an establishment-year. The time period comprises 1982, 1987, 1992, 1997, 2002, and 2007. Standard errors are clustered at the industry-year level. To normalize the values within each industry, we demean the variable and divide by its standard deviation.

deviation decrease in size (i.e. an increase in distortion) leads us to a 2.03-standard-deviation increase in common-markup productivity. Column 3 presents the opposite result for instances where we overstate the markup: a decrease in size leads to a 0.60-standard-deviation decrease in common-markup productivity.

F Measurement Error in Bils, Klenow and Ruane (2017)

Bils, Klenow and Ruane (2017), henceforth BKR, highlight the possibility that measurement error could be misinterpreted as misallocation in microdata. They propose a correction for *additive* measurement error in establishment revenue R and input bundles I. Their estimates suggest that measurement error has increased in U.S. Census microdata, and that accounting for this change in measurement error eliminates the upward trend in misallocation from a gross-output Hsieh-Klenow model.

In this appendix, we show that deviations from constant returns to scale look like measurement error in the BKR procedure, and that a decline in returns to scale over time looks like an increase in measurement error. Informally, our argument emphasizes two points. First, a procedure that does not explicitly account for *multiplicative* measurement error will pick up this multiplicative measurement error as *additive* measurement error. Second, overlooking deviations from constant returns to scale leads to multiplicative measurement error in the input bundle. For instance, if the true returns to scale in an industry were α_i, then the input bundle under constant returns to scale $I_{crts,ie}$ would relate to the true input bundle I_{ie}, as shown below. As a result, the BKR procedure could interpret deviations from constant returns to scale as measurement error.

$$I_{crts,ie} = \underbrace{I_{ie}^{\frac{1-\alpha_i}{\alpha_i}}}_{multiplicative\ measurement\ error} I_{ie}$$

$$I_{crts,ie} = I_{ie} + \underbrace{[I_{ie}^{\frac{1-\alpha_i}{\alpha_i}} - 1]I_{ie}}_{additive\ measurement\ error}.$$

Formally, we focus on the key parameter λ in BKR estimating equation [2], reproduced below. BKR show that $\lambda = 1$ if there is no misallocation. Larger deviations from unity indicate a greater extent of measurement error. The key estimating equation relates the time-series change in revenue, R, to the change in the input bundle I, the revenue productivity $TFPR$, and the product of I and $TFPR$. Both I and $TFPR$ depend on the assumed returns to scale. The measure of change Δ defined as the "growth rate of a plant variable relative to the mean of its sector."

$$\Delta\widehat{R} = \Psi \cdot \Delta\widehat{I} + \Phi \cdot f(\ln TFPR) + \Psi(1-\lambda) \cdot \Delta\widehat{I} \cdot g(\ln TFPR).$$

We derive the below relationship between λ_{crts}, estimated under assumed constant returns to scale, and true λ, where $g(\cdot)$ is some polynomial. In short, the BKR procedure correctly captures measurement error under one of two conditions: either $\lambda = 1$, so there

is no measurement error, or $\alpha_i = 1$, so that the assumed constant returns to scale hold in the data. As a result, if there is any measurement error in the data, the BKR estimates can conflate measurement error with model misspecification.

$$\lambda_{crts} = \lambda + (1-\lambda)[1-\gamma] \text{ where } \gamma = \frac{g\left(\ln R_{ie} - \alpha_i \ln I_{crts,ie}\right)}{g\left(\ln R_{ie} - \ln I_{crts,ie}\right)}.$$

We show that the mismeasurement of λ varies predictably with returns to scale α_i. Since $\widehat{\lambda}_{crts}$ in BKR takes values between 0.095 and 0.358 for U.S. data, we focus entirely on the case of $\lambda < 1$. In short, if λ_{crts} is closer to 1 than is λ, then we understate measurement error when we impose constant returns to scale. Indeed, this is the case when returns to scale are increasing: when $\alpha_i > 1$, then $1 > \lambda_{crts} > \lambda$, and we understate misallocation. By contrast, when returns to scale are decreasing, then we overstate misallocation, since $\alpha_i < 1$ leads to $1 > \lambda > \lambda_{crts}$.

Consider a setting in which measurement error does not change over time, but returns to scale decline from increasing to constant; imposing constant returns to scale in this setting would lead us to infer an increase in measurement error, even though no such increase has taken place. As detailed in the previous paragraph, overlooking increasing returns to scale leads us to understate measurement error. As returns to scale decline over time, our estimate of measurement error asymptotes to its true value from below. In short, imposing constant returns to scale here would lead us to understate measurement error early in the period and to see this measurement error grow toward its true value over time. However, by assumption, true measurement error has not changed; we only see it grow as the bias from imposing constant returns declines over time.

With the caveat that our estimates of returns to scale are for a value-added world, while BKR work in a gross-output world, we present the BKR estimates of λ_{crts} and our estimates of returns to scale α_i. By the arguments above, it is possible that a decline in returns to scale could explain the increase in measurement error over time that BKR find. If, as a result, there has not been a substantial change in measurement error over time, then measurement error is less capable of explaining the upward trend in misallocation.

Table 9: U.S. Manufacturing – Division of Value Added

	1978-1982	1983-1987	1988-1992	1993-1997	1998-2002	2003-2007
λ_{crts}	0.358	0.336	0.326	0.326	0.192	0.095
$\alpha_{average}$	1.23	1.20	1.20	1.12	1.11	0.96