No. 2022-006

DECOMPOSING THE RISE IN MARKUPS

By

Has van Vlokhoven

1 March 2022

ISSN 0924-7815
ISSN 2213-9532
Decomposing the Rise in Markups

Has van Vlokhoven*

February 24, 2022

Abstract

A recent literature argues that the rise in aggregate markups is due to economic activity reallocating toward high-markup firms rather than an increase in firm-level markups. I show that this result is biased by three types of measurement error. First, standard decomposition methods estimate a positive reallocation effect even when the allocation does not change. Second, firm-level markups are mismeasured when the production function is misspecified. If this is the case, the decomposition can give opposite results depending on whether actual or measured markups are used. Third, the decomposition is sensitive to measurement error of the variables sales and costs. Correcting for each type of measurement error suggests that the reallocation effect is smaller than previously thought and perhaps even zero or negative. Moreover, correcting for measurement error of sales lowers the estimated sales-weighted average markup, and I find that the estimated increase in the sales-weighted markup over the last 60 years in Compustat is in its entirety due to a rise in measurement error. The cost-weighted average markup is not affected by measurement error and is therefore a more robust estimate of market power.

Keywords: Markups, Market Power, Decomposition, Reallocation

JEL codes: D22, E25, L11, O40

*Tilburg University, h.vanvlokhoven@tilburguniversity.edu. I thank Joseph Stiglitz for useful comments.
Aggregate markups (De Loecker et al., 2020) and profits (Barkai, 2020; Van Vlokhoven, 2019) have been increasing over the last decades. In order to understand what is driving this rise in market power and what type of policy response is appropriate it is important to consider how the underlying firm-level distribution of markups has changed over time. For instance, is it the case that most firms have experienced a rise in markups or is the rise in markups due to reallocation of economic activity toward firms with a high markup? If it turns out to be the case that most firms have increased their markup, then this suggests that there is more anti-competitive behavior in general. In this case, stricter antitrust laws might be suitable. If, on the other hand, the rise in market power is due to some high-markup firms becoming larger because they have made efficiency gains then stricter antitrust laws might not be welfare improving.\(^1\)

The aggregate markup, \(\mu_t\), at time \(t\) is the weighted average of firm-level markups \(\mu_{it}\), where the weights are denoted by market shares \(m_{it}\): \(\mu_t = \sum_i m_{it} \mu_{it}\). The market share could, for instance, be based on costs, sales or value added. Several methods have been proposed to decompose the change in an aggregate variable, such as the markup or productivity, in a within firms effect and a reallocation between firms effect.\(^2\) One frequently used decomposition is a Haltiwanger (1997) decomposition. This decomposes the change in markups, \(\Delta \mu_t = \mu_t - \mu_{t-1}\), in the following three components,\(^3\)

\[
\Delta \mu_t = \underbrace{\sum_i m_{it-1} \Delta \mu_{it}}_{\text{Within}} + \underbrace{\sum_i \mu_{it-1} \Delta m_{it}}_{\text{Between}} + \underbrace{\sum_i \Delta m_{it} \Delta \mu_{it}}_{\text{Covariance}}.
\] (1)

The first component denotes the change in the aggregate markup when market shares would not have changed, but only firm-level markups had changed. This is called the within effect. The second component is the between effect and gives the change in markups if only market shares would have changed while holding markups at the firm level constant. The final component is the covariance between changes in market shares and markups. The covariance effect is positive when the firms that are experiencing an increase in their markup also tend to experience an increase in their market share and vice versa. The sum of the between effect and the covariance effect is usually referred to as the reallocation effect. De Loecker et al. (2020) perform such a decomposition of the markup for the United States using Compustat data. They find that the reallocation effect is positive and large, while the within effect is close to zero. At first sight, this suggests that although aggregate markups have increased substantially, markups at the firm-level have only increased modestly.

\(^1\)Obviously, if the reallocation channel dominates this does not necessarily mean that this is due to efficiency gains. It could also be the case that larger companies have become better at anti-competitive behavior. An example of anti-competitive behavior that favors larger firms is suing smaller firms for patent infringement.

\(^2\)I focus on markups. See Hsieh and Klenow (2017) and Nishida et al. (2017) for a discussion on how reallocation has affected productivity growth.

\(^3\)See also Baily et al. (1992), Griliches and Regev (1995) and Foster et al. (2001).
In this paper, I show that such a decomposition is plagued by three types of measurement error, and each type of measurement error leads to a positively biased reallocation term. Correcting for each type of measurement error I find that the reallocation effect is substantially smaller than what is found by De Loecker et al. (2020) and is perhaps even zero or negative.

The first type of measurement error is that the Haltiwanger decomposition gives a positive reallocation effect even when the joint distribution of markups and market shares is not changing. Thus, finding a positive reallocation effect does not necessarily imply that the allocation has changed. To see why this is the case, consider an economy that is in a steady state such that the joint distribution of markups and market shares is constant over time, but with firm-level markups and market shares changing over time. Furthermore, assume there is a positive relationship between market shares and markups at the firm level. Due to the stationarity of the distribution, the relatively smaller firms tend to increase their market shares, and therefore, these smaller firms tend to experience an increase in their markups due to the positive relationship between the two. For the same reason, larger firms tend to experience a decline in their markup. Hence, the within effect is negative. As the aggregate markup is not changing, the negative within effect implies a positive reallocation effect, even though the allocation is not changing. In order to understand how large this bias is I construct a counterfactual that gives the magnitude of the within and reallocation effects when the economic environment does not change over time. This counterfactual depends on the size of the firm-level shocks and the strength of the relationship between market shares and markups. For instance, if there were no changes in market shares the counterfactual would give a zero within effect and a zero reallocation effect, and the larger the changes in market shares the larger the effects in absolute value. To construct a model-based counterfactual, I solve a model that matches the relationship between market shares and markups, and that matches how firms move up and down in the sales distribution over time. I find that over a 35-year period the cumulative within effect would have been around -0.06, and the reallocation effect would have been around 0.06, if the macroeconomic environment would not have changed. These are large effects as the markup in 1980 is around 1.2. De Loecker et al. (2020) find that between 1980 and 2015 the within effect is about 0.03 and the reallocation effect is about 0.2 for the sales-weighted markup. Taking these estimates at face value it seems that the reallocation effect is indeed dominant. However, relative to the counterfactual, the gap between the within and reallocation effect is much smaller. Relative to the counterfactual the reallocation effect dominates the within effect by 0.05 instead of 0.17. Thus, after correcting for this first type of measurement error the within effect and reallocation effect are about equally much associated with the rise in markups, although the reallocation effect still seems to weakly dominate the within effect.

When constructing this counterfactual I assume that markups are observed, but in

---

4There is ample empirical evidence for a positive relationship between markups and market shares. See, for instance, Edmond et al. (2015, 2018), Brooks et al. (2016), Kikkawa et al. (2019) and Burstein et al. (2020).
practice we do not observe markups and they have to be estimated first. The typical approach to measure markups nowadays is the production approach that uses that the markup is equal to the output elasticity times the inverse revenue share of that respective input (De Loecker and Warzynski, 2012; Hall, 1988). The challenge lies in estimating the output elasticity for which a production function needs to be estimated first. Bond et al. (2021) show that standard approaches to estimate a production function yield severely biased estimates of the aggregate markup. Moreover, it is even more challenging to consistently estimate each term of the decomposition. It is not sufficient to estimate the average output elasticity consistently, but one needs to estimate firm-specific output elasticities consistently. Unmeasured heterogeneity in output elasticities is the second source of mismeasurement that affects the decomposition. I show that when the production function is misspecified and there is heterogeneity in factor-augmenting technology such that firm-level output elasticities and markups are mismeasured the decomposition can give opposite results depending on whether the estimated or true firm-level markups are used. To test whether output elasticities do indeed differ across firms, I estimate the output elasticity separately for each decile of the firm size distribution within Compustat. Consistent with the model, I find that output elasticities are increasing in firm size. Furthermore, changing the production function to allow for more heterogeneity in output elasticities than in the baseline lowers the size of the reallocation effect.

The third type of measurement error that affects the decomposition is measurement error of sales and costs. The data suggests that measurement error of both variables has increased over time in Compustat. I correct for measurement error of sales by using the first stage of the production function estimator as developed by Ackerberg et al. (2015). Doing so changes the decomposition substantially. After correcting for measurement error, the reallocation effect is close to zero and the within effect dominates. If, in addition to correcting for measurement error, we would also compare these results to a counterfactual the true reallocation effect might very well be negative.

Knowing whether the within effect or reallocation effect dominates is not only important for understanding what is causing the change in market power. It is also crucial for understanding the welfare consequences of the change in markups. A rise in markups is not necessarily detrimental to welfare. If a rise in markups is due to economic activity shifting to firms that have a high markup this might be welfare-enhancing as the high markup firms might be too small to begin with (Baqee and Farhi, 2020). For an economy in which the macroeconomic environment does not change over time, Baqee and Farhi (2020) provide the following formula to determine the change in allocative efficiency

$$\Delta \text{Allocative Efficiency} = - \sum_i \hat{\lambda}_{it-1} \Delta \log \mu_{it} - \Delta \text{Factor Income Shares},$$

where $\hat{\lambda}$ are cost-based Domar weights. Based on this formula and using Compustat data, Baqee and Farhi (2020) find that improvements in allocative efficiency due to the reallocation of market share to high-markup firms account for about a 7 percentage
point increase in aggregate TFP over the period 1997-2015. However, the first part of this formula looks very similar to the negative of the within effect in the Haltiwanger decomposition. Indeed, suppose that there are no input-output linkages such that the Domar weights are equal to the sales share. Then this term is equal to the negative of the (sales-weighted) within effect except that the change in markups is replaced by the change in the log markup. Thus, using the same argument as in the above, again considering an economy with no change in the allocation such that also the factor income shares do not change, allocative efficiency would increase over time according to the above formula. Turning back to the stationary counterfactual, I find that over an 18-year period the above measure of allocative efficiency would have increased by about 2 percentage points. Thus, this has the potential to explain a third of the 7 percentage point rise that is found by Baqee and Farhi (2020). This suggests that allocative efficiency might have improved, but by substantially less than 7 percentage points. The reason why I find different results than Baqee and Farhi (2020) is that their formula is based on exogenous markups while I consider endogenous markups. In addition to these results, when I correct for measurement error of sales as in the above I find that allocative efficiency has only increased by three percentage points between 1997 and 2015. Thus, this measured change in allocative efficiency is only one percentage point larger than the reallocation effect we would find for an economy in which the allocation is not changing.

A different type of decomposition that is often used as well is an Olley and Pakes (1996) decomposition. This decomposition does not look at changes over time within a firm, but looks at how moments of the cross-sectional distribution are changing over time,

\[ \Delta \mu_t = \Delta \overline{\mu}_t + \Delta \sum_i (m_{it} - \overline{m}_t) (\mu_{it} - \overline{\mu}_t), \]

where \( \overline{\cdot} \) denotes the unweighted average. This decomposition decomposes the change in the aggregate markup in the change in the unweighted average markup (within) and the change in the covariance between market shares and markups (reallocation). When the unweighted average markup is increasing, the within effect is positive, and the reallocation effect is positive when the covariance is increasing. One advantage of this decomposition is that when the joint distribution of markups and market shares is constant, both the within and reallocation effect are zero, as desired. However, following Melitz and Polanec (2015), a version of this decomposition that is frequently used takes entry and exit into account. For instance, Autor et al. (2020) use this adjusted decomposition to study the fall in the labor share over time and find that the fall in the labor share is mainly associated with a negative reallocation effect. The extension taking entry and exit into account means in practice that to calculate the within and reallocation effects between two consecutive periods only firms that are active in both periods are considered. This implies that also for this decomposition the within effect can be negative when the joint distribution of markups and market shares does not change over time. This is the case when entering
firms have on average a higher markup than exiting firms. Namely, for the markup distribution to be invariant in such an environment it must be that continuing firms experience a drop in their markup on average, and hence, the within effect is negative. Depending on how market shares are distributed among firms, the reallocation effect can be positive or negative. Thus, also for the Olley-Pakes-Melitz-Polanec (OPMP henceforth) decomposition a non-zero reallocation effect does not necessarily imply that the joint distribution of markups and market shares has changed. Moreover, it could be that the within term is negative while the unweighted average markup is increasing, and that the reallocation effect is positive while the covariance is declining over time. Empirically, this turns out to be the case in Compustat. Although the OPMP decomposition suggests otherwise, the unweighted average markup has been increasing while the covariance between market shares and markups has been declining. The rise in the aggregate markup is equivalent to the rise in firm-level markups. This is another indication that reallocation is not the major driver of the rise in markups among Compustat firms, but that the rise in markups is due to an increase in firm-level markups.

Finally, there is some discussion in the literature on what the appropriate weights are to construct aggregate markups. One alternative is to use market shares based on costs and another alternative is to use sales shares. De Loecker et al. (2020) find that the sales-weighted markup has increased substantially while the cost-weighted markup has only increased modestly. Edmond et al. (2018) argue that the cost-weighted markup is more relevant when studying welfare. I argue that there is another benefit of studying the cost-weighted markup over the sales-weighted markup. Namely, the sales-weighted markup is heavily influenced by measurement error while the cost-weighted markup is not. Suppose there is measurement error in sales. Firms for which sales is measured to be too high will also have a too high estimated markup. For the sales-weighted markup, the positively biased markups are over-weighted leading to a positively biased measure of the aggregate markup. Instead, the cost-weighted markup is not sensitive to idiosyncratic measurement error as it is equal to total sales divided by total costs times the output elasticity, and is therefore a more robust estimate of market power. When I correct for measurement error of sales I find that the corrected sales-weighted markup has not increased over the last 60 years. Thus, the estimated increased sales-weighted markup in Compustat is in its entirety due to a rise in measurement error.

The remainder of this paper is organized as follows. In Section I I construct a counterfactual that gives how large the Haltiwanger reallocation effect is when the joint distribution of markups and market shares is not changing over time. Section II decomposes the change in the markup using Compustat data and Section III discusses how mismeasurement of the markup affects the decomposition results. In Section IV I discuss the Olley-Pakes decomposition and Section V concludes.
I Haltiwanger decomposition

With a Haltiwanger decomposition as in equation (1) it could be that the reallocation effect is positive and the within effect is negative even when the joint distribution of firm-level markups and market shares is not changing. This is the case when there is a positive relationship between markups and market shares. That the distribution does not change over time does not mean that firms cannot grow or shrink, or change their markup, but implies that if some firms are growing some other firms are shrinking and likewise for the markup. The positive relationship between market shares and markups makes that the covariance effect is positive. Since the aggregate markup does not change, this means that the within effect and/or the between effect has to be negative. It turns out that both are negative in this economy. The stationary distribution plus the positive relationship between markups and market shares imply that firms that experience an increase in their markup are relatively small, while those that experience a decrease in their markup are relatively large. This leads to a negative within effect. Likewise, the between effect is negative because firms that experience an increase in their market share have a relatively low markup. That the within effect is negative implies that the overall reallocation effect is positive.

Proposition 1 (Haltiwanger Decomposition). Suppose that the joint distribution of markups and market shares is constant, and that there is a positive relationship between markups and market shares. Then, as firms move up and down in the market share distribution, the reallocation term in the Haltiwanger decomposition is positive and the within term is negative.

Thus, an empirically found positive reallocation effect does not necessarily imply that the allocation has changed over time. Likewise, an empirically found negative within effect does not necessarily imply that markups at the firm-level have fallen. The finding that the reallocation component of the Haltiwanger decomposition is positive in this stationary economy is correct in the sense that if market shares would not have changed over time, the aggregate markup would have declined, due to the change in firm-level markups which tend to be negative for larger firms in this model-economy. However, this interpretation misses that markups and market shares might be endogenous, and the positive reallocation effect cannot be used to conclude that the allocation has changed, simply because the allocation has not changed in this economy.

Proposition 1 relies on the assumption that there is a positive relationship between market shares and markups. If there would be no relationship between the two, then the reallocation term and within term would both be zero in a stationary environment. When the relationship is negative, a negative reallocation term and a positive within term would appear instead. A positive relationship between markups and market shares is found empirically in a variety of settings. See, for instance, Edmond et al. (2015, 2018), Brooks et al. (2016), Kikkawa et al. (2019) and Burstein et al. (2020). Furthermore, this assumption holds naturally in several models of firm dynamics. In Appendix A I consider
Cournot and Bertrand competition à la Atkeson and Burstein (2008) and show that a positive relationship between market shares and markups occurs both when changes to market shares are driven by productivity shocks and demand shocks. In the main text I study a monopolistic competition model with a Kimball (1995) aggregator in which this assumption also holds.

In the remainder of this section I will use the Kimball model to construct a counterfactual. This counterfactual quantifies how large the Haltiwanger within and reallocation terms are in the absence of any true reallocation. This counterfactual is helpful in assessing how large the empirically observed reallocation and within effects truly are. If, for instance, the empirically observed reallocation effect is smaller than the counterfactual reallocation effect this suggests that the true reallocation effect is negative. The reason why I study the monopolistic competition model in the main text instead of an oligopoly model is that I want to relate my results to the literature on estimating production functions, and therefore I treat capital as a state variable. A model in which capital is dynamic remains tractable under monopolistic competition but not under oligopolistic competition.\footnote{Under an oligopoly the decision of a firm on how much to invest depends on the capital stock and productivity level of its competitors, such that these become state variables. Hence, already with a handful of firms in a market this optimization problem becomes untractable.}

In Appendix A I study oligopoly when capital is a static input and find similar results as in the main text.

Model Suppose there is a continuum of varieties of mass 1 indexed by \( i \) and that the final good, \( Y \), is produced by a competitive firm according to the following Kimball aggregator

\[
\int_0^1 \chi_i \Upsilon \left( \frac{Y_i}{Y} \right) di = 1 ,
\]

where \( \Upsilon(q) \) is strictly increasing, strictly concave and satisfies \( \Upsilon(1) = 1 \). The Kimball aggregator nests the familiar CES aggregator with \( \Upsilon(q) = q^{1-1/\sigma} \). \( \chi_i \) is a preference shifter and \( Y_i \) denotes output of variety \( i \). Taking input prices as given, the first-order condition of the competitive firm gives the following inverse demand function for variety \( i \)

\[
P_i = \chi_i \Upsilon' \left( \frac{Y_i}{Y} \right) PY \left( \int_0^1 \chi_i Y_i \Upsilon' \left( \frac{Y_i}{Y} \right) di \right)^{-1} ,
\]

where \( D \) is a demand index. Each firm is the monopoly producer of a variety \( i \) with associated cost function \( C_i(Y_i) \). Within a period the firm chooses its output and price level to maximize flow profits \( \Pi_i = \max_{P_i, Y_i} P_i Y_i - C_i(Y_i) \) subject to (5). This gives the following first-order condition

\[
P_i = \frac{1}{1 + \frac{Y_i}{Y} \Upsilon'' \left( \frac{Y_i}{Y} \right) \frac{C_i'(Y_i)}{Y' \Upsilon' \left( \frac{Y_i}{Y} \right)}}.
\]
I use the Klenow and Willis (2016) specification for $\Sigma$, such that the markup becomes

$$\mu_i = \frac{\sigma \left( \frac{\Sigma_i}{\Sigma} \right)^{-\varepsilon/\sigma}}{\sigma \left( \frac{\Sigma_i}{\Sigma} \right)^{-\varepsilon/\sigma} - 1}.$$  

As already anticipated there is a positive relationship between market shares and markups, which is governed by the super-elasticity $\varepsilon$. The elasticity of demand of the average firm is denoted by $\sigma$. When the super-elasticity is zero this model equals the standard CES monopolistic model with a constant markup $\frac{\sigma}{\sigma - 1}$.

To close the model I need to specify the production function. Production is assumed to follow a Leontief between materials $M_i$ and a Cobb-Douglas aggregate of capital, $K_i$, and labor, $L_i$:

$$Y_i = \min \{ A_i K_i^{\alpha} L_i^{1-\alpha}, \alpha M_i \}.$$ \hspace{1cm} (6)

The reason why I use a gross output production function and thus have materials as an input is that this allows me to relate the model to Compustat data in which we only observe sales and a bundle of inputs consisting of labor and materials (i.e., cost of goods sold). A Leontief production function is reasonable for a gross output production function and is also commonly used in the literature (e.g., Ackerberg et al. (2015)) although an elasticity of substitution of 0 is of course a stark assumption. Productivity $A_i$ takes the form of factor-augmenting technology and not as Hicks-neutral technology which is consistent with the findings in Doraszelski and Jaumandreu (2018), David and Venkateswaran (2019), Raval (2019, 2020) and Demirer (2019). In Section III I show that allowing for heterogeneity in factor-augmenting technology can explain some of the patterns we observe in the data. These assumptions on the production function do not drive the results in this section. Similar results are obtained when studying a Cobb-Douglas production function in capital and labor and having Hicks-neutral technology differences.

Materials and labor are variable inputs that can be chosen within the period whereas the capital stock is chosen during the previous period. This leads to the following recursive profit maximization problem over the choice of investment $I$:

$$V(K_i, A_i) = \max_{I_i} \Pi_i - PI_i + \beta EV(K_i', A_i')$$

s.t. $K_i' = (1 - \delta)K_i + I_i$

---

The Klenow-Willis specification is the following

$$\Upsilon(q) = 1 + (\sigma - 1) \exp(1/\varepsilon) \varepsilon^{-1} \left[ \Gamma \left( \frac{\sigma}{\varepsilon}, \frac{1}{\varepsilon} \right) - \Gamma \left( \frac{\sigma}{\varepsilon}, \frac{q^{\varepsilon/\sigma}}{\varepsilon} \right) \right],$$

with $\sigma > 1, \varepsilon \geq 0$ and where $\Gamma(s, x)$ denotes the upper incomplete Gamma function: $\Gamma(s, x) = \int_x^\infty t^{s-1}e^{-t}dt$.

Market shares are here defined in terms of real output shares while in the Haltiwanger decomposition market shares refer to sales shares. Edmond et al. (2018) show that there is also a positive relationship between sales shares and markups, namely: $1/\mu_i + \log(1 - 1/\mu_i) = \frac{x}{\varepsilon} \log(m_i) + \text{constant.}$
\[
\log(A_i') = \rho \log(A_i) + \upsilon_i,
\]
where \(\Pi_i\) is the flow of profits depending on capital and productivity and which follows from the static optimization choice of choosing prices, output, labor and materials. Next period values are denoted by a prime. The logarithm of productivity follows an AR(1) process. Capital follows a standard accumulation equation with depreciation rate \(\delta\). The discount rate is denoted by \(\beta\). Finally, the materials and investment good producing sectors are assumed to be producing according to a perfectly elastic supply curve, and I normalize the price of materials and of the investment good to 1.

An equilibrium of this model consists of a set of firm-level prices, and quantities of output, materials and labor that are consistent with the static firm-optimization problem given capital, productivity and aggregate demand. Furthermore, the distribution of capital and productivity is consistent with the dynamic optimization problem and the law of motion for productivity. Total output is such that the labor market clears where labor supply is normalized to 1 and such that it is consistent with equation (4). See Appendix C for the solution algorithm.

Constructing a counterfactual
Considering the Haltiwanger decomposition, the size of the within and reallocation effect in an economy in which the joint distribution of markups and market shares is not changing depends on the relationship between markups and market shares and on the size of the firm-level shocks. The super-elasticity \(\varepsilon/\sigma\) governs the relationship between markups and market shares. Using the US Census of Manufactures Edmond et al. (2018) estimate this super-elasticity to be equal to 0.16 within 6-digit industries, and I calibrate my model to match this value.\(^8\) To match churning within the firm size distribution I set the standard deviation of the normally distributed productivity shocks \(\upsilon_i\) to 0.24 and the autocorrelation coefficient \(\rho\) to 0.99. This ensures that estimating an AR(1) process for sales in the simulated data gives the same coefficients as in Compustat. Namely, an autocorrelation coefficient of 0.98 and the variance of the error term is equal to 0.17. An alternative way of matching churning in the data is to directly target the standard deviation of changes in market shares. This is the relevant statistic as in a stationary environment the covariance term in the Haltiwanger decomposition is equal to \(\text{corr}(\Delta m_i, \Delta \mu_i) \times \text{std}(\Delta P_i Y_i) \times \text{std}(\Delta \mu_i)\) where \(P_i Y_i\) denotes average sales. In the data, the standard deviation of changes in sales over average sales (i.e., market share times the number of firms) equals 0.69. In the calibrated model targeting an AR(1) for sales this standard deviation is lower and equals 0.19. Thus, in that sense the calibration is conservative as instead targeting this standard deviation directly would lead to a larger counterfactual.\(^9\) Another reason for not targeting this standard deviation directly is that there is likely to be measurement error in the data that could affect this standard deviation.

---

\(^8\)See Section III for a discussion on why I do not estimate this super-elasticity directly within Compustat.

\(^9\)To get a counterfactual without solving a model one could also directly use this formula for the covariance term and plug in values for the correlation between changes in markups and market shares, and the standard deviations of these two terms. I choose to solve a model as this turns out to be helpful when discussing measurement error of output elasticities and variables in the data.
substantially. Here I assume that all variation across firms comes from productivity differences and set the preference shifter \( \chi_i \) to 1 for all firms. In Appendix A I do the opposite and assume that all variation comes from the preference shifter. This leads to a similar counterfactual.

The remaining calibration is summarized in Table 1. I set the the demand elasticity \( \sigma \) to 4 to target the sales-weighted aggregate markup and set \( \varepsilon \) to 0.64 to get a super-elasticity \( \varepsilon/\sigma \) of 0.16. I set the depreciation rate \( \delta \) to 0.1, which is equal to the average depreciation rate during the period 1980-2015. I set the discount rate \( \beta \) to 0.92 to get a cost of capital of 0.19 as I find it to be in Van Vlokhoven (2019) for this time period.\(^{10}\) I set \( \alpha_M \) to 0.88 to target a material share of sales of 50% as it is in the data (De Loecker et al., 2020), and I set the capital share \( \alpha \) to 0.31. In Section III it becomes clear which variable I target when calibrating \( \alpha \).

Figure 1 shows the resulting cumulative decomposition of the sales-weighted average markup over a 35-year horizon. The cumulative reallocation effect is 0.06 after 35 years, while the within effect is -0.06. Thus, although the allocation and the average markup do not change over time, this decomposition gives that if economic activity would not have been reallocated to higher markup firms, the markup would have fallen from 1.38 to 1.32. These numbers are substantial. De Loecker et al. (2020) find that within Compustat data between 1980 and 2015 the cumulative reallocation effect is about 0.2 and the within effect is about 0.03. Thus, compared to this counterfactual the within effect is equal to 0.09 \( (= 0.03 - (-0.06)) \) and the reallocation effect is equal to 0.14 \( (= 0.2 - 0.06) \). It is still the case that the reallocation effect dominates the within effect, but the gap is much smaller than what is suggested by these numbers at face value. Moreover, if shocks to firm-level productivity would be set higher in the calibration, then the counterfactual reallocation effect could dominate the observed reallocation effect.

Table 3 in Appendix A shows that the size of the counterfactual reallocation effect does not depend much on whether we consider the sales-weighted or cost-weighted markup. The same is true for the within effect as this is always the negative of the reallocation effect due to the aggregate markup not changing. Also, whether firm heterogeneity

---

\(^{10}\) In this model, the cost of capital is equal to \( \delta + 1/\beta - 1 \).
When considering an oligopolistic model with Cournot competition the counterfactual reallocation effect becomes somewhat larger.

Turning to welfare, using the formula provided by Baqee and Farhi (2020) and assuming there are no input-output linkages, the counterfactual change in allocative efficiency is similar, though somewhat smaller in magnitude, to the sales-weighted reallocation effect. The counterfactual cumulative change in allocative efficiency over a 35-year horizon would be equal to 0.04 when the allocation is not changing. In their empirical application, Baqee and Farhi (2020) find that allocative efficiency has increased by 7 percentage points over an 18-year period. The model-based stationary counterfactual of 2 percentage points over such a time period can thus explain a third of this found rise. Thus, according to this exercise allocative efficiency seems to have increased but likely by a number closer to five percentage points than seven percentage points.

II Decomposition in Compustat

The previous section has shown that the true within effect is likely to be larger than the empirically found within effect while the true reallocation effect is likely to be less than the empirically found reallocation effect. However, it is important to stress that these counterfactuals are model-based, and there might be important elements missing from the model. Here and in the next section, I study how robust the finding of a dominant reallocation effect is in the data.

Following De Loecker et al. (2020) I decompose the rise in the markup between 1980 and 2015 for the United States using Compustat data, which comprises mainly publicly listed firms. The markup is obtained from De Loecker et al. (2020). They estimate a time-varying output elasticity of an input bundle by industry by assuming a Cobb-Douglas over capital and this input bundle. The input bundle consists of materials and labor directly used in production, and is referred to as cost of goods sold. The markup is then
simply this output elasticity multiplied by sales over cost of goods sold. I follow the same data cleaning steps as De Loecker et al. (2020) do.\textsuperscript{11}

Compared to the above decomposition, there is an additional term that accounts for entry and exit into the data,

\begin{align}
\Delta \mu_t &= \sum_{t \leq t-1} m_{it} \Delta \mu_{it} + \sum_{t \leq t-1} \mu_{it} \Delta m_{it} + \sum_{t \leq t-1} \Delta m_{it} \Delta \mu_{it} \\
&+ \sum_{t \leq t-1} m_{it} (\mu_{it} - \mu_{t-1}) - \sum_{t \leq t-1} m_{it-1} (\mu_{it-1} - \mu_{t-1}) .
\end{align}

(7)

The net entry effect compares the markup of entering and exiting firms with the aggregate markup.\textsuperscript{12} Moreover, this decomposition lumps firms across all industries together. I do the following to disentangle between reallocation across industries and reallocation across firms within an industry. First, I calculate the within effect and reallocation effect among 2-digit naics industries using the above formula, where firm index $i$ is replaced by industry index $j$. Then, I use the above formula to decompose the change in the industry-level markup for each industry. To get the total (firm) reallocation, within and net entry effects I aggregate across industries using the weights $m_{jt-1}$. The first panel of Figure 2 shows the resulting decomposition of the sales-weighted markup as reported by De Loecker et al. (2020) and this is also what is referred to in the above. During this 35-year period the sales-weighted markup has increased by around 0.3 and this is not due to reallocation across industries. Considering the within firm and reallocation across firms effects within industries I find the following. The cumulative within effect is going up and down. If it were for the within effect alone the markup would have increased by less than 0.05. Instead, the reallocation term is found to be around 0.2. Also net entry has a positive contribution on the markup. This figure in itself suggests that the rise in the sales-weighted markup is mainly associated with a change in the allocation. But as shown in the previous section, relative to the counterfactual the difference between the reallocation and within effect is not that large.

An alternative to the sales-weighted markup is the cost-weighted markup, which is more appropriate when relating markups to welfare (Edmond et al., 2018). The cost-weighted markup has not increased as much as the sales-weighted markup (Edmond et al., 2018). The cost-weighted markup only increased by 0.1 between 1980 and 2015. As is well-known, the cost-weighted average markup is equal to the output elasticity times

\begin{itemize}
\item \textsuperscript{11}I drop firms without industry information, and I drop firms with negative sales, cost of goods sold or selling, general and administrative expenses. Furthermore, I trim the top and bottom percentile of the ratio of sales to cost of goods sold and of the ratio of cost of goods sold to all costs. Nonetheless, my results are slightly different from what De Loecker et al. (2020) report as I drop firms that report in Canadian dollars and only keep firms that report in US dollars. In addition, I do the decomposition at the industry level and then aggregate across industries.
\item \textsuperscript{12}The entry term refers to entry and exit into the dataset and not to entry and exit into the economy. For instance, a firm that goes from being public to being private is included in the exit term.
\end{itemize}
Figure 2: Haltiwanger decomposition of the markup in Compustat

total sales divided by total costs. Thus, in this sense the aggregate cost-weighted markup is insensitive to within-industry variation. However, this does of course not mean that changes in the cost-weighted aggregate markup cannot be due to reallocation of economic activity. It could in principal still be the case that the rise in the cost-weighted markup is due to high-markup firms becoming larger in terms of their total costs or wage bill. However, the results in Figure 2b clearly show that this is not the case. The cost-weighted decomposition is the opposite of what is found for the sales-weighted markup. The cumulative reallocation effect is negative while the within effect is equal to the aggregate rise in markups. What is especially remarkable about this result is that for the stationary counterfactual it does not matter much whether the sales-weighted or cost-weighted average markup is used, but for the empirical estimate it turns out to matter a lot. Also, relative to the counterfactual the within effect does more than fully explain the rise in cost-weighted markups. This is an indication that reallocation is not the major driver of the rise in markups.

III Mismeasured Markups

The above counterfactual and data analysis assume that markups are measured accurately. In this section, I show that mismeasurement of firm-level markups can largely account for the trends observed in the previous section. To study measurement error it is informative to first take a step back and disentangle the reallocation effect in the Haltiwanger decomposition in the between and covariance effect separately. Figure 3 shows that the covariance effect is zero in the sales-weighted decomposition and negative in the cost-weighted decomposition. This is a striking finding as in the model the covariance effect is positive and of about a similar size between the sales-weighted and cost-weighted decompositions. It could of course be the case that the model is misspecified, but here I will argue that model misspecification is unlikely to be driving the differential results between data and model. I show that these trends in the covariance term can instead be
explained by measurement error. The reason why model misspecification is unlikely to be driving the differential effect in the covariance terms between model and data is as follows. In general, when the shocks considered are demand shocks, the model will always generate a positive covariance effect for both the sales-weighted and cost-weighted markup. A positive demand shock increases the quantity sold and the markup. Due to there being decreasing returns to scale when holding the capital input constant, marginal costs are increasing and thus the price is increasing as well. Hence, sales and total costs are increasing as well leading—together with the increased markup—to positive covariance terms. To see what the effect is of productivity shocks on the covariance term it is insightful to consider for the moment Hicks-neutral technology changes and that all inputs are variable. In this case, when productivity increases by 1%, the marginal cost decreases by 1%. With imperfect pass-through this leads to a decline in the price of less than 1%, denote this by $\Delta p$ (in absolute value). Hence the markup increases by $(1 - \Delta p)\%$. With a price elasticity of demand $\tilde{\sigma}$ being larger than 1, the real quantity sold increases by $\tilde{\sigma}\Delta p\%$, such that nominal sales increase by $(\tilde{\sigma} - 1)\Delta p\%$. Thus, increases in the markup are associated with increases in nominal sales. Total costs, however, change by $(\tilde{\sigma}\Delta p - 1)\%$. Thus, if the price-elasticity of demand is low and/or the pass-through is low, changes in total costs can be negatively associated with changes in the markup. However, Amiti et al. (2019) find a cost pass-through of 0.6. Thus, for a negative covariance term to occur it must be the case that the price-elasticity of demand is less than 1/0.6 which implies a markup that is larger than 2.5 which seems implausible. Also in simulations in which productivity takes the form of factor-augmenting technology and capital is a fixed input I have not been able to generate a zero or negative covariance term for reasonable calibrations. The argument why model misspecification is unlikely to be driving the differential effect in the covariance terms between model and data depends, however, on there being a positive relationship between markups and firm size. Obviously, when the super-elasticity is equal to zero all firms have the same markup and hence the covariance terms would be equal to zero. Such

Figure 3: Haltiwanger decomposition of the markup in Compustat - separating out the covariance and between effect
a model could thus explain the zero sales-weighted covariance term but still would not be able to explain the negative cost-weighted covariance term. Moreover, a zero or negative relationship between sales and markups is inconsistent with the data as I show later in this section.

In the model I have assumed there are no adjustment frictions, expect for capital. Could it be the case that adjustment frictions to other inputs or prices explain the zero sales-weighted and negative cost-weighted covariance terms? When adjustment frictions are driving the decomposition results obtained so far, the decomposition will be very different when longer time lags are considered. Intuitively, adjustment frictions matter less when a longer horizon is considered. Figure 10 in Appendix B shows the decomposition in Compustat when three-year lags are taken instead of one-year lags. The motivation for doing this exercise is that it is likely that three years after a shock a firm has probably been able to re-adjust its levels of inputs and prices. It turns out that the resulting decomposition with three-year lags is similar to the decomposition with one-year lags, although the reallocation effect is somewhat lower for the sales-weighted markup. This strongly suggests that adjustment frictions are not driving the results I find. Therefore, in what follows I only consider models with no adjustment frictions except for capital.

Another option is that firm churning has other causes than changes in productivity or demand, which might potentially be able to explain the observed covariance terms. It could, for instance, be that the level of competitiveness within a market is changing over time. This could be modeled in two ways. The first is to change the elasticity of substitution between varieties. Alternatively, the number of firms active in a market can change. In markets in which the elasticity is increasing the largest firms experience an increase in market shares based on both sales and costs (as consumers become more price sensitive and the largest firms are large because they charge a low price), while the smaller firms become smaller. For small firms the markup goes down, while for larger firms the effect on the markup is ambiguous. The increase in the elasticity of substitution has a direct negative impact on markups for all firms, while the increase in firm size for the largest firms has a positive effect on markups. Thus, for small firms the covariance term is positive, while for large firms the sign of the covariance term is ambiguous. Therefore, it is theoretically possible that the overall covariance effect is negative. However, the above suggests that the covariance term is similar in magnitude for the sales-weighted and cost-weighted markup in this version of the model as well. Therefore, this seems to be an unlikely explanation for the zero sales-weighted covariance term and negative cost-weighted covariance term. Moreover, in simulations it turns out to be the case that the covariance term is positive when for some markets the elasticity of substitution is increasing while for other markets the elasticity is declining.\textsuperscript{13} Furthermore, changes in

\textsuperscript{13}Autor et al. (2020) show that the skewness of the productivity distribution affects whether the reallocation effect dominates the within effect when markets become more ‘tough’. Although the model studied here is different from the model studied in Autor et al. (2020), something similar is going on here as well. When in the oligopolistic Cournot model the elasticity of substitution between varieties increases (i.e., markets become more ‘tough’), the aggregate markup decreases when productivity is log-normally distributed and increases
the number of firms also lead to a positive covariance effect as an increase in the number of firms leads to lower market shares and lower markups, and vice versa when the number of firms falls.

The above discussion makes me confident to say that the differential results regarding the covariance term between the model and data are not due to model misspecification. Other possible explanations that turn out to be more potent are measurement error of markups and measurement error of market shares. Following De Loecker and Warzynski (2012), the production approach to measuring markups rewrites the first-order condition with respect to an input to obtain that the markup equals the product of the output elasticity, $\theta^X_i$, and the inverse sales share of an input $X_i$

$$\mu_i = \theta^X_i \frac{P_i Y_i}{P^X_i X_i}.$$  

(8)

It could be that the variables in this formula are mismeasured. For instance, there is an $i$ subscript on the output elasticity while, typically, when estimating a production function at best limited variation in the output elasticity is allowed for. Furthermore, sales or costs might be mismeasured. Especially the fact that the covariance term is negative for the cost-weighted markup suggests that mismeasurement might be important. The negative covariance term means that firms that increase their markup also experience a decline in their costs. By the above formula, this pattern could be due to mechanical reasons. In years in which costs are measured to be too high, markups are measured too low. Then, changes in measurement error over time yield a downward biased estimate of the covariance effect for the cost-weighted markup. On the other hand, measurement error in sales is unlikely to explain the zero sales-weighted covariance effect as this would lead to an upward biased sales-weighted covariance effect as for firms for which sales is measured too high also the markup is overestimated.

**Measurement Error Output Elasticities**

Before discussing measurement error in costs and sales in more detail, I first discuss measurement error in output elasticities as this turns out to provide an explanation for the zero sales-weighted covariance effect. The output elasticity, $\theta^X$, is typically estimated by estimating a production function that allows only for hicks-neutral differences in technology. If instead there are also differences in factor-augmenting technology across firms there will be unmeasured heterogeneity in the output elasticity. In the presence of heterogeneity in the output elasticity, the following relationship emerges between the with a Pareto distribution. However, for both the log-normal and Pareto distributions it is the case that the covariance term is positive when for some markets the elasticity of substitution is increasing while for other markets the elasticity is declining.
measured firm-level markup, $\hat{\mu}_i$, and the true markup, $\mu_i$:

$$\hat{\mu}_i = \hat{\theta}^X \frac{P_i Y_i}{P_i^X X_i} = \hat{\theta}^X \mu_i,$$

(9)

where $\hat{\theta}^X$ is the estimated output elasticity which is assumed to be common across firms. Thus, a firm for which a high markup is estimated might in fact be a firm with a low output elasticity. This means that the distribution of the estimated markup is not necessarily informative of the true distribution when there is heterogeneity in the output elasticity. And when firm-level output elasticities/factor-augmenting technologies are changing over time this is also problematic for interpreting the decomposition results as this might lead to a spurious relationship between changes in measured markups and market shares. Note that for simplicity I have assumed here that the estimated output elasticity is the same across firms. This is the case for a Cobb-Douglas production function, which is frequently used in empirical applications. Another frequently estimated production function is the translog production function which does lead to heterogeneous output elasticities. However, when estimating a translog production function full heterogeneity of output elasticities is typically not allowed for as it is usually assumed that factor-augmenting technologies are identical across firms. Thus, also with a translog production function there is unmeasured heterogeneity of output elasticities and thus mismeasured markups.

There is recent evidence that there is variation in factor-augmenting technologies across firms (David and Venkateswaran, 2019; Demirer, 2019; Doraszelski and Jaumandreu, 2018; Raval, 2019, 2020). The above Leontief production function (6) in the model does feature differences in factor-augmenting technology and thus differences in output elasticities. This production function yields that the larger productivity $A_i$ the fewer capital and labor is needed to produce one unit of real output, while the amount of materials that is needed to produce one unit of real output is not affected by productivity. To estimate the markup, De Loecker et al. (2020) take the variable input as cost of goods sold which corresponds to the combined cost of direct labor used in production and the use of raw materials. Thus, this is a bundle of inputs and in the above production function cost of goods sold of a firm $i$ correspond to $wL_i + P^M M_i$. When estimating a Cobb-Douglas production function in capital and this bundle of inputs, the output elasticity of this input bundle is identical across firms, but with the above production function they are clearly not when there is heterogeneity in productivity $A_i$ across firms. This leads to a mismeasured markup.

To see how mismeasured output elasticities affect the decomposition, consider a firm that experiences an increase in factor-augmenting productivity. This implies a decrease in the marginal cost, and therefore a decrease in the price level, and an increase in the

\[ \hat{\theta}^X \] is aware that the Leontief production function is not consistent with the necessary assumptions for the production approach to measuring markups to hold; namely that the production function is twice differentiable. However, the Leontief can be viewed as the limiting case of a CES production function for which the elasticity of substitution between materials on the one hand and capital and labor on the other hand approaches zero. For this CES production function the assumptions of the production function approach to measuring markups hold.
market share (as long as the elasticity of demand exceeds 1) and true markup, yielding a positive covariance effect as documented in the model above. What happens to the measured markup? Does it necessarily increase as well or could it also decline? Given that the estimated output elasticity does not change (as it is identical across all firms), we have to consider what happens to the inverse revenue share of costs of goods sold. Suppose the increase in productivity is such that the marginal cost decreases by 1% and the price falls by $\Delta p$ in absolute value. Hence, as before, the true markup increases by $(1 - \Delta p)\%$. The real quantity sold increases by $\tilde{\sigma} \Delta p\%$ and nominal sales increases by $(\tilde{\sigma} - 1)\Delta p\%$. Hence, expenditure on materials, $P^M$, increases by $\tilde{\sigma}\Delta p\%$, while expenditure on labor changes by $\frac{1}{1 - \alpha} (\tilde{\sigma} \Delta p - \Delta A)\%$. Thus, the numerator in the markup estimator (sales) increases by $(\tilde{\sigma} - 1)\Delta p\%$ while the denominator (cost of goods sold) increases by a percentage between $\frac{1}{1 - \alpha} (\tilde{\sigma} \Delta p - \Delta A)\%$ and $\tilde{\sigma}\Delta p\%$. When a substantial share of costs are materials costs, the increase in the denominator is closer to $\tilde{\sigma}\Delta p\%$ (this is the limiting case in which labor costs are zero). Thus, in this case, when productivity increases, costs increase faster than sales making that the measured markup declines. Hence, mismeasuring firm-level output elasticities biases the sales- and cost-weighted covariance terms negatively and can thus potentially explain why a zero sales-weighted covariance effect is observed.

In order to understand whether mismeasuring the output elasticity can quantitatively explain the zero sales-weighted covariance term I am going to do the decomposition on the measured markup in simulated data from the above model. In the above calibration, I exogenously set the capital intensity $\alpha$ equal to 0.31. This was done in anticipation of the current exercise in which I calibrate $\alpha$ to match the zero covariance term in the decomposition of the sales-weighted measured markup. When $\alpha$ is large, labor costs are low such that cost of goods sold predominantly consists of materials costs making that the covariance term becomes negative. On the other hand, when $\alpha$ is small, the covariance term is positive. It turns out that when $\alpha$ is equal to 0.31 the covariance term is zero. Is an $\alpha$ of 0.31 reasonable? To answer this question, we should have a look at the labor share of value added. The labor share in this model is only 31% which seems small at first sight compared to a labor share of around 60% as it is in the data. However, it should be noted that in this model labor only refers to labor used directly in production (i.e., the equivalent of the labor part of cost of goods sold in data). With a markup of 1.38 and constant returns to scale, profits gross of overhead costs are equal to 56% of value added.\(^{15}\) This is obviously unreasonably large and suggests that there are overhead costs. In Van Vlokhoven (2019) I find that profits net of all costs are about 15% of value added in the same data set. This suggests that in the current model around 40% of value added corresponds to overhead costs. If we would think that overhead is more labor intensive than production, such that—let’s say—three quarters of overhead costs refer to labor, then the labor share of value added according to the model is around 61% which is close to

---

\(^{15}\)Profits as a share of value added are equal to $\left(1 - \frac{1}{\phi}\right) \frac{1 - \text{material share of sales}}{1 - \text{material share of sales}}$ where $\phi$ are the returns to scale. In this formula, the markup refers to the cost-weighted markup, which in this model is similar to the sales-weighted markup.
empirical estimates. Thus, an $\alpha$ of 0.31 that is needed to generate a zero covariance effect seems plausible. This is by any means no proof that the zero covariance effect is due to a misspecified production function and therefore mismeasured micro-markups, but it is evidence that this is a plausible explanation. I provide more evidence that firm-level output elasticities are misspecified towards the end of this section.

Moreover, for completeness I have estimated a translog production function which allows for limited heterogeneity in output elasticities. Figure 11 in Appendix B shows that doing so gives a smaller between effect than in the baseline though it is still positive. The cumulative sales-weighted between effect over a 35-year horizon is about 0.15 compared to 0.25 in the baseline. And the cost-weighted cumulative between effect is about 0.15 compared to 0.2 in the baseline. The within effect according to the translog is quite erratic but about equally large as the between effect. That the between effect becomes smaller when addressing heterogeneity in the output elasticity is consistent with the model as the between effect according to the true markup is lower than the between effect of the measured markup. Estimating a translog production function, however, only partly addresses the bias due to mismeasured output elasticities. This is because the translog still assumes that firms do not differ in factor-augmenting technologies and therefore only captures part of the variation in output elasticities. Therefore, capturing full heterogeneity in output elasticities would probably lead to an even lower between effect.

**Measurement Error Costs**

The model does not only yield a zero covariance term for the sales-weighted measured markup but also a zero covariance term for the cost-weighted measured markup whereas the latter is large and negative in the data. Moreover, the slope of the covariance term is becoming larger over time. Between 1980 and 1990 the change in the covariance term is -0.03, during the 1990s the change is -0.06, during the 2000s it is -0.1 and between 2010 and 2015 it is -0.06. I am going to match this negative covariance term by introducing measurement error in costs. In years in which cost are measured too high, the markup is measured too low and vice versa in years in which costs are measured too low. This gives a negative bias for the cost-weighted covariance term. I model transitory measurement error from classifying some inputs in an adjacent year to when it actually occurred as follows. Each year a fraction of firms chooses to shift a fraction of their costs of that year to the next year. Thus, over a longer horizon firms are not able to underreport their costs but they can do so for an individual year.\textsuperscript{16} The firms that choose to do so are randomly chosen and thus changes from year to year. The fraction of costs that a firm can shift to the next year is randomly distributed and follows a censored lognormal distribution. To match the increasing slope of the covariance term I let measurement error increase over

\textsuperscript{16}Using the more standard assumption that measurement error is iid across years yields similar results but has as somewhat unsatisfying property that some firms might underreport their costs for multiple years in a row.
Figure 4: Haltiwanger decomposition of the measured markup in the model when there is measurement error of the output elasticity and costs.

Figure 4 shows the resulting decomposition of the measured markup when both the firm-specific output elasticity and costs are mismeasured. Recall that the covariance term for both the sales-weighted and cost-weighted markup are targeted by choosing $\alpha$ and the measurement error of costs, while the between and within effect are not targeted. As the cost-weighted covariance term is negative, the between and/or the within effect has to be positive since the aggregate markup is constant over time in these simulations. The right panel shows that both turn out to be positive. Thus, although the joint distribution is not changing over time, the decomposition of the measured markup gives both a positive between and within effect. This is opposite from the decomposition of the true markup which yields a negative between and within effect. The reason why the between effect is positive is that firms which in year $t$ underreport their costs have a high measured markup and a low market share. Thus, when they do not underreport their costs in the next year they will experience an increase in market shares and thus it appears as of high-markup firms are gaining market shares. The within effect is positive because firms with low costs tend to experience a decrease in their measured markup as one reason why they have a low cost this year is that they have underreported their cost this year. Comparing these results with the results obtained using Compustat it is apparent that these simulated results go in the same direction as in the data, but the gap between the between and within effect is larger in the data. In Compustat, over a 35 year period the between effect is 0.1 larger than the within effect while in the simulated data the between effect is 0.03 larger than the within effect.

The left panel in Figure 4 shows the sales-weighted decomposition. The between effect...
is positive and the within effect is negative. This is consistent with the empirically found between effect dominating the within effect. However, the difference is quantitatively limited as is the case for the cost-weighted markup. In the simulations the between effect dominates the within effect by 0.03 while in the data the between effect dominates the within effect by around 0.22. In this sense it might be possible that there empirically is a positive between effect compared to a counterfactual, but the point is to show that these decomposition results are sensitive to model misspecification. In the case of the cost-weighted markup the between and within effects for the measured markup are of the opposite sign of the results for the true markup. Likewise, for the sales-weighted measured markup the between and within effect are close to zero while for the measured markup they are negative. When $\alpha$ would be larger than 0.31 also for the sales-weighted markup the between and within effect would be of opposite sign depending on whether the true or measured markup is used.

**Measurement Error Sales**

In the above I have mostly focused on the decomposition, but my results also have implications for the estimate of the aggregate markup. In addition, I have assumed that sales are correctly measured, but since the above suggests that measurement error in costs has increased over time it might very well be the case that measurement error of sales is present as well and has also been increasing over time. This might be relevant for understanding the different evolution of the sales-weighted and cost-weighted average markup. The cost-weighted average markup is equal to total costs divided by total sales times the output elasticity

$$\sum \frac{P^X_i X_i}{P^X X} \hat{\mu}_i = \sum \frac{P^X_i X_i}{P^X X} \hat{\theta}^X \frac{P_i Y_i}{P^X X_i} = \hat{\theta}^X \frac{PY}{P^X X},$$

where totals are denoted when the $i$ subscript is omitted. Thus, as long as the output elasticity is consistently estimated, idiosyncratic measurement error of costs or sales does not affect the cost-weighted average markup. This is very different for the sales-weighted average markup

$$\sum \frac{PY_i}{PY} \hat{\mu}_i = \sum \frac{PY_i}{PY} \hat{\theta}^X \frac{P_i Y_i}{P^X X_i} = \sum \hat{\theta}^X \frac{(P_i Y_i)^2}{PY P^X_i X_i}.$$

By Jensen’s inequality, the larger measurement error the larger the estimated sales-weighted average markup. This is intuitive. Firms for which observed sales are too high are over-weighted while also having a higher estimated markup.

A way to correct for measurement error of sales is to use the first stage of the Ackerberg et al. (2015) procedure of estimating a production function. In the first stage of this procedure, measurement error of sales is purged by regressing sales on a flexible functional
Figure 5: The sales-weighted markup in Compustat both when corrected for measurement error sales and when not.

form of inputs

\[ y_i = f(x_i, k_i) + \xi_i \]  

(10)

where lower case letters denote logs and are in nominal terms, and \( \xi \) denotes measurement error in sales. Estimating the above equation gives predicted sales \( \hat{y}_i \). To correct for measurement error, we can simply use predicted sales instead of actual sales to estimate the markup and to obtain the market share. Using a third-degree polynomial in capital and cost of goods sold to purge measurement error gives the results as displayed in Figure 5. Correcting for measurement error changes the estimate of the sales-weighted average markup substantially. Instead of increasing by 0.3 between 1980 and 2015, the aggregate markup is only increasing by around 0.15. Moreover, zooming out and considering the period from 1960 onward the sales-weighted markup has not increased when correcting for measurement error while the markup increased by 0.2 in the baseline. This is not to say that markups have not increased during this time period. De Loecker et al. (2020) find that the cost-weighted average has increased by around 0.1 after 1960 and, as mentioned above, measurement error does not affect the cost-weighted average (as long as output elasticities are consistently estimated). The large sensitivity of the sales-weighted markup is an argument in favor of using the cost-weighted markup. Finally, correcting for measurement error makes that the cost-weighted average markup is about equally large as the sales-weighted average markup while in the baseline the sales-weighted average markup is substantially larger.

In addition to changing the sales-weighted average markup, Figure 6 shows that correcting for measurement error also alters the decomposition results dramatically. When correcting for measurement error the reallocation effect is close to zero and the entire rise in markups after 1980 is driven by a within effect. These patterns hold for both the sales-weighted and cost-weighted markup.\textsuperscript{19} The reallocation effect consists of a

\textsuperscript{19} Also, the cost-weighted average markup turns out to be somewhat lower when correcting for measurement error. This is because equation (10) is in logs and thus average predicted sales is not equal to average

23
positive between effect and a negative covariance effect which are about equally large in magnitude. This is different from the baseline sales-weighted between effect which is 0.2 over this 35-year horizon while after correcting for measurement error the between effect becomes around 0.1. Thus, also correcting for measurement error in sales makes that the reallocation effect is smaller than previously thought.

Correcting for measurement error of sales also affects the Baqee and Farhi (2020) estimate of the change in allocative efficiency. Instead of an increased allocative efficiency of 7 percentage points I find that allocative efficiency has increased by only 3 percentage points between 1997 and 2015 after correcting for measurement error of sales.\footnote{I calculate this quantity as follows. I use for the Domar weights the sales share. In the baseline, the first term of equation (2) equals 0.04. Using predicted sales this term falls to 0. Thus, using predicted sales lowers the estimate of the change in allocative efficiency by four percentage points since the factor income shares are not affected.}

Finally, measurement error might also affect the estimate of the markup by affecting the output elasticity. Standard methods to estimate production functions such as Ackerberg et al. (2015) do allow for measurement error in output, but not for measurement error in inputs.\footnote{Collard-Wexler and De Loecker (2021) is a notable exception that allows for measurement error in capital.} The recent literature on estimating markups has criticized the production approach as it does lead to biased estimates when firm-specific prices are not observed (Bond et al., 2021). I find that even when real variables are observed the bias can be large due to measurement error in costs and production function misspecification. To get the output elasticity I follow De Loecker et al. (2020), and using Ackerberg et al. (2015) I estimate a Cobb-Douglas production function with as inputs capital and costs of goods sold, allowing for Hicks-neutral technology differences. To see how production function misspecification and measurement error of costs affect the estimate of the markup individually I first estimate a production function when costs are correctly measured. In my simulations, the estimated markup is 20\% larger than the true markup. Thus, there is in this case a substantial positive bias when the production function is misspecified. On the other hand, adding measurement error in inputs does not alter this bias substantially.

---

Figure 6: Haltiwanger decomposition when corrected for measurement error sales in Compustat
Obviously, that there is a positive bias due to production function mismeasurement and only a small bias due to measurement error of costs is not a general result. However, it highlights that markups obtained from estimating a production function should be interpreted with care.

**Relationship Markups and Firm Size**

In the literature there have been some contradicting findings regarding the relationship between firm size and measured markups. I show in this section that these findings can be explained by mismeasured markups due to mismeasured output elasticities. Using the De Loecker et al. (2020) markup estimates within Compustat yields a negative super-elasticity of \(-0.05\), which is inconsistent with standard models of firm conduct. This negative relationship between markups and firm size is noisy as regressing log markups on log sales yields a positive, though small, coefficient. Thus, at best there is a weak positive relationship between measured markups and sales in Compustat. On the other hand, Edmond et al. (2018) find using the US Census of Manufactures a robust positive and large relationship between markups and market shares; namely a super-elasticity of 0.16. Although these results seem contradictory I argue here that they are consistent with each other. One difference between De Loecker et al. (2020) and Edmond et al. (2018) is that Edmond et al. (2018) construct the markup using labor revenue shares and thus the output elasticity of labor while De Loecker et al. (2020) use a bundle of materials and labor. As there is only limited labor data in Compustat it is not possible to test directly whether this difference in results is due to using different variables when measuring markups or due to using different data sets. However, in a recent paper Xhani (2021) finds within the same data set for the UK a negative relationship between markups and sales when markups are measured using an input bundle and a positive relationship when markups are measured using labor only.

One reason why these differential results appear depending on which input is used is that these papers assume that output elasticities are constant across firms. Instead if there are differences in factor-augmenting technologies the output elasticities will differ by firm size. To test this hypothesis I estimate the relationship between size and measured markups within my simulated data both when the markup is measured using labor only and when an input bundle of materials and labor is used. The elasticity between sales and markups is 0.04 for the true markup. When using the markup estimated using the input bundle the estimated elasticity becomes 0.004, whereas when using the markup estimated using labor the elasticity becomes 0.27. The reason for this result is that with this production function the output elasticity of the input bundle is increasing in firm size.\(^22\) For the smallest firms this output elasticity is 0.75 whereas for the largest firms this is 0.95. When this is not taken into account by assuming a constant output elasticity the

\(^{22}\)To calculate the output elasticity I invert equation (8) to get the output elasticity as the product of the true markup and the revenue share of the respective input.
Table 2: Relationship between firm size and markups in Compustat

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>log markup</td>
<td>Const $\theta$</td>
<td>Varying $\theta$</td>
<td>Const $\theta$</td>
<td>Varying $\theta$</td>
</tr>
<tr>
<td>log sales</td>
<td>$0.002^{***}$</td>
<td>$0.055^{***}$</td>
<td>$-0.046^{***}$</td>
<td>$0.126^{***}$</td>
</tr>
<tr>
<td></td>
<td>$(0.001)$</td>
<td>$(0.001)$</td>
<td>$(0.001)$</td>
<td>$(0.014)$</td>
</tr>
<tr>
<td>Year-Industry FE</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Observations</td>
<td>186713</td>
<td>185606</td>
<td>162487</td>
<td>101057</td>
</tr>
<tr>
<td>$R^2$</td>
<td>0.09</td>
<td>0.20</td>
<td>0.08</td>
<td>0.03</td>
</tr>
</tbody>
</table>

Robust standard errors in parentheses. The columns with a varying output elasticity $\theta$ allow $\theta$ to vary by decile of the sales distribution within an industry. The columns with a constant output elasticity take $\theta$ to be the same across firms within an industry.

$^*$ $p < 0.05$, $^{**} p < 0.01$, $^{***} p < 0.001$.

production approach underestimates the markup of the large firms compared to the small firms when the input bundle is used. For the output elasticity with respect to labor the opposite occurs. In this model, larger firms have a lower output elasticity with respect to labor. Thus, when using only labor the markups of large firms are overestimated compared to small firms. To summarize, a larger elasticity between sales and measured markups is found when only labor is used to estimate markups compared to when labor plus materials are used.

To further investigate how mismeasurement of the output elasticity affects the relationship between firm size and markups I estimate the output elasticity for each decile of the firm size distribution separately. As there are too few observations to estimate the output elasticity for each industry-year-decile I combine data across years. That is, for each industry-year I determine to which decile a firm belongs and then estimate the output elasticity for all firms of the same decile within an industry. The first two columns of Table 2 show that doing so gives an elasticity of markups with respect to sales of 0.06 whereas when the output elasticity is taken to be common across all firms the elasticity is 0.002. Figure 12 in Appendix B shows graphically the relationship between the median markup and firm size by decile. Except for the first decile the relationship between markups and firm size is found to be monotonically increasing when the output elasticity varies by size while it is found to be monotonically decreasing when the output elasticity is independent of firm size. From this figure it is also apparent that when allowing for a size-dependent output elasticity the estimated markup is lower. Moreover, the last two columns of Table 2 estimate the super-elasticity by regressing $\frac{1}{\mu} + \log(1 - \frac{1}{\mu})$ on the log of sales. The super-elasticity equals 0.13 when using a size-dependent output elasticity which is statistically indistinguishable from my calibration of 0.16. Instead, with a common output elasticity, the super-elasticity is estimated to be -0.05. The reason why the relationship between sales and markups becomes larger when allowing for varying output elasticities is that I estimate the output elasticity to be increasing in firm size. Figure 7
Figure 7: The output elasticity as a function of firm size in Compustat

shows that the output elasticity is below 0.6 for firms with a below-median size and above 0.6 for the upper half of firms.

These results are consistent with what is found in a recent paper by Raval (2020). Using data from several countries he finds that the relationship between firm size and markups is often negative or zero when using a translog production function to estimate the markup. However, similar to the results I find for Compustat, when he estimates the markup using a more flexible procedure that allows for differences in factor-augmenting technology he tends to find a positive relationship between firm size and markups. The way in which Raval (2020) deals with heterogeneity in factor-augmenting technologies is different from how I do this. Raval (2020) uses a cost share estimator and groups plants based on their labor cost to material cost ratio while I estimate a separate production function for each decile of the firm size distribution.

IV Olley-Pakes Decomposition

The above Haltiwanger decomposition looks at changes over time within firms. An alternative is to consider how the moments of the cross-sectional distribution change over time. This is done by the Olley and Pakes (1996) decomposition extended by Melitz and Polanec (2015) to take entry and exit into account. Elementary algebra shows that the weighted-average markup is equal to the unweighted average plus the sum of the product of the demeaned market share and demeaned markup (i.e., the covariance times the number of firms)

$$\mu_t = \bar{\mu}_t + \sum_i (m_{it} - \bar{m}_t) (\mu_{it} - \bar{\mu}_t) = \bar{\mu}_t + n_t \text{cov}(m_{it}, \mu_{it}),$$

where $n_t$ denotes the number of firms and $\bar{\cdot}$ denotes unweighted averages. Taking the difference between two periods, and taking entry and exit into account following Melitz
and Polanec (2015) yields

$$
\Delta \mu_t = \underbrace{\Delta \mu_t^C}_{\text{Within}} + \underbrace{n_t^C \Delta \text{cov}_{t}^C}_{\text{Reallocation}} + \underbrace{M_t^E (\mu_t^E - \mu_t^C)}_{\text{Net entry}} - \underbrace{M_{t-1}^X (\mu_{t-1}^X - \mu_{t-1}^C)}.
$$

(11)

The superscript $C$ denotes firms that continue between $t - 1$ and $t$. For example, $\Delta \mu_t^C$ is the change in the unweighted average markup between $t - 1$ and $t$ among continuing firms. The superscript $E$ denotes firms that enter and $X$ denotes firms that exit. Let $M_t^G = \sum_{i \in G} m_{it}$ represent the market share of a group $G$ of firms and define $\mu_t^G = \sum_{i \in G} \frac{m_{it}}{M_t^G} \mu_{it}$ as that group’s markup.

Thus, the within effect equals the change in the unweighted-average markup among continuing firms. The reallocation effect is the change in the covariance between market shares and markups among continuing firms. And the net entry effect compares the markup among entering and exiting firms with the markup of continuing firms. Figure 8 shows the resulting OPMP decomposition of the markup among Compustat firms between 1980 and 2015.23 The first panel shows that for the sales-weighted markup the firm-reallocation term is again positive and about as large as it is for the Haltiwanger decomposition. The OPMP decomposition gives that about 0.2 of the rise in markups is due to reallocation. And the overall within effect is close to zero, but now it is negative during the early years of the sample. Different from the baseline Haltiwanger decomposition, for the OPMP decomposition it does not matter much whether we consider the sales-weighted or cost-weighted markup. For the baseline Haltiwanger decomposition the reallocation effect is initially zero and becomes negative after 2000 when considering the cost-weighted markup, while the OPMP decomposition gives a positive reallocation effect for the cost-weighted markup initially (second panel). However, for the latter half of the sample the reallocation effect becomes negative making that the overall reallocation effect is close to zero. By construction, the within effects are identical across the sales-weighted and cost-weighted decomposition.

In the absence of firm entry and exit, the above criticism that applies to the Haltiwanger decomposition does not apply to the Olley-Pakes decomposition. When there are changes to the market share and markup at the firm level, but such that the joint distribution does not change over time, the Olley-Pakes decomposition identifies a zero within effect and a zero reallocation effect. However, this is not necessarily true anymore when the decomposition also takes entry and exit into account. To see this suppose again that among firms present in the market the joint distribution of markups and market shares is constant over time but suppose now that there is also entry and exit. The within effect will be negative when firms that enter have a larger markup than firms that exit. The reason is that, for the distribution to be stationary, it must be the case that continuing firms experience a decline in their average markup. Moreover, depending on how market shares

---

23Reallocation across industries is still based on a Haltiwanger decomposition at the industry level as there is no reason to believe that growing industries also experience a change in their markup in a systematic way.
are distributed among firms, also the covariance among continuing firms can change. For instance, if entering and exiting firms have a negligible market share (i.e., $M_E \approx 0$ and $M_X t - 1 \approx 0$) then the negative within effect implies a positive reallocation effect by equation (11). Thus, although the joint distribution of markups and market shares does not change, also this decomposition can yield a positive reallocation effect and a negative within effect. The reallocation effect can also be non-zero in a stationary economy while the within effect is zero. Suppose that the correlation between markups and market shares is positive and that the market share of entrants is equal to the market share of firms that exit. Then, when entering firms have less dispersed market shares and/or markups than exiting firms the reallocation effect is measured to be positive.

**Proposition 2 (Olley-Pakes-Melitz-Polanec Decomposition).** Suppose that the joint distribution of markups and market shares is constant, and that entering firms have on average a larger markup than exiting firms. Then, as firms enter and exit, the within term in the OPMP decomposition is negative. In addition, when the market shares of exiting and entering firms are sufficiently small, the reallocation term is positive. Furthermore, when the market share of entrants is equal to the market share of firms that exit, when there is a positive correlation between markups and market shares, and when entering firms have less dispersed market shares and/or markups than exiting firms, the reallocation effect is measured to be positive also when the within effect is zero.

Likewise to the Haltiwanger decomposition, this also means for the OPMP decomposition that finding a positive reallocation effect and a negative within effect does not necessarily imply that the allocation has changed. The reason is that in equation (11) the within and reallocation effects are calculated among continuing firms only. It could be the case that the unweighted average markup among continuing firms is declining between two periods while the unweighted average markup among all firms stays constant or even increases. Whether this is the case can easily be tested directly in the data. Despite that the OPMP decomposition suggests otherwise, Figure 9 shows that it is indeed the case that the unweighted average markup has been increasing over time while the covariance between
the markup and market share has been declining.\textsuperscript{24} This is true both when considering all firms present in a given year or only considering the firms that continue between two consecutive years. To see how these results are consistent with the above OPMP decomposition consider the decomposition between the years \(t - 1\) and \(t + 1\) during the early part of the sample period when the OPMP decomposition gives a negative within effect while the unweighted average markup is increasing. Firms that operate at both time \(t - 1\) and \(t\) have a lower markup on average at time \(t\), and also firms that operate at both time \(t\) and \(t + 1\) have a lower average markup at time \(t + 1\) (by the negative within effect in the OPMP decomposition). But firms that operate in both \(t\) and \(t + 1\) have a larger markup at time \(t\) than firms that operate in both \(t - 1\) and \(t\) have at time \(t - 1\). That in Figure 9a the average markup among all firms is larger than the average markup among continuing firms (between \(t - 1\) and \(t\)) means that firms that enter the data set have on average a larger markup than exiting firms.

The unweighted average markup has increased by around 0.4 which is larger than the rise in the aggregate markup. These results suggest that the rise in markups among Compustat firms is due to an increase in firm-level markups and not due to reallocation of economic activity toward firms with a high markup. The results here depend to some extent on how narrowly defined industries are. The average markup and covariance in Figure 9 are calculated within a 2-digit industry and aggregated using industry sales shares (as the other decompositions are). When considering 3-digit or 4-digit industries, the unweighted average markup increases by about 0.2 and the covariance is close to zero and displays an increasing trend over time where the total increase is about 0.05.\textsuperscript{25} Furthermore, that the covariance is negative does not necessarily mean that there is a negative relationship between the firm size and markup. Regressing the log markup on

\textsuperscript{24}The covariance here refers to \(\sum_i (m_{it} - \bar{m}_t) (\mu_{it} - \bar{\mu}_t)\). Only the covariance between markup and sales is plotted. The covariance between markups and costs has also been declining.

\textsuperscript{25}The Haltiwanger decomposition depends less on how narrowly defined industries are in this data set (except that the reallocation across industry term becomes more important the more narrowly defined industries are).
the log of sales gives a positive coefficient (see also the discussion in the previous section).

In Section III I have shown that correcting for measurement error of sales affects the Haltiwanger decomposition substantially. Likewise, Figure 13 in Appendix B shows that correcting for measurement error also affects the Olley-Pakes-Melitz-Polanec decomposition. Now for both the sales-weighted and cost-weighted average markup the within effect dominates while the reallocation effect is negative. This is another piece of evidence that the rise in markups among Compustat firms is not due to reallocation.

V Conclusions

This paper shows that standard decomposition methods yield biased results regarding whether changes in aggregate markups are due to reallocation or due to firm-level markups changing. This is in part due to the fact that decomposition methods can yield a positive reallocation effect even when the joint distribution of markups and market shares is not changing. I show this is the case for a Haltiwanger decomposition when there is a positive relationship between markups and market shares, and I show that an Olley-Pakes-Melitz-Polanec decomposition yields a negative within effect in a stationary environment when the markup of entering firms exceeds the markup of exiting firms. In addition, when firm-level markups are imprecisely estimated, markup decompositions give biased results. Firm-level markups can be mismeasured both due to production function misspecification and measurement error. Even if the aggregate markup would be consistently estimated it is unlikely that the decomposition—or for that matter other higher order moments of the firm-level distribution—is consistently estimated.

Although it is not the intention of this paper to make any final conclusions regarding whether reallocation is an important driver of changing markups, it seems to be the case that reallocation is at best an equally important driver of the change in markups among Compustat firms as changes in firm-level markups are. More likely is that the reallocation effect is close to zero. This conclusion is, amongst others, based on the following results. The baseline cost-weighted Haltiwanger decomposition yields a negative reallocation effect. Partially correcting for heterogeneity in output elasticities lowers the estimated reallocation effect. After correcting for measurement error of sales, the Haltiwanger reallocation effect is zero or negative for both the cost-weighted and sales-weighted markup. And finally, following the Olley-Pakes decomposition the rise in aggregate markups is associated with a rise in the unweighted markup.

There are several explanations possible for why I find a limited reallocation effect. In this paper I use Compustat data which does not cover many small firms. Thus, that I do not find much evidence of reallocation among Compustat firms does not mean there has not been any reallocation of economic activity between companies that are included in the Compustat database and firms that are not. Also, an alternative to studying markups directly is to consider the fall in the labor share for which, for instance, Kehrig and Vincent (2021) and Autor et al. (2020) find that reallocation is important. An advantage of studying
the labor share is that it is directly observed. However, a decomposition of the labor share is not necessarily informative about market power as firms might have a low labor share because they have automated more tasks and not because they have more market power.
References


Appendix A  Oligopolistic Competition

The model in the main text features monopolistic competition with varying markups. In this appendix I show that the same conclusions hold when considering oligopolistic competition à la Atkeson and Burstein (2008).

Suppose there is a continuum of sectors, indexed by $j \in [0, 1]$, that produce a good $y_j$. These sectoral outputs are used by a competitive firm to produce final output, $Y$, according to a CES production function with elasticity of substitution $\eta > 1$:

$$ C = \left( \int_0^1 y_j^{-\eta} \, dj \right)^{\frac{\eta}{\eta - 1}}. \quad (12) $$

The first-order conditions of the competitive firm give the inverse demand functions,

$$ \frac{P_j}{P} = \left( \frac{y_j}{Y} \right)^{-1/\eta}. \quad (13) $$

Each sector consists of $I$ firms, each producing a distinct variety in quantity $q_{ij}$. The output of a sector is given by the CES aggregate of these $I$ varieties with elasticity of substitution $\rho > \eta$:

$$ y_j = \left( \sum_{i=1}^I \chi_{ij} q_{ij}^{-\rho} \right)^{-\frac{1}{\rho-\eta}}, \quad (14) $$

where $\chi_{ij}$ is a preference shifter. As before, this gives as inverse demand function

$$ \frac{P_{ij}}{P_j} = \chi_{ij} \left( \frac{q_{ij}}{y_j} \right)^{-1/\rho}, \quad (15) $$

and as sectoral price index

$$ P_j = \left( \sum_{i=1}^I \chi_{ij}^{\rho} P_{ij}^{1-\rho} \right)^{\frac{1}{1-\rho}}. \quad (16) $$

Each firm produces output according to a Leontief production function as in the main text, but now capital is a variable input.

Cournot As there is only a limited number of firms in a sector, firms do have market power. Let’s assume that firms engage in Cournot competition, in which firms choose their quantities, $q_{ij}$, taken as given the quantities chosen by the other firms. Firms do internalize that their quantity supplied affects the sectoral quantity $y_j$. However, as each sector is infinitesimally small, firms take prices and the quantity $Y$ of final consumption as given.

The above implies that firm $ij$ solves the following maximization problem within a
period

$$\max_{P_{ij}, q_{ij}} P_{ij}q_{ij} - mc_{ij}q_{ij}$$

subject to the inverse demand function derived from (13) and (15):

$$\frac{P_{ij}}{P} = \chi_{ij} \left( \frac{q_{ij}}{y_j} \right)^{-\frac{1}{\rho}} \left( \frac{y_j}{Y} \right)^{-\frac{1}{\eta}}$$

where the firm takes into account that its quantity, $q_{ij}$, affects sectoral output, $y_j$, by (14). $mc_{ij}$ denotes the marginal cost of a firm and depends on its productivity. Due to constant returns to scale, total costs are simply equal to marginal costs times the quantity produced. Substituting out $P_{ij}$ from the objective, and taking the derivative with respect to $q_{ij}$ implies

$$P_{ij} = \frac{\varepsilon(s_{ij})}{\varepsilon(s_{ij}) - 1} mc_{ij} \mu_{ij}$$

where

$$\varepsilon(s) = \left[ \frac{1}{\rho} (1 - s) + \frac{1}{\eta} \right]^{-1}$$

and $s_{ij}$ is the market share of firm $i$ in its sector $j$.\(^26\) This gives that the markup is a function of the market share measured by sales,

$$\mu_{ij} = \frac{1}{1 - \frac{1}{\rho} - \left( \frac{1}{\eta} - \frac{1}{\rho} \right) s_{ij}}$$

with $\frac{1}{\eta} - \frac{1}{\rho} > 0$. Thus the larger the market share, the larger the markup. It does not matter whether changes in market share are due to changes in productivity or changes in the preference shifter.

**Bertrand** An alternative to Cournot competition is Bertrand competition in which the firm takes the prices, instead of the quantities, of other firms as given. This leads to the following constraint for the maximization problem (where I have now combined (13) and (15) to substitute out $y_j$):

$$\frac{P_{ij}}{P} = \chi_{ij} \left( \frac{q_{ij}}{C} \right)^{-\frac{1}{\rho}} \left( \frac{P_j}{P} \right)^{1-\eta/\rho}$$

Using this equation to substitute out the quantity, $q_{ij}$, in the objective, and taking the derivative with respect to $P_{ij}$, where the firm takes into account that changing its price

\(^{26}\)The derivation uses that $s_{ij} = \frac{P_{ij}q_{ij}}{\sum_i P_{ij}q_{ij}} = \frac{\chi_{ij}q_{ij}^{1-1/\rho}}{q_{ij}^{1-1/\rho}}$, where I have used (15) to obtain the second equality. Using (15) and (16), the market share can also be written as $s_{ij} = \frac{\chi_{ij}^{1-1/\rho} P_{ij}^{1-\rho}}{\sum_{i=1}^{N} \chi_{ij}^{1-1/\rho} P_{ij}^{1-\rho}}$.\(^{37}\)
affects $P_j$ by equation (16), gives

$$
\mu_{ij} = \frac{(\rho - \eta)s_{ij} - \rho}{(\rho - \eta)s_{ij} + 1 - \rho}.
$$

(23)

Thus, also Bertrand competition gives a positive relationship between changes in firm size and changes in markups.

**Counterfactual**

I first construct the counterfactual for the above Cournot model while only allowing for variation in productivity and not in the preference shifter. Then, I will also allow for variation in the preference shifter. I also construct the counterfactual for the Kimball model with variation in the preference shifter. The results are in Table 3.

**Table 3:** Counterfactual Haltiwanger reallocation effect of the markup over a 35-year horizon when the joint distribution of markups and market shares is not changing

<table>
<thead>
<tr>
<th>Model</th>
<th>Productivity shocks</th>
<th>Preference shocks</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Sales-weighted</td>
<td>Cost-weighted</td>
</tr>
<tr>
<td>Kimball</td>
<td>0.06</td>
<td>0.06</td>
</tr>
<tr>
<td>Cournot</td>
<td>0.09</td>
<td>0.08</td>
</tr>
</tbody>
</table>

I calibrate the Cournot model as follows. The production function is the same as in the main text. Namely, $\alpha_M = 0.88$ and $\alpha = 0.31$. The relationship between firm size and markups is determined by $\rho$ and $\eta$. I take the values of these parameters from Edmond et al. (2015) who obtain these by transforming the coefficients of a regression of the inverse firm-level markup on the market share. This means that I set $\rho$ to 10.5 and $\eta$ to 1.24. Furthermore, I set the number of firms in an industry to 7 to match that the sales-weighted markup is just below 1.4 on average. To match the AR(1) of sales, I set the autoregressive coefficient of productivity equal to 0.982 and the standard deviation of the productivity shock equal to 0.117. This gives a counterfactual reallocation effect of 0.09 for the sales-weighted markup and of 0.08 for the cost-weighted markup, which is slightly higher than the counterfactual reallocation effect of the Kimball model in the main text. When I allow for variation in the preference shifter while assuming that all firms are equally productive the counterfactual reallocation effect becomes 0.13 and 0.10 for the sales-weighted and cost-weighted decomposition respectively. The preference shifter here also follows an AR(1) with an autocorrelation coefficient of 0.986 and a standard deviation of the error term of 0.0585 to match the coefficients of the AR(1) of sales. Finally, the first row of Table 3 shows that with preference shocks the Kimball model gives a very similar counterfactual as when productivity shocks are driving firm churning.
Appendix B  Additional Figures

Figure 10: Haltiwanger decomposition of the markup in Compustat with three-year lags (the final period is based on a two-year lag)

Figure 11: Haltiwanger decomposition of the markup in Compustat when the output elasticity is estimated by a translog

Figure 12: Median markup as a function of firm size in Compustat. Both when the output elasticity does not vary by firm size (baseline) and when it does vary by firm size.
Appendix C  Solution Algorithm Kimball Model

I solve for the equilibrium as follows. Suppose we know the joint distribution of firm-level productivity $A_i$, capital $K_i$ and the preference shifter $\chi_i$. First, calculate the output, price and input use of each firm by solving the static firm optimization problem conditional on $A_i$, $K_i$ and $\chi_i$. To do so, first divide the static firm optimization problem by $DY$ to obtain

$$
\max_{q_i \geq 0} \chi_i \Upsilon'(q_i) q_i - w(q_i/A_i) \frac{1}{1-\alpha} A_i^{-\frac{\alpha}{1-\alpha}} K_i^{-\frac{\alpha}{1-\alpha}} Y^{\frac{\alpha}{1-\alpha}} D - \frac{q_i}{\alpha} \frac{P^M}{D},
$$

where $q_i$ is the market share $\frac{Y_i}{Y}$. This leads to the following first-order condition:

$$
\chi_i \Upsilon''(q_i) q_i + \chi_i \Upsilon'(q_i) = \frac{1}{1-\alpha} w q_i^{-\frac{\alpha}{1-\alpha}} A_i^{-\frac{1}{1-\alpha}} K_i^{-\frac{1}{1-\alpha}} Y^{\frac{\alpha}{1-\alpha}} D + \frac{1}{\alpha} \frac{P^M}{D}.
$$

(25)

That the markup equals $\frac{\Upsilon'(q_i)}{(1-\alpha) + \Upsilon''(q_i)}$ gives $\Upsilon''(q_i) = \frac{1}{\mu q_i} \Upsilon'(q_i)$. Plugging this into the first-order condition gives that

$$
\chi_i \Upsilon'(q_i) = \frac{\mu(q_i)}{1-\alpha} w q_i^{-\frac{\alpha}{1-\alpha}} A_i^{-\frac{1}{1-\alpha}} K_i^{-\frac{1}{1-\alpha}} Y^{\frac{\alpha}{1-\alpha}} D + \frac{\mu(q_i)}{\alpha} \frac{P^M}{D}.
$$

And using the Klenow-Willis aggregator gives

$$
\frac{\sigma - 1}{\sigma} \frac{1}{\sigma} q_i^{\frac{\alpha}{1-\alpha}} = \frac{1}{1-\frac{\alpha}{\sigma} q_i^{\frac{\alpha}{1-\alpha}}} \chi_i \left[ \frac{1}{1-\frac{\alpha}{\sigma} q_i^{\frac{\alpha}{1-\alpha}}} w q_i^{-\frac{\alpha}{1-\alpha}} A_i^{-\frac{1}{1-\alpha}} K_i^{-\frac{1}{1-\alpha}} Y^{\frac{\alpha}{1-\alpha}} D + \frac{1}{\alpha} \frac{P^M}{D} \right].
$$

Using this equation we can solve for $q_i$ for each combination of $A_i$, $K_i$ and $\chi_i$ conditional on the aggregate states $Y$ and $D$. Then, we can solve for the equilibrium values of $Y$ and $D$ by solving the following system of two equations

$$
\int \chi_i \Upsilon(q(A_i, K_i, \chi_i; Y, D)) dH(A_i, K_i, \chi_i) = 1
$$

40
\[ Y = \left( \int \left( \frac{q(A_i, K_i, \chi_i; Y, D)}{A_i} \left( \frac{K_i}{K} \right)^{-\alpha} \right) \frac{1}{1-\alpha} \frac{dH(A_i, K_i, \chi_i)}{K^\alpha L^{1-\alpha}} \right) \]

where \( Z \) is average (factor-augmenting) technology, and \( K \) and \( L \) are total capital and labor supply respectively.\(^{27}\) The first equation is the Kimball aggregator, and the second equation makes sure that \( Y \) is consistent with aggregate production.

This gives us the market share \( q(A_i, K_i, \chi_i) = q(A_i, K_i, \chi_i; Y, D) \) and markup. Knowing \( q_i \) and \( Y \) we can calculate firm-level output, use of variable inputs labor and materials, and flow profits conditional on \( A_i, K_i \) and \( \chi_i \). With this in hand we can solve the dynamic optimization problem of choosing investment \( I \) using value function iteration. Given the choice of investment we can calculate the joint distribution of capital, productivity and the preference shifter. Iterate until this joint distribution has converged.

\(^{27}\)To get the expression for aggregate technology, let \( Y = \min\{ZK^\alpha L^{1-\alpha}, \delta_M M\} \) and use that aggregate labor used in production equals \( L = \int L_i(A_i, K_i, \chi_i) dH(A_i, K_i, \chi_i) \). Substituting out \( L \) and \( L_i \) using the production functions and rewriting gives the expression for aggregate technology \( Z \).