Plant-to-Table(s and Figures): Processed Manufacturing Data and Measured Misallocation

Martin Rotemberg[*] and T. Kirk White[†]

[*]New York University
[†]Center for Economic Studies, U.S. Census Bureau

July 2021

Abstract

We describe differences between the commonly used version of the U.S. Census of Manufactures and what establishments themselves report. The originally reported data has substantially more dispersion in measured establishment inputs, output, and productivity. Even after trimming, measured allocative efficiency is substantially higher in the cleaned data than in the raw data: around 5x higher in 2002 and 2007, and 50x in 2012. Without trimming, the changes are substantially larger. We describe a Bayesian approach for editing and imputation that can be used across contexts, discussing how to incorporate analysts' manual edits and tax records, as the Census currently does.

The research in this paper was conducted while the second author was an employee of the Census Bureau. The views expressed in this paper are those of the authors and not the U.S. Census Bureau. All results have been reviewed to ensure that no confidential information is disclosed. The paper includes output approved for release under Disclosure Review Board release numbers DRB-B0115-CDAR-20181016, CBDRB-FY20-CES007-002, CBDRB-FY19-CMS-7909, CBDRB-FY21-CES007-005, and CBDRB-FY21-CES007-006. We are grateful to many colleagues and seminar audiences for their feedback. Thanks to Mitsu Nishida and Amil Petrin, with whom we started this project, to Hang Kim and Jake Blackwood for their helpful comments, to Allan Collard-Wexler for a thoughtful discussion, to Emek Basker for helpful comments and careful disclosure avoidance review, and to Lee Tanenbaum and Weitao Lin for their expert research assistance. We are also grateful to countless colleagues at the U.S. Census Bureau who helped us understand the underlying data. Contact information: mrotemberg@gmail.com & thomas.kirk.white@census.gov.

I Introduction

Over the past twenty years, many economists (including us) have written papers and lecture notes highlighting that the within-industry misallocation of factors can help explain cross-country differences in productivity (Restuccia and Rogerson, 2008; Hsieh and Klenow, 2009). The source of this belief is a robust stylized fact that developing countries like China and India have more measured within-sector dispersion in establishment behavior than wealthier countries like the US. This paper describes some challenges with the measurement of this stylized fact.

The confidence we have in our claims about dispersion in establishment behavior - either the "true" values for a particular country, or of cross-country differences - depends on how worried we are about measurement error (Bartelsman et al., 2021). In principle, measurement error has an ambiguous effect on measured dispersion, since non-classical noise can either push establishments' reported values towards or away from what is typical in their sector. In this paper, we discuss two potential sources of measurement error: establishments potentially misreporting their own characteristics, and subsequent data processing potentially introducing new errors. Across a variety of methods to clean the data, we find that measured allocative efficiency is multiple times higher in the cleaned data.

Most statistics agencies initially ask establishments to verify (or send in) information, but the subsequent steps vary across surveys. Many statistical agencies in developed countries, including the U.S. Census Bureau, both edit and impute responses. The exact procedures vary across industries and time (White, 2014), but broadly take two forms. First, the Census Bureau *edits* some outliers. If a reported variable fails one or more edit rules, then it may be temporarily replaced with a missing value. Second, the Census Bureau *imputes* missing information, using other information reported by the plant (both

in that year and, when available, in previous years) and other plants in the same industry.[1] For 2002, 2007, and 2012 we have access to the original values and cleaned values for plants in the Census of Manufactures, as well as the relevant edit flags.

First, we describe extensive margin changes: the extent of edit flags for total value of shipments,[2] capital, payroll, and materials is large: around 80 percent of plants have a value in the final data that is different than in the raw "captured" data.[3] Both fixes of obvious reporting errors (around half of plants have at least one missing value, around 10 percent report distinct values for the same outcome, which leads to a "logical" edit), as well as more subjective changes (such as manual edits by industry experts, which affect five percent of plants) have large effects on measured dispersion. Across all four primitives, the captured data has thicker upper and lower tails than the final data, although the final values tend to be larger than the captured ones. Around a third of plants have an edited value that is at least 10% larger or smaller than the originally reported value.

We then turn to describing how measures of productivity change. Many measures of misallocation use statistics that are broadly similar to the spread in revenue productivity (Bartelsman and Wolf, 2018; Asker et al., 2019).[4] We show visually that cleaning has large effects on the distribution of TFPR, dramatically lowering the mass in the tails.

[1] Editing a variable normally involves deleting the original response, and then treating it as if it had never been reported.

[2] Correctly specified, gross output is not equal to total value of shipments but needs to be adjusted for changes in inventories (this adjustment also should be made for materials). However, because inventories are often missing (White et al., 2015) and their contribution to output is normally fairly small, we do not make this correction for the U.S. data. Otherwise the share of plants affected by editing and imputation would be much larger.

[3] There are well-known conceptual and practical concerns with the measurement of capital (Hicks, 1981; Hulten, 1991; Kehrig, 2015; Collard-Wexler and De Loecker, 2016). Capital measurement does affect our results, but it is not the primary driver. A majority of establishments have at least one characteristic besides capital which is affected by the cleaning process.

[4] Revenue productivity (TFPR) and quantity productivity (TFPQ) are somewhat complicated to measure, since neither production function elasticities nor quality-adjusted quantities are directly reported by plants (Olley and Pakes, 1996; Levinsohn and Petrin, 2003; Demirer, 2020). We focus on simple methods to estimate TFPQ and TFPR (Foster et al., 2008): using cost shares by 6-digit-NAICS industry to back out production functions, and using the structure of CES demand to translate from revenues to quantities (Hsieh and Klenow, 2009).

Some distribution statistics are unambiguously sensitive to tails, and indeed we find that the standard deviation of TFPR falls by half in the cleaned data. However, the data processing undertaken by the U.S. Census Bureau does not only affect the tails: it lowers the measured interquartile range of TFPR by almost as much. Both the effect of data processing and measured dispersion (in the raw & cleaned data) are increasing over time.[5]

The average absolute difference between captured and cleaned TFPR for the largest (or oldest) plants is around 2/3 of that for the smallest (or youngest) plants, with a fairly monotonic relationship in between. Nevertheless, there is still a large difference between captured and cleaned productivity even for the largest and oldest plants. Across the distribution, around a third of plants have a TFPR value that is lower in the final data, and around a sixth have a larger value in the final data.

In order to translate the raw variance and covariance numbers into aggregate productivity losses, we use the same type of model as in Bils et al. (2021) and Blackwood et al. (2021), where establishments have Cobb-Douglas production functions and face idiosyncratic distortions on their inputs, while consumers have CES demand.[6] This generates relatively intuitive relationships between aggregate productivity losses and frictions or distortions. Without trimming, using the originally reported data instead of the final data lowers measured allocative efficiency by a factor of a thousand (with 1 percent trimming, the captured value is about twenty times different than the final data). The exact measured productivity loss is sensitive to institutional assumptions, for instance the extent of roundabout production and the returns to scale (Haltiwanger et al., 2018; Blackwood et al., 2021). Across a variety of such specifications, we continue to find large changes in

[5] We cannot say with certainty if the Census Bureau's data cleaning is getting us closer or farther from the truth. However, we do know that treating singly-imputed data as if it were true data underestimates the amount of uncertainty in the resulting estimates (Rubin, 2004).

[6] There are other reasonable and commonly used measures of productivity dispersion as well, such as the covariance of TFPQ and TFPR (Hopenhayn, 2014) and the relationship between the size-weighted and unweighted average productivity in the economy (Baily and Hulten, 1992; Olley and Pakes, 1996; Foster et al., 2001). We also show how data processing affects those measures.

measured U.S. misallocation between the captured and final data. We do not take this result literally - we do not think that captured data gives compelling evidence that the U.S. manufacturing sector is characterized by a thousand (or twenty) times more misallocation than had been previously understood. Instead, we consider our results a "smoking gun" that measurement (and data processing in particular) is important to the study of misallocation.

Given that the overall cleaning effort has a large effect on measured misallocation, we return to describing the specific process of cleaning the data, by measuring how much each type of edit affects measured dispersion. We do so by describing their Shapley (1953) value; essentially the share of the total change that can be attributed to each type of edit. The most important edits are the logical and analyst edits, which jointly explain around two fifths of the total change.

Given the importance of these edits, it is difficult to interpret measured cross-country differences that use different processing methods. Many types of edits done by the U.S. Census Bureau are infeasible in many other countries. For instance, administrative data on payrolls requires a broad payroll tax base. However, even across Europe, only some countries augment their manufacturing surveys with tax data (Bartelsman and Wolf, 2018). Even beyond administrative edits, the Indian data is not edited at all in the central office.[7]

Given the extent of measurement error, it is also difficult to trust differences in unprocessed data. Since we are not nihilists, we describe the effect of methods which com-

[7] We have confirmed that there was no editing or imputation of the Indian data both in the data documentation, and in email communications with the Ministry of Planning and Statistics. O.P. Sharma, a former Deputy Director of Census Operations in India, writes "although the data are thought to be characteristic of firms in the organized sector, there are important caveats. ASI [Annual Survey of Industries] survey data are presented in raw form without adjustments to the ways that employers reported them; there are no attempts to contact employers to fill in missing or incomplete data or to correct for data that seem out of line with other data" (Sincavage et al., 2010). There is some imputation of prices in the Indian data post-2006, but we do not use the price information in our results.

monly clean data across contexts. Trimming is a popular and straightforward approach to data cleaning. However, it varies paper to paper, even within datasets. Hsieh and Klenow (2009) trim static measured distortions, while Bils et al. (2021) trim both productivity measures and growth rates.[8] Trimming is also a blunt instrument, and may remove correct responses from the data. For instance, in the U.S., the analyst-verified values are often relatively extreme values. When we trim the final data,[9] we are around twice more likely to remove observations that have been verified by a professional analyst, raising doubts if trimming the data brings us closer to or farther from the truth.[10]

We describe and implement a more reproducible approach than trimming: a theoretically-motivated data cleaning exercise which could then be used across establishment-level datasets without further need for data processing (Kim et al., 2015). Unlike trimming, which drops outliers from the sample, the Kim et al. (2015) method uses a Bayesian statistical model to simultaneously edit and impute the data. First, we look at the ratios of reported variables and flag the outliers of the ratios–this is a standard first step in the literature (Fellegi and Holt, 1976; Thompson et al., 2004). This step unfortunately remains a little ad-hoc across countries. We use the actual bounds used by the U.S. Census Bureau, and try to approximate their equivalents in the Indian data.[11] We then impute entries in order for the cleaned data to pass the edit checks. Unlike the imputation methods the Census Bureau uses most frequently in the Census of Manufactures, the Kim et al. (2015) method uses a Bayesian nonparametric approach (Ishwaran and James, 2001; Kim et al.,

[8] Nishida et al. 2015, Allcott et al. (2016) and Martin et al. (2017) also trim growth rates in the Indian ASI, but using different rules.

[9] We trim the extremes of TFPR and TFPQ.

[10] Laws governing access to confidential Census Bureau and IRS data prohibit confirming or denying the existence of any particular firm in the data. However, given what we know from publicly disclosed research, it seems plausible that outliers in business microdata (not just in Census data) may be especially important in an era of "rising superstar firms" (Autor et al., 2020).

[11] There are many reasons why it is impossible and unhelpful to use the same bounds in the US and India. One reason is that the US bounds are industry specific, and the industry classifications are different across countries. Another is that some of the bounds are nominal (e.g. wage bill per worker), for which the underlying units are different across countries.

2014) to favor making edits that are likely given the model for misreporting, and similarly impute values that are likely given the underlying model for the data. The imputation step works for missing values as well as those which are flagged as outliers. Because the model uses probabilistic imputation, we can draw many implicates from the estimated model in order to show uncertainty in our results (Rubin, 2004). We find that the amount of uncertainty in the Bayesian-edited data has been increasing over time.[12]

While the baseline Bayesian editing and imputation approach does not take advantage of the entire set of resources available to the U.S. Census Bureau, it nevertheless has some appealing properties. Around a third of plants in the data have at least one of materials or total value of shipments imputed using regression methods. Both for the edits and imputes, the resulting variance in the final data is smaller in the regression imputed data than for the rest (White, 2014). This is less true for the data imputed following Kim et al. (2015), suggesting, perhaps not surprisingly, that the regression-based imputation methods under-state the variance in the data (White et al., 2018; Van Buuren, 2018).

An important addendum to our results: there are also large undocumented differences in enumeration across countries. In the United States plants can fill out the form online, [13] while in India (especially in the early 2000s), the vast majority of plants do not report having a computer, and many plants need the help of enumerators.[14] In neither country is it documented in the data when statistics agencies contact plants to get updated responses (in the U.S. it is considered an unedited response if, after being prompted, a

[12] The method can handle richer data generating processes than we use in our main results. We show how the results change when we feed in additional sources of data for the U.S., in particular the high-quality administrative information available for revenues and labor, and a "hybrid" that additionally allows for the manual analyst edits.

[13] Until the 2017 Economic Census, smaller plants also had the option to fill out paper questionnaires. Starting in 2017 the Economic Census moved to all-online data collection except in Puerto Rico and other island areas.

[14] For the relevant variables, the US survey autofills units as in thousands of dollars, which some respondents may not notice and therefore effectively overstate their size for those variables.

plant responds with updated values),[15] so we do not know either the extent of verification (although, in general, analysts in the U.S. try to get validated responses from the biggest plants in lieu of editing). We also do not know the extent to which bookkeeping practices vary across contexts (Barrios and Gallemore, 2021; Zwick, 2021; Almunia et al., 2021).

Economists have a long tradition of studying (mis)measurement (Frisch, 1934; Griliches, 1974). Our paper is firmly in the spirit of Romer (1986a,b, 1988, 1989) and Balke and Gordon (1989), who study how differences in data quality matter for understanding the historical incidence of business cycles and unemployment (Burns, 1960).[16] We also complement current efforts to understand measurement error in labor force surveys (Abowd and Stinson, 2013; Kambourov and Manovskii, 2013; Meyer and Mittag, 2019; Medalia et al., 2019; Vom Lehn et al., 2020). In the misallocation (Banerjee and Duflo, 2005; Restuccia and Rogerson, 2008; Hopenhayn, 2014) literature in particular, Velayudhan (2018), Gollin and Udry (2019), and Esfahani (2019) study the role of measurement. Their approaches use economic theory to distinguish between measurement and "true" misallocation, for instance by arguing that farmers are unlikely to misallocate resources across their own plots, or that firms misreport to avoid additional taxes. Bils et al. (2021) explicitly studies measurement error in manufacturing, leveraging theory and panel data in order to argue that measurement error is increasing in the U.S. (using the already-processed final data).[17] One value of our approach is we provide direct evidence of measurement error. Furthermore, researchers can use our proposed data cleaning procedure regardless

[15] Sincavage et al. (2010) say that re-contacting never happened in India during the sample period of our data, although informally we have been told that it does happen currently.

[16] Similar issues have recently been discussed in the asset pricing literature, since measuring the correlation of consumption with asset prices is difficult (Savov, 2011; Kroencke, 2017). Data from international tests also can be imputed (Jerrim et al., 2017; Troccoli, 2020).

[17] The intuition of their theoretical result is that measurement error lowers the correlation of *changes* in inputs and revenue, which they can then use to correct the naive cross-sectional measurement of misallocation. The captured data for the annual panel surveys that Bils et al. (2021) use are not available, so we cannot speak directly to their results. We describe in Section IV.F other differences.

of their question, and even in settings with less extensive data collection and cleaning efforts than the US.

II A Theory of Misallocation

In this section, we briefly recap the theory of misallocation that underpins our empirical results (Bils et al., 2021; Blackwood et al., 2021). In order to introduce the notation and the intuition, in the main text we focus on describing the main outcomes in the paper. In Appendix D we describe how the calculation changes when we include sector-specific demand elasticities, non-constant returns to scale in production, and roundabout production. (Blackwood et al., 2021).

The wage, rental rate, and intermediate price are constant in the economy, but plants in each sector $s \in S$ face idiosyncratic distortions on each input.[18] As a result, each plant's profits are:[19] $\pi_{si} = P_{si}Q_{si} - \left(1+\tau_{L_{si}}\right) wL_{si} - \left(1+\tau_{K_{si}}\right) RK_{si} - \left(1+\tau_{M_{si}}\right) p_M M_{si}$.

Demand is CES, so profit maximization implies that the plant's output price is a fixed markup over its marginal cost. Within a sector, the input share of each input will be proportional to the distortion, and there is complete pass-through of improvements in A_{si} to prices. As a result, revenue productivity, $TFPR_{si} = \frac{P_{si}Q_{si}}{\left((rK)_{si}^{\alpha_s}(wL)_{si}^{1-\alpha_s}\right)^{\gamma_s}(p_M M)_{si}^{1-\gamma_s}}$, only varies due to the distortions. In particular,

$$TFPR_{si} = \frac{\sigma}{\sigma-1}\left[\left(\frac{\left(1+\tau_{K_{si}}\right)R}{\gamma_s\alpha_s}\right)^{\alpha}\left(\frac{\left(1+\tau_{L_{si}}\right)w}{\gamma_s\left(1-\alpha_s\right)}\right)^{1-\alpha}\right]^{\gamma_s}\left[\frac{\left(1+\tau_{M_{si}}\right)p_M}{\left(1-\gamma_s\right)}\right]^{(1-\gamma_s)}. \quad (1)$$

[18] We try to keep our notation standard - plant i in sector s produces quantity $Q_s i$ which it sells at a price of P_{si}. It uses three inputs, labor (L_{si}), capital (KL_{si}), and intermediate materials (ML_{si}), which respectively have prices w, r, and p_M. Plant si faces an exogenous multiplicative distortion for each input x $(1+\tau_{x_{si}})$ which generates a wedge between the national input price and the price paid by plant si.

[19] While we focus most of our attention on gross-output production functions, Hsieh and Klenow (2009) use a value added specification, and we show results for value added models as well. Without roundabout production, there is no conceptual difference between materials and the other inputs.

TFPQ can be inferred by taking advantage of the fact that the markup is known:

$$TFPQ_{si} \equiv A_{si} \propto \frac{(P_{si}Q_{si})^{\frac{\sigma}{\sigma-1}}}{\left((RK)_{si}^{\alpha_s}(wL)_{si}^{1-\alpha_s}\right)^{\gamma_s}(p_M)^{1-\gamma_s}} \tag{2}$$

Aggregate productivity is a CES aggregator of TFPQ and (relative) TFPR,[20]

$$TFP_s = \left(\sum_{i \in s} A_{si}^{\sigma-1} \widetilde{TFPR_{si}}^{1-\sigma}\right)^{\frac{1}{\sigma-1}}. \tag{3}$$

Since we know from Equation 1 that $\widetilde{TFPR_s}$ is equal to $TFPR_{si}$ unless plants have idiosyncratic distortions, the "efficient" counterfactual TFP is $\overline{A}_s = \left(\sum_{i=1}^{M} A_{si}^{\sigma-1}\right)^{\frac{1}{1-\sigma}}$. The ratio of observed to potential TFP is our measure of aggregate productivity:

$$\widetilde{TFP_s} = \left(\sum_{i \in s} \widetilde{TFPQ_{si}}^{\sigma-1} \widetilde{TFPR_{si}}^{1-\sigma}\right)^{\frac{1}{\sigma-1}}. \tag{4}$$

For our main outcomes, we calculate a Cobb-Douglas aggregator over the sectors using the gross output share of each sector (θ_s):

$$\widetilde{TFP} = \prod_{s \in S} \widetilde{TFP_s}^{\theta_s} \tag{5}$$

The three main outcomes that we describe in the paper are $\ln(\widetilde{TFPR_{si}})$, $\ln(\widetilde{TFPQ_{si}})$, and $\widetilde{TFP}$. For notational convenience, although we present results for log and industry-scaled productivity measures, we describe them as "TFPR" and "TFPQ" in our results.[21] We also describe the distributions of inputs, output, and the input shares (which we nor-

[20] We define sectoral $\widetilde{TFPR_s}$ as the revenue weighted harmonic mean of $TFPR$, and $\widetilde{TFPR_{si}} \equiv \frac{TFPR_{si}}{\widetilde{TFPR_s}}$. See Equation A. 3 for the formal expression.

[21] When we report distributional 90/10 and 75/25 "ratios," we calculate the log-differences within each industry for the respective percentiles and report the average across all industries.

malize using industry averages). We describe which types of establishments experience a larger change in measured characteristics when comparing the captured to the final data. We focus on two measures, the age and the number of employees of the business. For both, we run a local polynomial regression comparing the difference (either regular difference or the absolute value of the difference) between TFPR measured in the cleaned data vs. the captured data.

Instead of measuring how sensitive the measures of aggregate productivity are to different underlying economic assumptions, which has been the primary focus of much of the recent methodological literature on misallocation (Asker et al., 2014; Haltiwanger et al., 2018), we focus on calculating productivity using different versions of the data, including introducing several alternative edited and imputed datasets. In Appendix D, we describe how data cleaning affects the measured aggregate productivity under alternative assumptions on demand elasticities, roundabout production and the returns to scale, following Blackwood et al. (2021).

II.A Alternative Measures of Misallocation of Factors

While our focus is in the Hsieh and Klenow (2009) tradition, there are several other measures of dispersion which also relate to the allocation of factors. The most well-known alternative approach uses the revenue-share weighted average of productivity as a measure of aggregate productivity (Baily and Hulten, 1992; Foster et al., 2001; Bartelsman and Wolf, 2018), and then decomposes it into "within" and "between" terms, where the latter is interpreted as a measure of allocation (Olley and Pakes, 1996). We consider the simplest version of this measure: the share of the weighted average productivity driven by the unweighted average: $\frac{\frac{1}{N}\sum TFPR_i}{\sum \frac{P_iQ_i}{\sum P_jQ_j} TFPR_i}$.

We additionally consider two somewhat more model-free correlations in the data: between labor expenditures and output, and between TFPR and TFPQ (for both measures,

we show the R^2 of a regression after residualizing six-digit-industry fixed effects). The former is a measure of dispersion in labor productivity (inspired by Cunningham et al. 2018, although they report output per hour instead of output per wage expenditure), the latter is a simpler way of thinking about what is driving Equation 5, inspired by the discussion in Blackwood et al. (2021).

III Data Cleaning in the United States

We primarily use micro-data from the United States, from the 2002, 2007 and 2012 U.S. Censuses of Manufactures (CMF), augmented with capital constructed by Cunningham et al. (2018).[22] We describe the data sources in more detail in Appendix Section A, and present sample sizes for our main tables in Appendix Table 25.

As in most surveys, not all respondents to the CMF answer all of the questions, and some responses are inconsistent with each other or inconsistent with administrative records data (e.g., IRS payroll tax records) from the same firms. The Census Bureau has created imputation and edit rules for this data, the development of which are described in Sigman (1997) and Thompson and Sigman (1999). However, until recently, it was difficult for researchers to identify which, if any, responses for a given plant were imputed.[23] We go beyond the imputation flags and use the newly available actual responses from the establishments themselves.

We focus on the four variables used in Section II to measure plant-level total factor productivity: total value of shipments, total cost of materials (which includes the cost of

[22] For some of our analysis, we use data from the Annual Survey of Manufactures (ASM) subsamples in the CMFs, revenue data from the Census Bureau's Business Register, and employment and annual payroll from the Longitudinal Business Database (LBD).

[23] There is a tradition of researchers trying to back out imputations from the final data, see for instance Davis and Haltiwanger (1991), Roberts and Supina (1996) ,Syverson (2004a,b), and Collard-Wexler (2011). It was easiest to identify "hot deck" imputes, which led to duplicated observations. There are no hot deck imputes in the data we use, but item-level edit/impute flags for the 1987 and later censuses are available to researchers in the Federal Statistical Research Data Centers (FSRDCs) (White, 2014). It is worth noting that when Hsieh and Klenow (2009) was published, neither imputation flags nor the original plant responses were available.

energy), total wages, and the capital stock. The first three variables are directly measured in the Census of Manufactures, and are present in both the raw data and the final data. Capital is more difficult to measure: the plants report fixed assets at the beginning and end of the year, but for productivity what matters is the flow of capital services. We use the measures of real stock of capital carefully constructed by Cunningham et al. (2018), multiplied by nominal rental rates to measure capital flows (Kehrig, 2015).[24]

The captured data differs from the final data in two respects. First, missing values due to non-response in the reported data are imputed in the cleaned data,[25] using a variety of industry-specific regression-based and other imputation strategies. Second, responses which fail edit rules in the reported data are normally imputed or changed in the final data. The Census Bureau primarily uses two types of edit rules: balance rules, which require entries to add up[26] and a set of ratio edit rules which bound the ratios of any two variables. For the CMF, the Census Bureau develops upper and lower bounds for 12 contemporaneous variables (9 ratios) in every industry; these so-called explicit ratio bounds imply other implicit bounds (Fellegi and Holt, 1976).[27]

Edit-rule-failing responses are replaced using a variety of methods, described in Appendix Table 1. There are too many types of edit categories for us to describe the effects of them individually (plus several are fairly rare), so we group them into in eight categories (plus three residual categories) in Table 1, describing the categories in more detail in Appendix Section B. We define "replicable" edits as ones which (generously) could be

[24] We use the ratio of captured and final book assets to calculate an equivalent measure for the captured data. See Appendix A for more details.

[25] Throughout the paper we use "final data" and "cleaned data" or "Census-cleaned data" interchangeably. Likewise we use "raw data", "captured data", and "reported data" interchangeably.

[26] The balance rules cover total cost of materials (must equal the sum of five other variables), total inventories beginning and end of year (sum of three components each), average number of production workers (sum of quarterly production workers divided by 4), total employment (sum of production workers and non-production workers), and total salaries and wages (sum of production worker wages and non-production worker wages).

[27] The Census Bureau does statistical analysis of outliers in the data itself to determine the industry-specific ratio bounds for pairs of highly correlated variables (Thompson and Sigman, 1999).

done in the Indian data.[28] Table 1 additionally describes the share of plants affected by each edit category.[29] The most common types are imputes for missing, logical edits for payroll,[30] and analyst edits.[31] Including both edits and imputes for missing data, around a third of plants have at least one variable in the final data which is calculated using regression-based methods. We formally describe the logic of data cleaning in Section V.

Table 2 Panel A describes the overall exposure of plants to edits and imputes. Around 80 percent of plants have at least one value in the final data that differs from its captured counterpart.[32] Around a quarter of the plants with any change have exactly one change. Half of plants have a missing value.[33] The any-missing share is around 10 percent larger in 2012 than 2002. Appendix Table 5 shows that conditional on no missing values, around two fifths of plants have no edited values.

There are a few important parameters that are not directly reported in the data: the production function elasticities and the demand elasticity of substitution. We use cost shares for the former (and do not update the elasticities when using different cuts of the data), and 4 for the latter (Redding and Weinstein, 2020).[34]

[28] Just because an edit type is possible does not mean in practice that it would be used. For instance Bartelsman and Wolf (2018) note that only some European countries use administrative records to clean manufacturing census data, and even the ones who do use different approaches.

[29] Table 1 shows the average share over 2002, 2007, and 2012. The underlying year-specific values are in Appendix Table 2.

[30] Logical edits are described more formally in Appendix Section B, but are related to the balance rules - with redundant question, some values can be imputed with a linear combination of others.

[31] The flags indicate the type of model but not the finer details, for instance while the "B" flag denotes a regression edit, it does not indicate the relevant coefficients, which can vary within industry-years.

[32] White et al. (2018) and Foster et al. (2017) report different values. They do not consider capital but do consider edits to inventories (the former) and production hours (both).

[33] While our sample is the mail sample of plants that are supposed to fill out a survey, around a fifth of plants are entirely imputed. See Gauthier (2011) for a discussion of efforts the U.S. Census Bureau made to prevent delinquency in the 2007 Economic Census.

[34] We use an elasticity of substitution of 3 for the Value Added results, as in Hsieh and Klenow (2009), and show our results are similar using industry-specific elasticities of substitution (Ahmad and Riker, 2019).

IV Changes in Reported Variables

We start by describing changes to the primitives: the book value of assets (capital), the cost of materials, payroll, and the total value of shipments.[35] The results are in Table 2 Panel B, where we describe the changes at the variable level. Capital is the most likely to be changed, especially in 2007, although all variables are changed for at least ten percent of plants.[36] While capital is most commonly changed, Panel A shows that in every year over half of plants have a value changed for a characteristic besides capital. A large share of the changes are sizeable: around a third of plants have at least one change that is over 10% away from the originally reported response. While the capital changes are most likely to be big, for all characteristics the edits are larger than ten percent around five percent of the time.

IV.A Trimming the final data

It is not uncommon for researchers to trim outliers, even in the already-cleaned final data. For instance, Hsieh and Klenow (2009) trim the tails of plant productivity and distortions in each country-year.[37] When we explore trimming, we follow Blackwood et al. (2021).[38] Worryingly, trimming drops data that seems fairly reliable; plants with analyst-verified information are about about twice as common in the "trimmed" part of the data.[39] One value of having the administrative flags now available in the FSRDCs is that even naive

[35] Throughout the paper we use the terms "payroll", "salaries and wages" and "total wages" interchangeably, and use "payroll" in the tables.

[36] In 2007 in particular, many respondents did not respond that their end of year assets were equal to beginning of year assets plus capital expenditures less capital retirements and depreciation (due to a misinterpretation of the questions) (White, 2014). As a result, almost every plant has an edit flag for capital that year.

[37] Bils et al. (2021) and Blackwood et al. (2021) only trim the productivity outliers, although (Bils et al. 2021 additionally trim plants with big changes and update the production function elasticities after trimming.

[38] Trimming the 1 percent upper and lower extremes for TFPQ and TFPR shrinks removes around 3 percent of the plants and trimming 2 percent removes around 6 percent of plants. This is because there isn't perfect overlap at the extremes of each distribution.

[39] This value has been falling over time: the share of analyst-verified values is about 2.5 times as common in the trimmed portion of the data in 2002, and around 1.5 times as common in 2012.

trimming can be supplemented by refusing to change analyst-verified values.

IV.B Changes in Dispersion of Inputs and Output

Table 3 reports the distribution of the captured/final ratios calculated for Table 2.[40] Across all variables, when different, the captured values tend to be lower than their cleaned counterparts: The tenth percentile tends to be lower than one, while only the ninety-fifth tends to be above. For all variables —especially payroll and capital —the captured and final values are the most different in 2012.

In Table 4 we describe the effect of data processing on the dispersion of the plant characteristics themselves. We report the dispersion of each characteristic, normalized by the industry mean.[41] Table 4 shows the value in the captured data divided by the value in the final data, Appendix Table 3 shows the underlying dataset-specific values. It is perhaps not surprising that the standard deviation of the variables is consistently around 20 percent larger in the captured data, shown in Column 3. Columns 1 and 2 show that this is not just the behavior of extremes in the data, as the 90/10 ratio and interquartile range are also substantially larger in the captured data.

The Census Bureau's decisions about which variables are in error are based on ratios, not levels, of characteristics. So too does the model of misallocation - what matters is the input/output ratios, which are proportional to the distortions in Section II. Appendix Table 4 shows the same dispersion statistics as Table 4, but for the input/output ratios (so, for instance, Panel C of Appendix Table 4 shows the results for labor productivity pQ/wL.) The input shares have relatively more dispersion in the captured data than in the final data, and their dispersion is also increasing over time. Column 4 of Table 4 shows how related the same variable is in the captured and final data. We ran a regression of

[40] In order to prevent disclosure of information for any particular plant or firm, when we report the value "xth percentile", we calculate the average value of all plants in the xth centile.

[41] For instance, for total value of shipments we report the dispersion of $\frac{P_{si}Q_{si}}{\frac{1}{N_s}\sum P_{si}Q_{si}}$.

the (ln) variable in the captured data on its final counterpart, with industry fixed effects, and reporting the within R^2. The R^2 is lowest for capital, and highest for payroll and shipments, and for all four is falling over time.

IV.C The Distribution of TFPR

We now turn to the main object of interest: productivity. Before quantifying the changes, it is useful to give a visual sense of how they change. In Figure 1, we plot the density of TFPR in 2002, 2007, and 2012.[42] The difference in the distributions is not limited to outliers. There is substantially more spread of TFPR in the captured tails (especially in the upper tail), with the spread increasing in 2012.

In Figure 2, we plot the relationship between the absolute difference in measured TFPR between the captured and final data, focusing on firm age and firm-level employment since those are well measured at the firm level (Decker et al., 2020).[43] The pattern is similar for the three years for which we have data; the absolute difference is consistently falling in both firm age and firm size. The average absolute difference for the largest firms is large - and is often around 50% - but it is larger for the younger and smaller firms. The absolute gap could be positive even if the cleaning were mean zero.[44]

[42] Since we normalize by $\widetilde{TFPR_s}$, it is not mechanically the case that the distribution be centered around 0 (nor centered around the same value across datasets). In the interest of space, we only plot figures for TFPR.

[43] To comply with Census Bureau disclosure avoidance practice, we drop the 5% left and right tails from these graphs. We use firm characteristics on the x-axis. Firm age comes from the Census Bureau's Longitudinal Business Database (Chow et al., 2021). That dataset begins in the year 1976, so the age of firms born before 1976 is censored. According to the Census Bureau's Business Dynamics Statistics (which are tabulated from the LBD) 23% of manufacturing firms' ages were censored in 2002, 20% were in 2007, and 18% were in 2012.

[44] Appendix Figure 1 shows that the edited data consistently has larger measured TFPR than the captured data. The size of the average gap is decreasing in firm size and age, although the pattern for both, especially firm age, is somewhat flatter than for the absolute difference.

IV.D Dispersion of Measured Productivity

In this section, we describe the effects of data cleaning on alternative measures of productivity dispersion. As in Table 4, the values are the ratio of the value in the captured data to the corresponding value in the final data. The underlying values are in Appendix Table 6. The first two columns in Table 5 report the 90/10 ratio and the interquartile range, and the third shows the standard deviation. The relative dispersion is much bigger for measured productivity than it was for any of the directly reported inputs shown in Table 4. For TFPR, the standard deviation in the cleaned data is around half of that in the original data, and it's around a third for TFPQ. For all three measures, and for both TFPQ and TFPR, the ratio is largest in 2012 (Appendix Table 6 shows that this is mostly due to changes in the captured data, although TFPR dispersion is also increasing in the final data). Column 4 of Table 5 shows the within R^2 of a regression of (ln) variable in the captured data on its final counterpart, with industry fixed effects.[45] The R^2 for productivity are substantially lower than those for the directly reported inputs, with only around a third of the residual variance in captured TFPR explained by final TFPR.

Taken together, Tables 4 and 5 show that the cleaning undertaken by the U.S. Census Bureau is more nuanced than trimming tails, since it affects ratios of interior quantiles. Bartelsman and Wolf (2018) explicitly justify focusing on quantiles in order to avoid measurement issues. Our results suggest that in the U.S. data, this may not be sufficient.

IV.E Measured Misallocation in the Raw U.S. data: alternative approaches

Appendix Table 7 shows how data cleaning affects a few alternative measures of factor allocation, outside of the Hsieh and Klenow (2009) tradition. Panel A shows the effect on the "within" Olley and Pakes (1996) measure: $\frac{\frac{1}{N}\sum_i TFPR_i}{\sum_i s_i \times TFPR_i}$, where s_i is plant i's share of aggregate gross output. This value is essentially zero in the captured data, and close to

[45] Since there are industry fixed effects, the normalization doesn't affect the R^2.

one in the final data, leading to opposite implications about the role that the covariance of size and productivity (the residual of the value shown in the table) has for aggregate TFP.

In many models with no distortions (e.g. Melitz 2003; Boar and Midrigan 2021),[46] within a sector inputs and output are strongly (often perfectly) correlated, since for any given size the more productive plant will have a higher marginal product of labor. Appendix Table 7 Panel B shows the relationship between (ln) payroll and (ln) shipments and Panel C shows the relationship between all three (ln) inputs and (ln) shipments. For both panels, we run regressions with 6-digit-NAICS fixed effects, and report the within-R^2. Payrolls are around 20 percentage points worse at explaining shipments in the captured and final data, the difference is about 10 percentage points for all inputs. For both outcomes, the gap between datasets is growing over time.

Panel D of Appendix Table 7 shows the within-R^2 of an equivalent regression comparing TFPQ and TFPR. As Hsieh and Klenow (2009) point out, with constant markups, TFPQ and TFPR are uncorrelated within sectors, since falling prices cancel out rising productivity. We find this this pattern is substantially stronger in the final data: the within-R^2 is around 50 percent larger in the captured data.[47]

IV.F Measured Misallocation in the Raw U.S. data

For our final set of results comparing the cleaned and captured data, we use the model described in Section 2. While further removed from the raw data than Tables 4 and 5, the advantage of the calculation is that it gets closer to thinking about (measured) welfare costs: in most models, frictions are particularly important if they affect establishments' plant size ranking (Hopenhayn, 2014), which in the undistorted equilibrium is a function of (only) TFPQ. First, we consider the effects on measured misallocation of replacing

[46] See, for instance, Holmes and Stevens (2014) for an alternative to what they call the "standard model."

[47] See Haltiwanger et al. (2018) for a discussion about the correlation between TFPQ and TFPR and model misspecification.

captured data with cleaned data in the U.S. manufacturing sector in 2002, 2007 and 2012. Papers in this literature (Hsieh and Klenow, 2009; Bils et al., 2021; Blackwood et al., 2021) tend to calculate the extent of misallocation in trimmed data. For comparability, we calculate aggregate productivity not only in the untrimmed data, but also after trimming the extremes of TFPR and TFPQ.[48] Table 6 shows the results of calculations across a range of trimming percentages, showing the ratio of allocative efficiency in the captured to final data (the underlying dataset-specific values are in Appendix Table 8). Panel A shows the values for a gross-output model, Panel B for a value-added model, and Panel C uses only plants in the Annual Survey of Manufacturers sample (with the weights).[49] In the Census, for both Panel A and B, measured allocative efficiency in the captured data is never above .1 percent of the value in the final data (the ratio is somewhat higher in the ASM). In 2012, the gap is the largest: gross output allocative efficiency in the captured data is 0.00007 of the value in the final data. Trimming outliers increases the ratio by over a factor of 100, but measured allocative efficiency in the captured data is never above a third of its final data counterpart in the Census (again, the ratios are higher in the ASM), and the gap is largest in 2012. The difference between the final and captured data are substantially larger than the change from the U.S. to any South American (Busso et al., 2013) or Sub-Saharan African (Cirera et al., 2020) country found in the literature. It is also substantially larger than the gap found by Bils et al. (2021) comparing the U.S. to India, both before and after

[48] Differences in sample selection, industry definitions, and exact trimming definitions (see subsection IV.A) means in practice that our measured values in the final data are different than others. For instance, we use current 6-digit-NAICS codes, Bils et al. (2021) use constant Fort and Klimek (2016) 3-digit NAICS, and Blackwood et al. (2021) use 4-digit SIC codes (which correspond to 6 digit NAICS, but there have been classification changes). Similarly, our sample in the final data is larger than in earlier working paper versions of this paper, since we no longer constrain the data to be as balanced across samples.

[49] In the Annual Survey of Manufactures, plants above a certain size are sampled with certainty every year. Below the size threshold, plants are sampled with probability roughly proportional to size in a 5-year rotating panel. Due to Census Bureau disclosure avoidance rules, we cannot disclose the 2002 gross output number with 0% trimming.

applying their corrections.[50]

A potentially important difference between the final and captured data is the sample: the captured sample is around half of the final sample (since we cannot calculate productivity for plants with missing values). Intuitively, the effect of imputation is ambiguous. The data may not be missing completely at random (for instance smaller firms may have more missing data than larger firms), and the plants with missing data may have higher or lower dispersion in the unobserved true data. Similarly, the imputations themselves have an ambiguous effect relative to the true distribution (although the current Census methods likely lead to spuriously low measured dispersion, see White et al. 2018 and subsection V.A).

In Appendix Table 9, we calculate productivity and dispersion measures only for the plants in the final data that are also in the captured sample (reporting the ratio of the value in the final data in the captured sample to the value in the final data full sample, and without trimming). None of the differences are close to explaining the overall final/captured gap, although on average there is slightly more dispersion in just the captured sample (especially for materials). Measured allocative efficiency is around twice as high in the final data when constraining only to the captured sample.

In addition to sampling, another important set of considerations are details about model assumptions. In Appendix Table 10, we show how the change in measured misallocation is a function of underlying data assumptions, in particular the returns to scale and if production is roundabout (creating a knock-on effect of misallocation: improving the allocation of factors increases the materials available for everyone to use, further increasing aggregate productivity, similar to the logic of Hulten 1978). Using reported capital values instead of those in the BLS-Census Multifactor Productivity project on average

[50] As in Bils et al. (2021), we find in that allocative efficiency fell after 2002 in the final data. The decline is much larger in the captured data.

doubles the gap between the captured and final values. Allowing for heterogeneous demand elasticities (following Ahmad and Riker 2019) lowers relative allocative efficiency in the captured data. Using Demirer (2020) production function elasticities raises the relative gains in the captured data a little, and using Blackwood et al. (2021)'s elasticities has a similar effect.[51] Raising the returns to scale raises (relative) allocative efficiency in the captured data, roundabout production dramatically lowers it.

IV.G Quantifying the effects of different types of edits and imputations in the U.S. data

The results of Table 6 are unsatisfying - cross-country comparisons of measured misallocation in datasets which have been cleaned differently potentially may be driven by the quantitatively important data processing. However, while comparing raw data solves the latter problem, it does so at the expense of introducing new errors, since there are transparently incorrect responses in the raw data (such as what gets changed by the divide-by-1000 edit). The natural solution is to compare datasets which have been commonly cleaned (Romer, 1986a,b, 1988, 1989).

There are some relatively common edits done by the Census Bureau (shown in Table 1) that cannot even conceptually be replicated in India. For example, the U.S. Census Bureau makes use of payroll tax data, but India's Ministry of Statistics does not have access to comparable information. We quantify the extent to measure the contribution of each category of the edits shown in Table 1, using a Shapley (1953) decomposition. That is to say, we calculate allocative efficiency for every possible combination of the flags, and report the average marginal contribution.[52] We report the values for the gross

[51] Demirer (2020) report estimates at the two digit level. Blackwood et al. (2021) estimate production functions for detailed industries, but only use a subset of the economy. We also show that the sample isn't driving the results using the Blackwood et al. (2021) elasticities.

[52] When we turn on a flag, we change all of an affected plant's values to the final value, not just the one targeted by the particular flag.

output allocative efficiency values in Table 7, keeping the same order as Table 1.[53] As in Appendix Table 9, imputing missing values has the opposite sign effect as the actual change from captured to final.

The most important edit, especially in 2012, is the Analyst Edits, followed by the logical edits.[54] On the whole, around around $^2/_3$ of the total change in measured misallocation is due to changes that are difficult if not impossible to replicate in other contexts (changes due to logical imputes, analyst corrections, administrative record edits, and the non-replicable not-elsewhere classified set).

In Appendix Table 11, we repeat the Shapley exercise for the standard deviation of TFPR and TFPQ, as well as the within-R^2 of the the regression of TFPR on TFPQ with industry fixed effects. The analyst and logical edits tend to continue to be the most important. For these statistics, the imputes for missing values have the same sign as the actual captured to final change.

Since the details of the data cleaning process matter, and are difficult to replicate, in the next section we describe and then implement an algorithm for editing and imputing for raw establishment-level information. We apply this common edit-imputation algorithm to both the U.S. and Indian data and show how it changes measured misallocation in each country. For the U.S., we also show how it changes measured productivity dispersion and other statistics, and how those results change when we include in the editing process administrative records data that is not available in India.

V A Bayesian Approach to Cleaning Plant-level Data

In this section, we focus on the intuition behind a Bayesian edit-imputation algorithm and why we think this approach is particularly useful for large datasets with highly skewed,

[53] We report the share of the total change each edit is responsible for. But for rounding, the columns sum to one.

[54] Those flags (and the divide by 1000 edits) are not very common. The correlation of how common each flag is (from Table 1) and the average magnitude of its Shapley value (from Table 7) is $-.1$

highly heterogeneous business data (like the Census of Manufactures). We also highlight aspects of the Bayesian method that contrast with the Census Bureau's current editing and imputation methods. We provide further details of the Kim et al. (2015) algorithm [for this section, KCKRW] and details of our implementation of it in Appendix C.

For clarity, we begin with a bit of notation. Establishment i reports p characteristics, $\boldsymbol{y}_i = \{y_{i1}, y_{i2} \ldots y_{ip}\}$ (where items could be missing). The corresponding true values are $\boldsymbol{x}_i = \{x_{i1}, x_{i2} \ldots x_{ip}\}$, with an underlying distribution f. s_{ij} indicates if response j for establishment i is incorrect. Given the dataset of reported values $\boldsymbol{Y} = \{\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots \boldsymbol{y}_n\}$ the goal of data cleaning in principle varies depending on the objective of the statistical agency or the researcher. For example, the goal of the statistical agency might be to produce unbiased estimates of industry totals of individual variables, cross-tabulated with geography and establishment size.[55] For studying productivity dispersion, we would like our edited-imputed data to reflect the underlying joint distribution of plant characteristics $\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots \boldsymbol{x}_n\}$. The Bayesian approach of KCKRW is designed to do this while simultaneously insuring that the data satisfy internal consistency and plausibility checks provided by the statistical agency or the researcher, in particular that ratios of a given plant's variables fall within industry-year specific bounds and that certain sets of variables that should add up do add up.

The ratio bounds and adding-up constraints (a.k.a., balance edits) determine a feasible region for the data. If any of a plant's reported data falls outside the feasible region, then that plant's record is considered in error. Once a record is determined to be in error, the next step is to determine which variable or variables in that record need to be corrected ("error localization"). In this step, the KCKRW algorithm takes advantage of information that the Census Bureau's error localization method does not. By simultaneously using the

[55] However, see Cunningham et al. (2018) for a description of an joint BLS-Census experimental data product which publishes industry level measures of productivity dispersion.

joint distribution of the edit-passing "good" data and the edit constraints a faulty record is failing, the KCKRW algorithm chooses probabilisticly (a) which variables in the faulty record should be replaced with imputations and (b) what those imputations should be.

For example, suppose plant i reports $\{y_{i1}, y_{i2}, y_{i3}, y_{i4}\}$ where $\frac{y_{i1}}{y_{i2}}$ fail the ratios bounds. Depending on the joint distribution of $\boldsymbol{y}_i$, the KCKRW algorithm might replace both y_{i1} and y_{i2} with values that are highly likely given the reported values for y_{i3} and y_{i4}. The Fellegi and Holt (1976) method that the Census Bureau (and many other statistical agencies) uses seeks to minimize the number of edits, and ignores the joint distribution of the variables. In this example, if the Census could satisfy all the edit constraints by replacing only y_{i1} or only y_{i2} it would do so, without estimating the distributions of $\hat{f}\{x_{i1}|x_{i2}, x_{i3}, x_{i4}\}$ or$\hat{f}\{x_{i2}|x_{i1}, x_{i3}, x_{i4}\}$. One reason for wanting to minimize the number of changes to the data rather than preserve the joint distribution may be philosophical: the statistical agency may wish to keep as much of the originally reported data as possible.[56]

Once a decision has been made about which variables in a record are in error, the final step in the Census edit-imputation algorithm is to replace the variables in error with imputations.[57] In this step the Census Bureau uses a variety of methods, depending on what data is available for use in imputations. For example, for payroll, the Bureau has access to alternative data from IRS payroll tax records. For materials, no such administrative records data are available, so the Bureau most frequently uses the predicted value from a regression of materials on shipments to predict missing materials. The Census method is hierarchical, in the sense that it first tries the "best" method in its arsenal, and if that imputed value does not satisfy all the edit constraints, it moves on to the next method. One advantage of the Census method is that it is guaranteed to eventually produce im-

[56] Another reason may be that when the Fellegi and Holt (1976) method was developed, computational power was a small fraction of what it is today.

[57] Note that while these are separate steps in the Census algorithm, in the Kim et al. (2015) algorithm, error localization and imputation are done simultaneously.

putations for every plant.

In contrast to the Census Bureau's approach, the KCKRW algorithm uses a statistical model of the joint distribution of the data for a given sample of plants (normally an industry or industry-year).[58] It does this using a Bayesian non-parametric approach. Rather that assuming the data come from any particular distribution, the approach uses a truncated Dirichlet process mixture of normals (Ishwaran and James, 2001). The algorithm takes a finite number of draws of the Dirichlet process, where each realization is itself a normal distribution with a particular mean and variance. Each of these realizations is a *component* distribution of the overall mixture distribution for the given sample. The number of components is determined probabilistically. This feature of the algorithm allows it to flexibly model many different types of distributions with little input from the modeller. For example, for a sample with a small number of observations that appear to come from a gaussian distribution, the algorithm might choose a single component distribution. For a sample with a large number of observations that appear to come from a highly skewed multi-modal distribution, the algorithm will draw multiple component distributions, each with different mean vectors and different variance-covariance matrices to fit the data.[59]

In our baseline imputation models, we use four of the variables that are subject to ratio edits in the CMF (employment, total wages, cost of materials, and total value of shipments), given our experience that most manufacturing data sets around the world normally at least record those variables.[60] We then use the actual Census ratio bounds

[58] One potential downside of the KCKRW approach relative to the current Census Bureau approach is that the KCKRW algorithm requires that at least two variables used in the imputation model are observed for every plant. (This does not have to be the same two variables for every plant.) For our augmented models described in subsection V.B this does not create much of a selection bias, since the Census Bureau has employment and payroll from administrative records data for almost every plant in the CMF.

[59] There are of course still other researcher choices in the implementation described in Appendix C: for instance we build separate models for each industry-year, where an industry is defined at the 6-digit NAICS level.

[60] One question is what to do with capital. For the US, The Census Bureau edits and imputes the capital

for those four variables.[61] Since we are not including in our model any of the component variables involved in balance edits (e.g., production workers wages and non-production worker wages), we do not use the Census Bureau's balance edits in our implementation.

V.A Implementation of the Bayesian Approach

For each 6-digit NAICS industry-year, we run a single chain of Markov Chain Monte Carlo with a burn-in of 2000 iterations. In the first step, we estimate a model of the joint distribution of the edit-rule-passing data. The second step takes draws from the model to fill in the missing data and edit-failing data. We then re-estimate the model parameters on the combined dataset for each iteration. Note that imputations will differ across iterations both because the draws for missing data are different and because the draws s_{ij} for which ratio-edit-rule-failing variables are in error may be different across iterations. For inference, after burn-in, for each industry-year we continue the chain another 50,000 iterations, and sample 100 different implicates (i.e., completed datasets), keeping every 500th iteration.

In Table 8, we show the relationship between the actual final data and the first implicate of the Kim et al. (2015) implementation (called the "Bayesian Data" in the tables). Even though we handicapped the Bayesian models by not using the logical, analyst, or administrative edits, the ratio of measured allocative efficiency in the final vs Bayesian data is fairly close. For the gross-output specification, the average final value across all three trimmings is .9 of its Kim et al. (2015) counterpart, and the ratio is 1.1 for value added (within the ASM sample the gap is larger, the average ratio is around .7). Without trimming, Kim et al. (2015) measured allocative efficiency is much larger than the final

variables in a separate module from all the other variables we use, and does not use those variables to edit or impute for capital. In addition, in 2007, the Census Bureau decided that the majority of respondents had misinterpreted one of the capital asset questions, and decided to make a mass correction to the capital microdata. To avoid dealing with these issues, we use the final values and leave it out of of the model.

[61] The relevant ratios that are explicitly bounded (above and below) by Census are $\frac{shipments}{materials}$, $\frac{employment}{wages}$ and $\frac{shipments}{wages}$.

value - for instance, it is about twice as big for gross output. The gap shrinks with trimming, as the Kim et al. (2015) data is consistently less affected by trimming outliers than the final data.

In Appendix Tables 12-19 (and Appendix Figures 2-4), we reproduce the main outcomes of the paper, comparing the Bayesian-edited data to the captured and final data. The values are broadly in line with the changes to measured productivity shown in Table 8. Three features to highlight: Appendix Table 12 and 13 show that the Kim et al. (2015) approach suggests changing fewer values than the U.S. Census Bureau,[62] Appendix Table 20 Panel A shows that in the Bayesian data the Olley and Pakes (1996) "within" share is close to one, as in the final data, and Appendix Figures 2 and 3 show similar patterns to the final data for how the changes vary by age and employment.

By running the chain for our main edited-imputed data for 52,000 iterations and sampling 100 implicates of the U.S. data, we can show the extent of uncertainty driven by the simulation variance (Rubin, 1978; Schafer and Graham, 2002; Allison, 2009; Van Buuren, 2018). Following Rubin (2004), in Appendix Table 21 we report the coefficient of variation of the estimates across the different implicates.[63] The values are mostly small (less than one percent), which implies that the model fit is relatively tight. However, this is not always the case: the estimates for the Olley and Pakes (1996) "within" share have large (and increasing) uncertainty. The ability to show this type of result cannot be done using current Census methods (since, for instance, analysts do not provide confidence intervals for their edits).

[62] Materials is the same in the Bayesian and captured data 94 percent of the time, and for the other variables the percentage unchanged is even higher.

[63] Formally, Rubin (2004) derives two additional terms, one of which is driven by sampling uncertainty (less of a concern in the Census) and the second of which comes about because of the simulation variance in the standard deviation of estimates across the different implicates (because the mean is estimated). This last term is proportional to the (reported) standard deviation, and decreasing in the number of implicates, so we do not adjust for it. With 100 implicates, the "correct" estimates for the variance in the data are 1 percent larger than the values implied by Appendix Table 21.

Appendix Table 22 provides a final comparison of the captured, final, and Bayesian-edited data, where we think that modern editing and imputation methods are unambiguously a promising step forward: as a replacement for regression imputations for materials and shipments (recall that Table 2 shows that around a third of plants have at least one value in the final data which is generated using a regression).[64] We show the ratio of the standard deviation of each variable for which the regression flag equals 1 to the plants for whom the regression flag equals zero (so a value smaller than one implies less dispersion in the regression-imputed data). In Panel A, we show the values for the edited plants: where captured data exists, but is changed in the final data (and which sometimes is changed by the Bayesian algorithm). Panel B shows the values for the imputes, where the captured data is missing. For the final data, for both types of plants, there is around three-quarters as much variance in materials for the regression-imputed plants as for the rest (and around 90 percent as much for shipments). The Bayesian-imputed data has much more similar variance in the two datasets. The under-imputing of variance for the regression-imputed final data is intuitive: values are literally imputed on to a regression line, leaving little scope for residual uncertainty.

V.B Implementation with Additional Information

While our primary focus is constraining the information in the U.S. data to be the same as what we can use in the Indian data, it is also valuable to take advantage of the administrative data and Census Bureau analyst corrections available in the U.S.: we still want to edit (and impute) values, but we know more than what is just reported to the Census Bureau by the plants themselves. We do this in a few different ways, progressively adding in more information from administrative records. Administrative tax records are collected for tax IDs (EINs), and there are many multi-unit EINs in U.S. manufacturing sector, so

[64] White et al. (2018) show that within-industry TFPR dispersion is significantly smaller in the regression-imputed final data (which is a subset of all the imputed data) than in the non-imputed final data.

some care is required when using the administrative records.

Payroll is straightforward to include. Annual payroll and March 12th employment are available from IRS payroll tax records, and the CMF questionnaire uses the same definitions of these variables that IRS uses. The Census Bureau's Business Register (BR) processing allocates the EIN-level payroll and employment to the plants associated with those EINs, using allocations for the same EINs in prior years of Census of Manufactures or ASM data. For imputation we use annual payroll and March 12th employment from the Census Bureau's Longitudinal Business Database (LBD), which pulls data from the BR.

IRS Revenue/receipts are also available in the BR. Revenue on the income tax form is not always the identical concept as total value of shipments in the CMF, but it is still potentially useful for imputation. IRS tax data is available only at the EIN level, which can cover multiple plants. Building on work described in Haltiwanger et al. (2019) we use IRS income tax data in the BR to construct a plant-level revenue measure. We describe the construction of this plant-level revenue variable in more detail in Appendix A. When we include the administrative records, we use the same ratio bounds as in the regular data, and include the ratio edit that the "truth" is that the administrative and reported values are within 10% of each other.[65]

Our first supplement to the CMF data takes seriously the issue the mapping from EIN to plant-level information is more reliable for EINs only associated with a single unit. For the "single-unit administrative records" version of the data, we add in administrative records only for the single unit plants.

We then add in administrative records for the entire manufacturing sector. The "big administrative records" version of the data has more observations than the regular edited-

[65] The value of payroll in the CMF final data is only edited with administrative records if it differs from IRS payroll by more than 10%; we follow the same type of logic.

imputed data: there are plants who have too much missing data for us to credibly impute any values for the rest, but the additional administrative records make imputation possible. The trade-off is that we have to impute EIN-level revenue data to plants, which we do assuming that labor productivity is constant within firms (Kehrig and Vincent, 2020).

In our third supplement, we include additional information from the Census final data creation: the analyst edits. This version of the data supplements the "big administrative records" version of the data with the analyst edited values when possible. We call this the "hybrid" data.

In Appendix Figure 4, we augment Figure 1, showing the densities of TFPQ and TFPR for the baseline Bayesian edited and the "big administrative records" data as well as captured and final.[66] While there is substantially less dispersion in the Bayesian-edited datasets than in the captured data, the mode is smaller than in the final data (although the mode is higher in the Bayesian-edited data with administrative records than the one without).

In Appendix 23, we show the main values in the paper for the three alternative datasets (comparing to the baseline Kim et al. 2015-imputed model). There is less dispersion of the plant characteristics, and higher allocative efficiency, in the "single-unit administrative records" and "big administrative records" datasets. However, adding in the analyst edits in the "hybrid" data dominates, and lowers the measured allocation of factors relative to the baseline model.

V.C Validation and Comparison to the Final data

While for the most part we do not have a measure of the "ground truth," we do have one source of validation for both the Census-final and our own edited-imputed data: we can calculate output and output-per-worker from administrative records for the single-

[66] We also calculated distributions for the other two versions of the Bayesian edits with administrative records data. The results are available in our project folder on the Census server.

plant firms ("single-units"), for whom tax data is collected at the plant level. Since output is not generically collected at the plant level for all of manufacturing, it is not used for administrative record imputation in the CMF or ASM.[67]

In Appendix Table 24, we compare labor productivity across the different data versions described above: captured, final, baseline Bayesian, single-unit administrative records, big administrative records, and hybrid. Panel A shows the results for the main datasets (the first three). The captured data is fairly different than mean shipments and labor productivity, especially in 2012. The final and baseline Bayesian broadly match the administrative records, with the final data slightly closer. The final data on average are below the standard deviation and skew in the administrative records, the baseline Bayesian data is above. We then show the within R^2 of a regression of shipments on employment, with 6-digit-NAICS fixed effects. The final and baseline Bayesian data both slightly overstate the predictability. Finally, we show the within R^2 of a regression of the administrative values on their corresponding values in the survey data. Again, the final and baseline Bayesian data perform similarly, and neither are extremely close to the administrative data, especially for labor productivity.[68]

V.D Application to the Indian context

We apply the same Kim et al. (2015) algorithim to the Indian Annual Survey of Industries (ASI) in order to understand how the data cleaning procedure affects measured misallocation in a context outside the U.S. We are not able to replicate the exact method, since we do not have equivalent ratio bounds for the Indian data. We discuss data processing in Appendix Subsection A.I. The procedure changes more values in the ASI than it did in

[67] Annual payroll from administrative records are used by the U.S. Census Bureau for other purposes (Haltiwanger et al., 2016; Decker et al., 2020).

[68] Panel B shows the effects for the augmented Bayesian models. Since all of the models use the administrative data, it is not surprising that the R^2s in the last two rows of Panel B are higher than the ones in Panel A. However, it is worth noting that the datasets are not perfectly correlated with the administrative records. This is possible to improve in the model, by increasing a "reliability weight" for administrative data. This approach is a topic of future research.

the CMF - around 10 percent of each variable is edited, with the most changes for capital. The results for the change in measured misallocation for 2002 and 2010 are presented in Appendix Table 26. There are two major differences relative to to the captured/Bayesian comparison in Appendix Table 19. First, the effect of the editing process is larger by around a factor of 10. Second, while trimming lowers the relative change in the U.S. data, it is not as important in the Indian data.[69]

VI Discussion

In this paper, we use previously unexplored versions of the United States Census of Manufactures for 2002, 2007, and 2012 in order to investigate the role that measurement plays for estimating misallocation. We have two complementary goals. The first is to quantify the importance of data cleaning done by the Census Bureau. Four fifths plants have differences between their original responses and the final data, and over half for at least one difference besides capital. Even trimming generously, in the captured data measured TFP in the United States manufacturing sector is less than an third of what it is in the final data normally used by researchers. Measured allocative efficiency in the untrimmed captured data averages 0.1 percent of the value in the untrimmed final data. We also see large differences for other measures of dispersion, such as the interquartile range, and the share of weighted-average TFPR driven by the unweighted average (Olley and Pakes, 1996). While editing matters more for younger and smaller firms, Census Bureau editing changes measured TFPR on average by 50 percent for even the oldest firms and around a third for the largest.

Many of the important edits undertaken in the U.S. are infeasible for researchers using other datasets (especially from developing countries without access to administrative tax records), because they either use multiple responses for the same information or because

[69] These results contrast with those of Bils et al. (2021), who find that their measurement error correction matters more in the U.S. than in India. However, unlike the results we report, they use the same method in both contexts.

they rely on U.S. Census Bureau industry experts. We describe a new method which can be similarly applied across contexts, and show how it dramatically lowers measured misallocation in both the U.S. and India (relative to the original data). One important limitation of our approach highlighting the effect of data processing done by the U.S. Census Bureau is that it does not provide any insight into why there appears to be so much measurement error in the United States (nor do we have reduced form evidence that there is more or less measurement error in the U.S. relative to other countries). The method we propose is also sensitive to implementation details: adding administrative records as an input to our data-driven algorithmic approach raises measured allocative efficiency, but it is is lowered if we additionally include the changes made by analysts at the U.S. Census Bureau.

There is a large scope for different measurement choices to affect the estimation of misallocation and dispersion in manufacturing, and the results are sensitive to those choices: the differences in across-dataset measured misallocation within the US are orders of magnitude larger than the cross-country differences that are typically found Bils et al. (2021).

Neither ad-hoc approaches nor blissful ignorance are necessary. In this project, and in ongoing work, we show the value in working with experts in data cleaning (who currently tend to associate with other disciplines). To that end, we suggest an alternative approach for cleaning establishment-level data using a hierarchical Bayesian approach. Part of the success of the "farm to table" movement was encouraging diners to understand how food gets from farms to dinner tables. Researchers should similarly understand how the data statistical agencies collect from manufacturing plants is processed before it reaches researchers' table(s and figures).

References

Abowd, J. M. and M. H. Stinson (2013). Estimating measurement error in annual job earnings: A comparison of survey and administrative data. *Review of Economics and Statistics 95*(5), 1451–1467.

Ahmad, S. and D. A. Riker (2019). A method for estimating the elasticity of substitution and import sensitivity by industry. *U.S. International Trade Commission Working Paper 05*, 1–14.

Allcott, H., A. Collard-Wexler, and S. D. O'Connell (2016). How do electricity shortages affect industry? evidence from india. *The American Economic Review 106*(3), 587–624.

Allison, P. D. (2009). Chapter 4: Missing data. *The SAGE Handbook of Quantitative Methods in Psychology*, 72–89.

Almunia, M., J. Hjort, J. Knebelmann, and L. Tian (2021). Strategic or confused firms? evidence from âmissingâ transactions in uganda. *Working Paper*.

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

Asker, J., A. Collard-Wexler, and J. De Loecker (2019, April). (mis)allocation, market power, and global oil extraction. *American Economic Review 109*(4), 1568–1615.

Autor, D., D. Dorn, L. F. Katz, C. Patterson, and J. Van Reenen (2020, 02). The Fall of the Labor Share and the Rise of Superstar Firms. *The Quarterly Journal of Economics 135*(2), 645–709.

Baily, M. N. and C. Hulten (1992). Productivity dynamics in manufacturing plants. *Brookings Papers on Economic Activity 1992*, 187–267.

Balke, N. S. and R. J. Gordon (1989). The estimation of prewar gross national product: Methodology and new evidence. *Journal of Political Economy 97*(1), 38–92.

Banerjee, A. V. and E. Duflo (2005). Growth Theory through the Lens of Development Economics. *Handbook of Development Economics 1*(05), 473–552.

Barrios, J. M. and J. Gallemore (2021). Tax planning knowledge diffusion via the labor market. *University of Chicago, Becker Friedman Institute for Economics Working Paper* (2019-64).

Bartelsman, E., M. Hälbig, and F. di Mauro (2021). Methodological report on cross-country analysis of newly developed firm-level indicators. *Working Paper*.

Bartelsman, E. J. and Z. Wolf (2018). Measuring productivity dispersion. In *Oxford Handbook of Productivity Analysis*. Oxford University Press.

Bils, M., P. J. Klenow, and C. Ruane (2021). Misallocation or mismeasurement? Technical report, Working Paper.

Blackwood, G. J., L. S. Foster, C. A. Grim, J. Haltiwanger, and Z. Wolf (2021). Macro and micro dynamics of productivity: From devilish details to insights. *American Economic Journal: Macroeconomics 13*(3), 142–72.

Boar, C. and V. Midrigan (2021). Markups and inequality. *Working Paper*.

Burns, A. F. (1960). Progress towards economic stability. *The American Economic Review 50*(1), 1–19.

Busso, M., L. Madrigal, and C. Pages (2013). Productivity and Resource Misallocation in Latin America. *B.E. Journal of Macroeconomics 13*(1), 903–932.

Chow, M. C., T. C. Fort, C. Goetz, N. Goldschlag, J. Lawrence, E. R. Perlman, M. Stinson, and T. K. White (2021). Redesigning the longitudinal business database. *Working Paper*.

Cirera, X., R. Fattal-Jaef, and H. Maemir (2020). Taxing the good? distortions, misallocation, and productivity in sub-saharan africa. *The World Bank Economic Review 34*(1), 75–100.

Collard-Wexler, A. (2011). Productivity dispersion and plant selection in the ready-mix concrete industry. *US Census Bureau Center for Economic Studies Paper No. CES-WP-11-25*.

Collard-Wexler, A. and J. De Loecker (2016). Production function estimation with measurement error in inputs. Technical report, Working Paper.

Cunningham, C., L. Foster, C. Grim, J. Haltiwanger, S. W. Pabilonia, J. Stewart, and Z. Wolf (2018, April). Dispersion in Dispersion: Measuring Establishment-Level Differences in Productivity. Working Papers 18-25, Center for Economic Studies, U.S. Census Bureau.

Davis, S. J. and J. Haltiwanger (1991). Wage dispersion between and within us manufacturing plants, 1963-86. *Brookings Papers on Economic Activity*.

Decker, R. A., J. Haltiwanger, R. S. Jarmin, and J. Miranda (2020). Changing business dynamism and productivity: Shocks versus responsiveness. *American Economic Review 110*(12), 3952–90.

Demirer, M. (2020). Production function estimation with factor-augmenting technology: An application to markups. Technical report, Working Paper.

Esfahani, S. (2019). Agricultural misallocation: Mismeasurement, misspecification, or market frictions. *Working Paper*.

Fellegi, I. P. and D. Holt (1976). A systematic approach to automatic edit and imputation. *Journal of the American Statistical association 71*(353), 17–35.

Fort, T. C. and S. D. Klimek (2016). The effects of industry classification changes on us employment composition. *Working Paper*.

Foster, L., C. Grim, and J. Haltiwanger (2016). Reallocation in the great recession: cleansing or not? *Journal of Labor Economics 34*(S1), S293–S331.

Foster, L., J. Haltiwanger, and C. Syverson (2008). Reallocation, firm turnover, and efficiency: Selection on productivity or profitability? *American Economic Review 98*(1), 394–425.

Foster, L., J. C. Haltiwanger, and C. J. Krizan (2001, January). *Aggregate Productivity Growth: Lessons from Microeconomic Evidence*, pp. 303–372. University of Chicago Press.

Foster, L. S., C. A. Grim, J. Haltiwanger, and Z. Wolf (2017). Macro and micro dynamics of productivity: From devilish details to insights. *Working Paper*.

Frisch, R. (1934). *Statistical confluence analysis by means of complete regression systems*, Volume 5. Universitetets Økonomiske Instituut.

Gauthier, J. (2011). History of the 2007 economic census. *EC07-00R-HIST*.

Gollin, D. and C. Udry (2019). Heterogeneity, measurement error and misallocation: Evidence from african agriculture. *NBER Working Paper* (w25440).

Griliches, Z. (1974). Errors in variables and other unobservables. *Econometrica: Journal of the Econometric Society*, 971–998.

Haltiwanger, J., R. Jarmin, R. Kulick, and J. Miranda (2017). High growth young firms: Contribution to job, output, and productuvity growth. *CARRA Working Paper Series* (2017-03).

Haltiwanger, J., R. Jarmin, R. Kulick, J. Miranda, V. Penciakova, and C. Tello-Trillo (2019). Firm-level revenue dataset. *CES Technical Note Series* (2019-02).

Haltiwanger, J., R. S. Jarmin, R. Kulick, and J. Miranda (2016). *High Growth Young Firms: Contribution to Job, Output, and Productivity Growth*, pp. 11–62. University of Chicago Press.

Haltiwanger, J., R. B. Kulick, and C. Syverson (2018). Misallocation measures: The distortion that ate the residual. *University of Chicago, Becker Friedman Institute for Economics Working Paper* (2019-35).

Hicks, J. (1981). *Collected Essays on Economic Theory: Wealth and welfare*. Cambridge, Mass: Harvard University Press.

Holmes, T. J. and J. J. Stevens (2014). An alternative theory of the plant size distribution, with geography and intra-and international trade. *Journal of Political Economy 122*(2), 369–421.

Hopenhayn, H. A. (2014). On the Measure of Distortions. *Working Paper*.

Hsieh, C.-T. and P. J. Klenow (2009). Misallocation and Manufacturing Tfp in China and India. *Quarterly Journal of Economics 124*(4), 1–55.

Hulten, C. R. (1978). Growth accounting with intermediate inputs. *The Review of Economic Studies 45*(3), 511–518.

Hulten, C. R. (1991). The measurement of capital. In *Fifty years of economic measurement: The jubilee of the Conference on Research in Income and Wealth*, pp. 119–158. University of Chicago Press.

Ishwaran, H. and L. F. James (2001). Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association 96*(453), 161–173.

Jerrim, J., L. A. Lopez-Agudo, O. D. Marcenaro-Gutierrez, and N. Shure (2017). What happens when econometrics and psychometrics collide? an example using the pisa data. *Economics of Education Review 61*, 51–58.

Kambourov, G. and I. Manovskii (2013). A cautionary note on using (march) current population survey and panel study of income dynamics data to study worker mobility. *Macroeconomic Dynamics 17*(1), 172–194.

Kehrig, M. (2015). The cyclical nature of the productivity distribution. *US Census Bureau Center for Economic Studies Paper No. CES-WP-11-15*.

Kehrig, M. and N. Vincent (2020). Good dispersion, bad dispersion. *Working Paper*.

Kim, H. J., L. H. Cox, A. F. Karr, J. P. Reiter, and Q. Wang (2015). Simultaneous edit-imputation for continuous microdata. *Journal of the American Statistical Association 110*(511), 987–999.

Kim, H. J., J. P. Reiter, Q. Wang, L. H. Cox, and A. F. Karr (2014). Multiple imputation of missing or faulty values under linear constraints. *Journal of Business & Economic Statistics 32*(3), 375–386.

Kroencke, T. A. (2017). Asset pricing without garbage. *The Journal of Finance 72*(1), 47–98.

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

Martin, L. A., S. Nataraj, and A. E. Harrison (2017). In with the big, out with the small: Removing small-scale reservations in india. *American Economic Review 107*(2), 354–86.

Medalia, C., B. Meyer, V. Mooers, and D. Wu (2019). The use and misuse of income data and extreme poverty in the united states. *University of Chicago, Becker Friedman Institute for Economics Working Paper*.

Melitz, M. J. (2003). The impact of trade on intra-industry reallocations and aggregate industry productivity. *Journal of Political EcoEconometricanomy 71*(6), 1695–1725.

Meyer, B. D. and N. Mittag (2019). Using linked survey and administrative data to better measure income: implications for poverty, program effectiveness, and holes in the safety net. *American Economic Journal: Applied Economics 11*(2), 176–204.

Nishida, M., A. Petrin, M. Rotemberg, and T. K. White (2015). Are We Undercounting Reallocation's Contribution to Growth? *Working Paper*.

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

Redding, S. J. and D. E. Weinstein (2020). Measuring aggregate price indices with taste shocks: Theory and evidence for ces preferences. *The Quarterly Journal of Economics 135*(1), 503–560.

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

Roberts, M. J. and D. Supina (1996). Output price, markups, and producer size. *European Economic Review 40*(3-5), 909–921.

Romer, C. (1986a). Spurious volatility in historical unemployment data. *Journal of Political Economy 94*(1), 1–37.

Romer, C. D. (1986b). Is the stabilization of the postwar economy a figment of the data? *The American Economic Review 76*(3), 314–334.

Romer, C. D. (1988). World war i and the postwar depression a reinterpretation based on alternative estimates of gnp. *Journal of monetary Economics 22*(1), 91–115.

Romer, C. D. (1989). The prewar business cycle reconsidered: New estimates of gross national product, 1869-1908. *Journal of Political Economy 97*(1), 1–37.

Rotemberg, M. (2019). Equilibrium effects of firm subsidies. *American Economic Review 109*(10), 3475–3513.

Rubin, D. B. (1978). Multiple imputations in sample surveys-a phenomenological bayesian approach to nonresponse. In *Proceedings of the survey research methods section of the American Statistical Association*, Volume 1, pp. 20–34. American Statistical Association.

Rubin, D. B. (2004). *Multiple imputation for nonresponse in surveys*, Volume 81. John Wiley & Sons.

Savov, A. (2011). Asset pricing with garbage. *The Journal of Finance 66*(1), 177–201.

Schafer, J. L. and J. W. Graham (2002). Missing data: our view of the state of the art. *Psychological methods*.

Shapley, L. S. (1953). A value for n-person games. *Contributions to the Theory of Games 2*(28), 307–317.

Shorrocks, A. F. (2013). Decomposition procedures for distributional analysis: a unified framework based on the shapley value. *Journal of Economic Inequality*, 1–28.

Sigman, R. S. (1997, October). Development of a "plain vanilla" system for editing economic census data. Number 24 in Conference of European Statisticians Working Paper.

Sincavage, J. R., C. Haub, and O. Sharma (2010). Labor costs in india's organized manufacturing sector. *Monthly Lab. Rev. 133*, 3.

Syverson, C. (2004a). Market structure and productivity: A concrete example. *Journal of Political Economy 112*(6), 1181–1222.

Syverson, C. (2004b). Product substitutability and productivity dispersion. *Review of Economics and Statistics 86*(2), 534–550.

Thompson, K. J., J. T. Fagan, B. L. Yarbrough, and D. L. Hambric (2004). Using a quadratic programming approach to solve simultaneous ratio and balance edit problems. In *Proceedings of the Survey Research Methods Section, American Statistical Association*, pp. 4485–4490.

Thompson, K. J. and R. S. Sigman (1999). Statistical methods for developing ratio edit tolerances for economic data. *Journal of Official Statistics 15*(4), 517.

Troccoli, C. (2020). Comment on "paid parental leave and children's schooling outcomes". *Working Paper*.

Van Buuren, S. (2018). *Flexible imputation of missing data*. CRC press.

Velayudhan, T. (2018). Misallocation or misreporting? evidence from a value added tax notch in india. *Working Paper*.

Vom Lehn, C., C. Ellsworth, and Z. Kroff (2020). Reconciling occupational mobility in the current population survey. *IZA Discussion Paper No. 13509*.

White, T. K. (2014). Recovering The Item-Level Edit And Imputation Flags In The 1977-1997 Censuses Of Manufactures. Working Papers 14-37, Center for Economic Studies, U.S. Census Bureau.

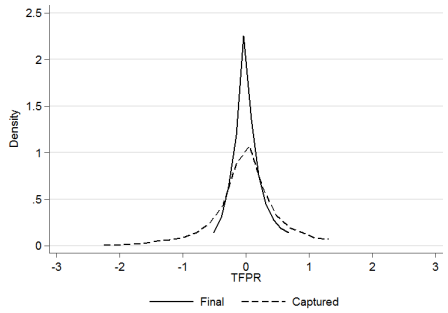
White, T. K., J. P. Reiter, and A. Petrin (2015). Imputation in U.S. Manufacturing Data and Its Implications for Productivity Dispersion. Research conference paper, Federal Conference on Statistical Methodology.

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

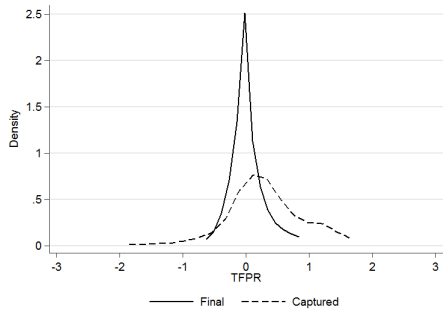
Zwick, E. (2021). The costs of corporate tax complexity. *American Economic Journal: Economic Policy 13*(2), 467–500.

Figure 1: Productivity Density, Captured vs Final

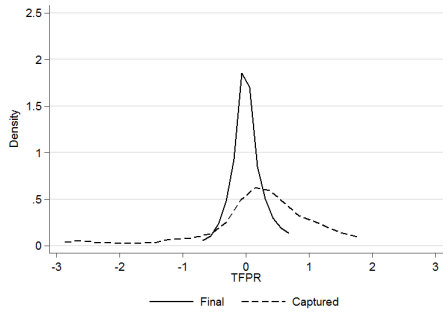
(a) **2002**



(b) **2007**



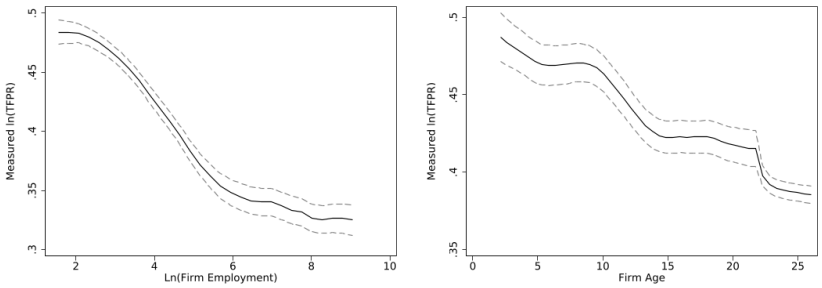
(c) **2012**



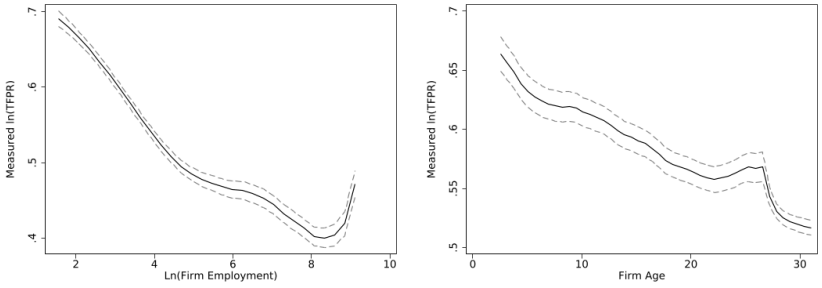
This figure plots the kernel densities for (scaled) TFPR in the captured and final data. For comparability across graphs, and to comply with Census disclosure rules, we trim one percent of the left tail and five percent of the right tail

Figure 2: Absolute Difference between Captured and Final Productivity, by Firm Employment and Age

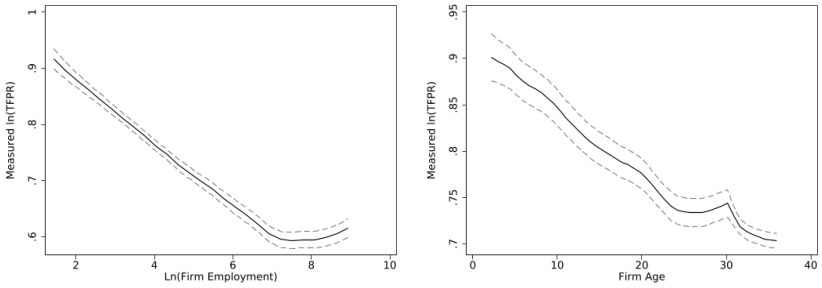
(a) **2002**



(b) **2007**



(c) **2012**



This figure plots a local polynomial regression, predicting the (absolute) difference between ln(TFPR) in the final vs captured data, over firm age and (ln) firm employment, as described in the text. To comply with Census disclosure rules, the 5 percent tails of each graph have been trimmed.

Table 1. Categorization of Changes Made to the U.S. Census of Manufacturers

Edit/Impute Label	Share of Plants	Underlying Flags
(1)	(2)	(3)
Impute Missing Value	0.485	Missing Captured Response (other than Capital)
Logical Edit for Payroll	0.089	Payroll in (RL)
Analyst Edits	0.055	Any variable in (RC)
Logical Edit for Shipments	0.04	Shipments in (RL)
Logical Edit for Materials	0.037	Materials in (RL)
Regression Edit for Materials	0.026	Materials in (RB) or (RW)
Edit from Administrative Records	0.018	Any variable in (RA)
Divide by 1000	0.005	Any variable in (RN)
Other Capital Edits	0.203	Changes to Capital (other than those described above)
Other Replicable Change	0.018	Shipments in {(RB), (RE), (RG), (RH), (RJ), (RM), (RS), (RV), (RX)} or Materials in {(RE), (RH), (RM)} or Payroll in {R(H), (RHQ), (RJ), (RG), (R2), (R4)}
Any Change, Not Elsewhere Classified	0.021	The remainder

Notes: Flags are defined in Appendix Table 1. Categorization done by authors. The "Share of Plants" is the share of plants who have been affected by the corresponding label, averaged over 2002, 2007, and 2012; the year-specific shares are reported in Appendix Table 2. The value for "Any Change, Not Elsewhere Classified" could not be disclosed in 2007, we impute 0 for calculating the average. The sample is plants for whom it is possible to calculate productivity in the final data.

Table 2. Characteristics of Changes Between Final and Captured Data

Type of Change	Share of Plants (2002)	Share of Plants (2007)	Share of Plants (2012)
(1)	(2)	(2)	(3)
Panel A. Overall Changes			
Any Change	0.722	0.894	.824
Any Change (other than capital)	0.62	0.574	.567
Any Missing	0.464	0.443	0.546
Any Regression Flag	0.376	0.349	0.349
Any Edit >10%	0.287	0.32	0.311
Exactly One Change	0.226	0.363	0.31
All Missing	0.228	0.202	0.243
Panel B. Characteristic-Specific Changes			
Capital:			
… Final < Captured	0.031	0.403	0.069
… Unchanged	0.729	0.259	0.499
… Final > Captured	0.24	0.338	0.432
… Edit >10%	0.165	0.208	0.191
Materials			
… Final < Captured	0.04	0.047	0.026
… Unchanged	0.894	0.875	0.869
… Final > Captured	0.066	0.078	0.104
… Edit >10%	0.045	0.059	0.059
Payroll:			
… Final < Captured	0.046	0.033	0.04
… Unchanged	0.797	0.895	0.883
… Final > Captured	0.157	0.073	0.077
… Edit >10%	0.076	0.063	0.088
Shipments			
… Final < Captured	0.014	0.028	0.029
… Unchanged	0.897	0.865	0.899
… Final > Captured	0.089	0.107	0.072
… Edit >10%	0.058	0.07	0.059

Notes: The "Share of Plants" is the share of plants who have been affected by the corresponding label in each year. Any Regression Flag includes both imputed and edited values. The sample for Panel A is plants for whom it is possible to calculate productivity in the final data. The sample for Panel B is plants for whom it is possible to calculate productivity in the captured data.

Table 3. Distribution of the Captured/Final Ratios

	First Percentile (1)	Fifth Percentile (2)	Tenth Percentile (3)	Ninetieth Percentile (4)	Ninety-Fifth Percentile (5)	Ninety-Ninth Percentile (6)
Panel A. Capital						
2002	0.46	0.498	0.523	1	1	1.918
2007	0.006	0.075	0.217	1.023	1.035	1.514
2012	0.004	0.053	0.135	1	1.01	17.43
Panel B. Materials						
2002	0.012	0.889	1	1	1	1.725
2007	0.008	0.524	1	1	1	2.505
2012	0.011	0.461	0.993	1	1	7.046
Panel C. Payroll						
2002	0.255	0.829	0.994	1	1	1.553
2007	0.222	0.951	1	1	1	2.034
2012	0.201	0.809	1	1	1	403.4
Panel D. Shipments						
2002	0.1	0.812	1	1	1	1.024
2007	0.076	0.755	0.988	1	1	1.163
2012	0.09	0.931	1	1	1	1.608

Notes: We calculate the value in the captured data divided by its final counterpart. This Table reports the distribution of those ratios. For disclosure purposes, the reported values are not the cutoff at the exact percentile, but the average value of all of the plants within the centile. The sample is plants for whom it is possible to calculate productivity in the captured data.

Table 4. Dispersion of Plant Characteristics, Captured vs. Final

	Captured Data / Final Data			
	90/10 ratio	75/25 ratio	Standard Deviation	R^2
	(1)	(2)	(3)	(4)
Panel A. Capital				
2002	1.122	1.178	1.112	0.957
2007	1.361	1.448	1.365	0.783
2012	1.439	1.442	1.446	0.595
Panel B. Materials				
2002	1.083	1.123	1.087	0.839
2007	1.244	1.263	1.275	0.793
2012	1.492	1.419	1.493	0.797
Panel C. Payroll				
2002	1.168	1.19	1.184	0.938
2007	1.144	1.161	1.165	0.901
2012	1.21	1.255	1.218	0.853
Panel D. Shipments				
2002	1.064	1.118	1.063	0.925
2007	1.059	1.097	1.091	0.917
2012	1.279	1.262	1.282	0.861

Notes: The first three column of this table reports moments of the distribution of the inputs and shipments. The variables are logged and demeaned within each industry. For each of the statistics in columns 1, 2, and 3, we calculate the relevant value in the captured data divided by its final counterpart, and report the ratio. The underlying values are in Appendix Table 3. Column 4 reports the within R^2 of a regression of the corresponding (logged) variable in the captured data on its value in the final data, with 6-digit NAICS fixed effects. Within each dataset, the samples are balanced (it must be possible to calculate productivity), but the samples in the captured and final data are different.

Table 5. Productivity Dispersion, Captured vs. Final

	Captured Data / Final Data			
	90/10	75/25	Standard Deviation	R^2
	(1)	(2)	(3)	(4)
Panel A. TFPQ				
2002	3.711	5.895	3.776	0.538
2007	2.788	4.179	2.791	0.555
2012	4.141	5.713	4.394	0.527
Panel B. TFPR				
2002	2.14	2.024	2.254	0.334
2007	2.116	2.565	2.192	0.318
2012	3.13	3.082	2.941	0.247

Notes: The first three column of this table reports moments of the distribution of TFPQ and TFPR. TFPR and TFPQ are calculated using sectoral cost shares (and TFPQ uses the model to convert from revenues to quantities). The variables are logged and demeaned within each industry, as described in the text. For each of the statistics in columns 1, 2, and 3, we calculate the relevant value in the captured data divided by its final counterpart, and report the ratio. The underlying values are in Appendix Table 6. Column 4 reports the within R2 of a regression of the corresponding (logged) variable in the captured data on its value in the final data, with 6-digit NAICS fixed effects. Within each dataset, the samples are balanced (it must be possible to calculate productivity), but the samples in the captured and final data are different.

Table 6. Measured Allocative Efficiency, Captured vs. Final

	Captured Data / Final Data		
	No Trimming	1 Percent Trimming	2 Percent Trimming
	(1)	(2)	(3)
Panel A. Gross Output			
2002	0.0004	0.157	0.238
2007	0.001	0.184	0.278
2012	0.00007	0.02	0.037
Panel B. Value Added			
2002	0.004	0.225	0.305
2007	0.006	0.181	0.313
2012	0.0003	0.027	0.032
Panel C. Annual Survey of Manufacturers (Gross Output)			
2002	-	0.283	0.294
2007	0.009	0.38	0.566
2012	0.022	0.251	0.392

Notes: The equations for calculating allocative efficiency are as described in the text. For the trimming, we drop the (upper and lower) extremes for TFPQ and TFPR. The values are balanced within each dataset (it must be possible to calculate productivity), but unbalanced across. For each of the statistics in the table we calculate the relevant value in the captured data divided by its final counterpart, and report the ratio. The underlying values are in Appendix Table 8. Panel A uses a gross output specification as in Bils et al. (2021) and Blackwood et al. (2021). Panel B uses a value added specification as in Hsieh and Klenow (2009). Panel C uses only the ASM plants (and the corresponding weights), the ASM value for the captured data with no trimming could not be disclosed in 2002.

Table 7. Effect of Flags on Measured Misallocation

Edit/Impute Label	Shapley Value (2002)	Shapley Value (2007)	Shapley Value (2012)
(1)	(2)	(2)	(3)
Impute Missing Value	-0.055	-0.054	-.137
Logical Edit for Payroll	0.06	0.022	.027
Analyst Edits	0.154	0.29	0.559
Logical Edit for Shipments	0.186	0.255	0.098
Logical Edit for Materials	0.047	0.049	0.018
Regression Edit for Materials	0.09	0.078	0.096
Edit from Administrative Records	0.008	0.009	0.008
Divide by 1000	0.087	0.1	0.035
Other Capital Edits	0.013	0.063	0.102
Other Replicable Change	0.213	0.189	0.196
Any Change, Not Elsewhere Classified	0.197	-0.001	-0.00002

Notes: Lables are defined in Table 1. We calculate gross output misalloaction (as in Table 6 column 1) with no trimming for every possible combination of flags (if a flag is turned on, we use only final values for the plant, not just for the particular characteristic affected by the flag). We then calculate the Shapley (1953) value for each flag, and report the share of the change from captured to final measured misallocation attributable to each flag (so the columns would sum to one if not for rounding). Negative values imply that the Shapley (1953) value of the flag is the opposite of the actual captured-to-final difference. The sample is plants for whom it is possible to calculate productivity in the final data.

Table 8. Measured Allocative Efficiency, Bayesian vs. Final

	Final Data / Bayesian Data		
	No Trimming	1 Percent Trimming	2 Percent Trimming
	(1)	(2)	(3)
Panel A. Gross Output			
2002	0.456	1.182	1.279
2007	0.315	0.907	0.813
2012	0.892	1.175	1.178
Panel B. Value Added			
2002	0.975	1.264	1.204
2007	0.235	0.997	1.009
2012	1.399	1.652	1.462
Panel C. Annual Survey of Manufacturers (Gross Output)			
2002	0.396	0.857	0.961
2007	0.724	0.689	0.682
2012	0.767	1.061	1.048

Notes: The equations for calculating allocative efficiency are as described in the text. For the trimming, we drop the (upper and lower) extremes for TFPQ and TFPR. The values are balanced within each dataset (it must be possible to calculate productivity), but unbalanced across. For each of the statistics in the table we calculate the relevant value in the final data divided by its Bayesian-edited counterpart, and report the ratio. Panel A uses a gross output specification as in Bils et al. (2021) and Blackwood et al. (2021). Panel B uses a value added specification as in Hsieh and Klenow (2009). Panel C uses only the ASM plants (and the corresponding weights), the ASM value for the captured data with no trimming could not be disclosed in 2002. The Bayesian edited data is constructed following Kim et al. (2015), as described in the text.

Online Appendix

Plant-to-Table(s and Figures): Processed Manufacturing Data and Measured Misallocation

Martin Rotemberg and T. Kirk White

July 2021

A U.S. Census and IRS Income and Payroll Tax Data

The quinquennial **Census of Manufactures (CMF)** covers roughly 300,000 manufacturing plants.[70] Every year, the **Annual Survey of Manufactures (ASM)** collects information on a rotating panel of establishments, with the largest establishments surveyed with certainty, and smaller plants in the frame with some probability (which is increasing with size). The sampling weights are available in the U.S. Censuses of Manufactures,[71] and we show how measures of misallocation change when using the ASM sample instead of the Census.

To measure plant-level TFPQ and TFPR we use four plant-level variables: the total value of shipments, the total cost of materials (which includes costs of energy), total salaries and wages (i.e., annual payroll), and capital. In our imputation models (but not in our measure of plant-level productivity) we also use March 12 employment, because it is highly correlated with the other variables and because high quality administrative records data are available to impute it when it is missing.

To measure capital in the final data, we use the real capital stock constructed as part of the BLS-Census Multifactor Productivity project, as described in Cunningham et al. (2018).[72] We then multiply the real stock values by nominal rental rates (Kehrig, 2015) in order to measure capital flows. While Cunningham et al. (2018) provide separate measures of capital for equipment and structures, we use the plant-level sum of the two measures.

For the captured data, we use the fact that the procedure above provides an implicit rental rate for each plant in the data: the relationship between the cleaned measured flow cost of capital services and the final stock value in the Census data.[73]

[70] Information for plants with fewer than 5 employees - roughly one third of the sample - are almost entirely imputed. We follow standard practice of excluding these so-called administrative records plants (Foster et al., 2016).

[71] Sample are selected at 5-year intervals beginning in years ending with 4 or 9.

[72] An internal Census memo (available to researchers with approved access to the manufacturing data in the FSRDCs) written by Jake Blackwood and Cody Tuttle as part of that project was very helpful for us. See also Chow et al. (2021) for a description of the Longitudinal Business Database.

[73] Specifically, we calculate the stock of capital as the average of total assets at the beginning and end of the year. We then calculate $K_{\text{flow,captured}} = \frac{K_{\text{flow,BLS}}}{K_{\text{stock,final}}} \times K_{\text{stock,captured}}$. In the original version of this paper, we instead multiply the capital stock value in the Census by 10% to impute the flow costs, and we show measured counterfactual productivity gains for that set-up in Appendix Table 10. In India there is no BLS-type of information for capital, so we use a 10 percent interest rate. Because the counterfactual we are interested removes deviations from the industry average, whatever interest rate we pick cancels out and therefore doesn't affect our misallocation measure (as long as the chosen interest rate is constant within each sector).

For some of our analyses we use revenue data from the Census Bureau's Business Register (BR), ultimately from IRS income tax forms. Haltiwanger et al. (2017) use this data to construct a firm-level revenue measure using the revenue data from various IRS income tax forms. The details of the construction of the firm-level revenue variable are described in the data appendix of Haltiwanger et al. (2017) as well as in Haltiwanger et al. (2019). There are two revenue data issues worth mentioning. First, Haltiwanger et al. (2019) note that about 20% of businesses file their payroll and income tax reports under different EINs, and in these cases the Census Bureau has no direct way of linking the two EINs. Second, while the Business Register allocates EIN-level employment and payroll to establishments, the Census Bureau's Business Register does not do this for revenue. In order to create plant-level imputations of the administrative revenue data, we first we modified the underlying code used by Haltiwanger et al. (2019) to aggregate revenue to the EIN level instead of the firm level.[74] We then used the plant-level allocations described in the previous paragraph to allocate EIN-level revenue to each plant, heroically assuming that all plants within the same EIN have the same labor productivity (Kehrig and Vincent, 2020).

A.I Indian Data

For India, we use the Annual Survey of Industries (the ASI). Factories with over 100 workers are surveyed every year, while smaller establishments are surveyed every few years (the ASI is designed to be representative at the State by Industry level, so establishments without local competitors are more likely to be surveyed). Hsieh and Klenow (2009) and Bils et al. (2021) use the same dataset, and we follow standard practice in generating measures of gross output, intermediate inputs, capital, and payroll.[75] We use cost shares by two digit industry to back out production function parameters. Industries are grouped using India's NIC (National Industrial Classification) codes, and we report the value of reallocation for 2002 and 2010, which are the start and end of the dataset we are using. While there is no administrative data we can use for variable validation in India, there is at least some direct evidence of measurement error. In the repeated cross-section, if we observe a plant twice we can calculate both the change in reported age and the time between observations. Over a third of the time, the change in reported plant age is either at least 2 years bigger or smaller than the actual change. Some of this is clearly misunderstanding the question - for instance, in some years reporting age and other years reporting the year of initial production. Dropping the plants with a gap over 1000, the gap is still clearly wrong a fifth of the time.

Since the industry codes do not line up with the U.S.'s (not that we would necessarily think that the same bounds should apply), we cannot use the same feasible region in India as in the US. We define the feasible region by following the resistant fences method, which is the starting point for how Census chooses its ratio bounds (Thompson and Sigman, 1999). Within each industry, we calcuate the log ratio r_{jk} for all of the inputs j and only

[74] We are grateful to Cristina Tello-Trillo for helping us with this.

[75] We use the code from Rotemberg (2019), which essentially copied the code from Allcott et al. (2016), whose appendix goes into the data in substantially more detail, as does the appendix of Bils et al. (2021).

use output for k. We calculate the ratio's 25th and 75th percentiles, Q_{jk}^{25} and Q_{jk}^{75}, and the interquartile range IQR_{jk}. We then flag all ratios that are either smaller than $Q_{jk}^{25} - C \times IQR_{jk}$ or larger than $Q_{jk}^{75} + C \times IQR_{jk}$ where C is a pre-specified threshold (for our application, $C = 3$, which is relatively large). The variables we use in India are the same as the ones in the U.S., adding in capital (since we have no external data source) and the sample weight. We run the estimation separately by 2-digit industry and year, and we constrain the sample to include only the plants with no missing values.

B Effects of Each Type of Edit on Measured Misallocation and Productivity Dispersion

In Table 1 and 7 (as well as Appendix Table 2), we describe the effect of eleven supercategories of the edit flags. We describe the mapping in Table 1, in this section we go into more detail, in the same order as the tables.

The flag impute missing value applies to all variables (besides capital), and we use it whenever the final data is not missing, but the captured data is missing.[76]

We split out logical edits for payroll, shipments, and materials. Logical edits are done when there are multiple survey questions which ask for the same information. For example, the CMF asks respondents for the plant's production worker wages, non-production worker wages, and total salaries and wages (which should be the sum of the first two). To give a simple example, if reported non-production worker wages are, e.g., only half of the reported total salaries and wages, and both are plausible for the given plant in the given industry, but non-production worker wages were not reported, then the Bureau will impute non-production worker wages for this plant as total salaries and wages minus production worker wages.

Analyst corrections rely on the expertise of full-time industry specialists employed by the U.S. Census Bureau. Note that analysts can try to use primary sources (such as calling a plant) to verify or edit values, this is not categorized as an analyst correction (or as an edit at all). In addition to changing values, analysts can also "goldplate" information to ensure that it does not get edited using other methods. We do not consider "goldplating" a type of edit for considering the Shapley values, but we do use them in the "hybrid" data and when comparing the exposure of trimmed and untrimmed data to analyst-verified values.

Regression Edits (which we report for materials) are used to edit data when alternative sources of information are not available. The U.S. Census Bureau uses a variety of industry-specific regression-based imputation strategies. Since they do not require any observed alternative value, regression imputes are often used to impute missing values when no other information is available for a given variable, preventing the use of logical-type imputes. In general, regression edits create predictions using one other variable, and (for plants surveyed in the Annual Survey of Manufactures), one-year lags of the imputed variable. Unlike administrative and logical edits, there is not necessarily any direct

[76] There are some observations with missing capital data but where the corresponding flag does not correspond to impute missing; we overrule those flags for this categorization.

evidence that the value reported by the the establishment may be incorrect. Instead the Census uses the ratio bounds we use in Section V.

Administrative edits are similar to logical edits, but differ in that the alternative source of information is from administrative records. The administrative records come primarily from IRS payroll tax data.

The divide by 1000 edits happen when the ratio of a dollar-valued variable (e.g., annual payroll) over employment is 1000 times greater than what is typical for the plant's industry. In these cases the Census Bureau suspects that the respondent answered in dollar units even though the questionnaire asks for values in thousands of dollars.

We separately split out the effect of capital edits from the other residual categories. This includes all changes to capital that aren't defined by one of the flags above (besides impute missing).

We manually categorize the residual changes into ones that are replicable and ones that are not. For the most part, this is regression-types of edits that we don't categorize in the Regression Edits for Materials (such as regressions for the other variables, or directly imputing the midpoint of the ratio bound).

C Bayesian Simultaneous Edit-Imputation

In this section, we describe how we implement the Kim et al. (2014) approach. First, we define the feasible region $\mathcal{D}$ of plausible reports. In our implementation we define this region using only a set of ratio edit rules which bound the ratios of any two variables.[77] The ratio edit rules can come either from industry specific knowledge, or from outliers in the data itself. Fellegi and Holt (1976) note that the set of explicit ratio edit rules can imply additional ones as well.[78] While $\boldsymbol{s}_i$ is not directly observed, A_i indexes the failed ratio edit rules.If we were using balance edits in our implementation, A_i would also indicate failed balance edit rules. If, e.g., y_{i1} fails multiple edits and y_{i2} fails only one, then, other things equal, y_{i1} is more likely to be faulty than y_{i2}.

For the baseline Bayesian model, we use the relevant ratios that are explicitly bounded (above and below) by the Census Bureau: $\frac{shipments}{materials}$, $\frac{employment}{wages}$ and $\frac{shipments}{wages}$. When we add in the administrative records, we add in the ratio edit rules that the administrative and corresponding survey values should be within 10% of each other.

The Bayesian model is very much driven by the data, and requires setting only a few parameters in the R code (which is available on the FSRDC server and in our replication package). Following Kim et al. (2015), we set the maximum possible number of components distributions, K. We set K=50, which is large enough that no data are in the lowest probability components in any industry-year. For the Markov Chain Monte Carlo, we

[77] In addition to ratio edit rules, the Kim et al. (2015) algorithm is designed to also use balance edits which require entries to add up. For instance, total wages = production worker wages + non-production worker wages, or more generally $\left(x_{iT_\ell} - \sum_{j\in\beta_\ell} x_{ij} = 0\right)$ for x_{iT_ℓ} as the total for the ℓth balance rule for the set of component variables β_ℓ. In their application of the algorithm to one industry in the 2007 CMF, Kim et al. (2015) use 12 variables that are subject to ratio edits, six of which are totals subject to balance edits, and 15 component variables that are subject to balance edits but not ratio edits.

[78] For instance, rules $x_1 \leq x_2$ and $x_2 \leq x_3$ imply $x_1 \leq x_3$.

choose a burn-in of 2000 iterations, which is long enough that even the largest industries' distributions appear to converge and have good mixing properties.[79] Following Kim et al. (2015) we then run the chain another 50,000 iterations, keeping every 500th iteration so that we have 100 completed datasets ("implicates"). Other than telling the algorithm the folder structure, which variables to use, and what ratio and balance edits the variables are subject to, no other choices need to be made by the modeller.

There is one caveat which applies to any edit-imputation method that uses ratio edits, including the Census Bureau's methods. In the U.S. CMF, the questionnaires ask the respondent to report in thousands of dollars. In some cases, the respondents appear to report in dollars. The Census Bureau determines this by looking at the ratio of each the dollar-valued variables over employment. If for every dollar-valued variable X, a plant's ratio of $\frac{X}{employment}$ fails the ratio edit, but $\frac{X/1000}{employment}$ satisfies the edit, then the Bureau replaces the dollar-valued variable X with $\frac{X}{1000}$. Note that if the Census Bureau did not do this, its Fellegi and Holt (1976) algorithm would choose to replace the employment value for this plant and keep all the (very large) reported dollar valued variables, since that would minimize the number of changes to the data. Before feeding the U.S. captured data into the Bayesian edit-imputation algorithm, we replace the captured value with the final value for any variable that was edited by the Census Bureau using this divide-by-1000 edit. Note that the Indian ASI data does not appear to have this problem.

In our baseline edit-imputation model we have 4 variables subject to 6 ratio edits (3 ratios, each with an upper and lower bound). For the U.S. CMF we have separate models for each of our 1,412 industry-years. On the FSRDC server, which only allows us to run up to 10 jobs in parallel, we are able to run all of the MCMC chains (52,000 iterations for each industry-year) in less than 24 hours.

C.I Motivation for the Editing/Imputed Model

In this subsection, we briefly motivate the approach of the algorithm. Kim et al. (2015) describe in detail how the method works, in particular in their Appendix A.

The goal of the editing and imputing process is for the cleaned data to be likely given a model for reporting error, likely given a model for error indicators, and likely given a model for the underlying data.

More formally,

$$f\left(\boldsymbol{x_i}, \boldsymbol{s_i} | \boldsymbol{y_i}, A_i\right) \propto f\left(\boldsymbol{y_i} | \boldsymbol{x_i}, \boldsymbol{s_i}, A_i\right) f\left(\boldsymbol{s_i}, A_i | \boldsymbol{x_i}\right) f\left(\boldsymbol{x_i}\right). \tag{A. 1}$$

For the model of reporting error, we maintain the U.S. Census Bureau's (implicit) approach to ratio and balance edits: data reported with error provides no information on the true value.[80] Therefore, $f\left(\boldsymbol{y_i} | \boldsymbol{x_i}, \boldsymbol{s_i}, A_i\right)$ is uniform over the support of feasible values

[79] Note that each iteration for a given industry-year involves taking draws for all of the missing and faulty data until all the observations we are modelling have been filled in.

[80] One important exception to this rule is the divide by 1000 edit (also known as units errors or "rounded" edits), where the original reported value is divided by 1000 and then rounded to the nearest unit. Our approach did poorly replicating those edits (partially because we imposed a flat prior for which variable

if $y_{ij} \neq x_{ij}$.

However, unlike the Census Bureau, our prior is a uniform distribution for the errors. That is to say, we do not start with weights on which variables are more likely to be reported with error, so all candidates $\boldsymbol{s_i}$ that result in feasible solutions are initially equally likely.

For the model for the underlying data, we assume that each establishment belongs to one of K mixture components (z). After assuming K, we need to estimate the probability of membership in each component (π_z), and within each component the mean vector $(\boldsymbol{\mu})$ and covariance matrix $(\boldsymbol{\Sigma})$. In order to ensure that all of the draws will pass the ratio edits, we impose that the distribution of $\boldsymbol{x_i}$ conditional on μ, Σ, z_i, given feasible region $\mathcal{D}$ is

$$f\left(\boldsymbol{x}_i|\theta_i\right) = \mathcal{N}\left(\boldsymbol{x}_i|\boldsymbol{\mu}_{z_i}, \boldsymbol{\Sigma}_{z_i}\right) \mathbb{1}\left[\boldsymbol{x}_i \in \mathcal{D}\right]$$

where $\mathbb{1}$ is the indicator function.

D Measuring Aggregate Productivity

In this section we describe the results of a more flexible model of misallocation than in the text, using results from Bils et al. (2021) and (particularly) Blackwood et al. (2021).

The first change is allowing heterogeneous sectoral demand elasticities. This turns out to be fairly straightforward - instead of σ in the expression for sectoral efficiency, each sector has its own σ_s

$$\widetilde{TFP_s}^{\text{Sector-specific}} = \left(\sum_{i=1}^{M} \widetilde{TFPQ_{si}}^{\sigma_s - 1} \widetilde{TFPR_{si}}^{1-\sigma_s}\right)^{\frac{1}{\sigma_s - 1}}. \tag{A. 2}$$

The aggregation to sector-wide aggregate efficiency is the same as in Equation 3. The second change is to allow for non-constant returns to scale. In this case, expected output (divided by TFPQ) is $\left(\left(K_{si}^{\alpha_s} L_{si}^{1-\alpha_s}\right)^{\gamma_s} M_{si}^{1-\gamma_s}\right)^{\xi_s}$ where $\xi_s > 0$.

$$\widetilde{TFPR_s}^{\text{NCR}} = \frac{\sum_{i \in s} A_{is}^{\frac{\sigma_s - 1}{\sigma_s - (\sigma_s - 1)\xi_s}} TFPR_{is}^{\frac{1-\sigma_s}{\sigma_s - (\sigma_s - 1)\xi_s}}}{\sum_{i \in s} A_{is}^{\frac{\sigma_s - 1}{\sigma_s - (\sigma_s - 1)\xi_s}} TFPR_{is}^{\frac{\sigma_s}{(\sigma_s - 1)\xi_s - 1}}} \tag{A. 3}$$

$\xi_s = 1$ simplifies to the $\widetilde{TFPR_s}$ used for the main results in Equation 3. The new equation for aggregating to sectoral efficiency is

$$TFP_s^{\text{NCR}} = \left(\sum_{i \in s} A_{si}^{\frac{\sigma_s - 1}{\sigma_s - (\sigma_s - 1)\xi_s}} \widetilde{TFPR_{si}}^{\text{NCR}\,\frac{\xi_s(1-\sigma_s)}{\sigma_s - (\sigma_s - 1)\xi_s}}\right)^{\frac{1}{\sigma - 1}}. \tag{A. 4}$$

With fixed prices, aggregate efficiency combines any of the above measures of sectoral

was more likely to be reported with error), and so we accepted all of the Census Bureau's rounding edits and otherwise used raw data.

efficiency with a Cobb-Douglas aggregator using the sectoral gross output shares.

$$\widetilde{TFP} = \prod_{s \in S} \widetilde{TFP_s}^{\text{NCR}_s^{\theta}} \tag{A. 5}$$

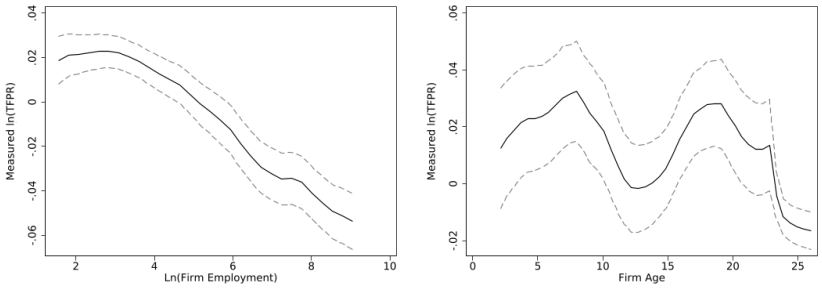
If production is roundabout (so some production gets used as an intermediate good), then the aggregation essentially scales as a function of the intermediates share.

$$\widetilde{TFP} = \prod_{s \in S} \widetilde{TFP_s}^{\text{NCR}^{\frac{\theta_s}{\sum_{k \in S}\left(\theta_k\left(1-\frac{\gamma_s}{\zeta_k}\right)+\theta_s\frac{\gamma_s}{\zeta_s}(1-\zeta_s)\right)}}} \tag{A. 6}$$

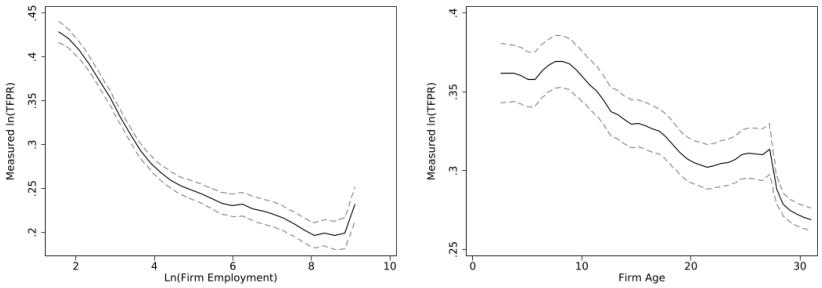
The results using these more general models are in Appendix Table 10.

Figure A. 1: Difference between Captured and Final Productivity, by Firm Employment and Age

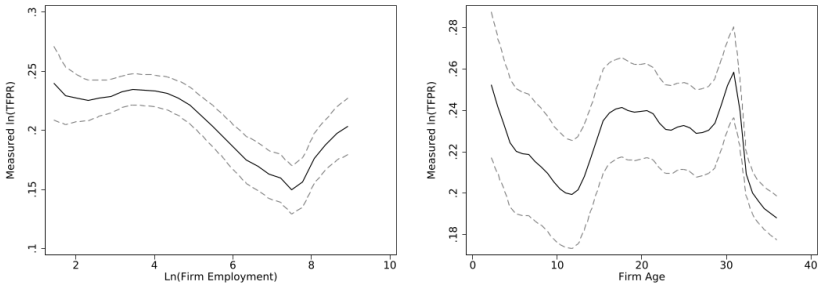
(a) **2002**



(b) **2007**



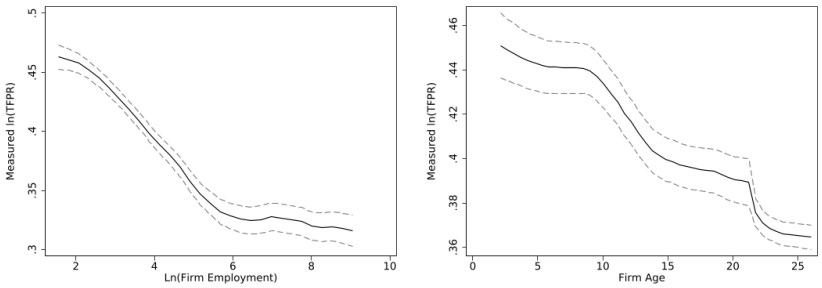
(c) **2012**



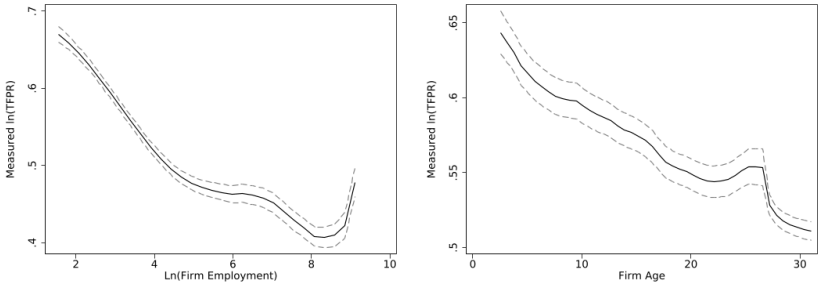
This figure plots a local polynomial regression, predicting the value of ln(TFPR) in the captured data minus the value in the final, over firm age and (ln) firm employment, as described in the text (Figure 2 instead shows the absolute difference, so is bounded below by 0 for each plant). To comply with Census disclosure rules, the 5 percent tails of each graph have been trimmed.

Figure A. 2: Absolute Difference between Captured and Bayesian edited Productivity, by Firm Employment and Age

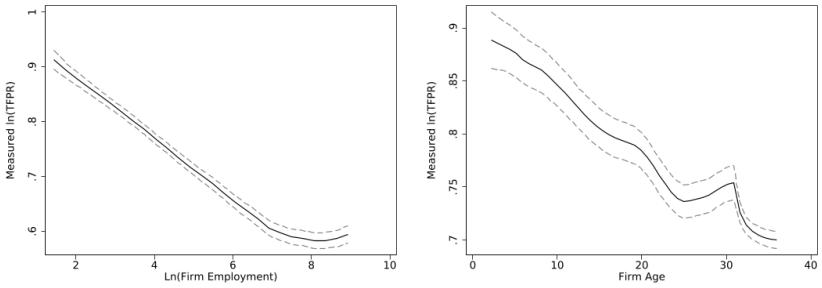
(a) **2002**



(b) **2007**



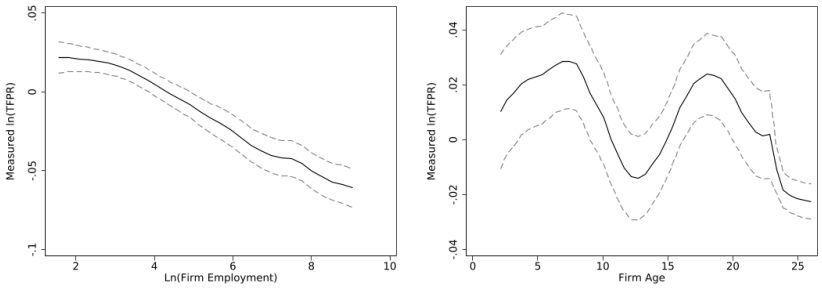
(c) **2012**



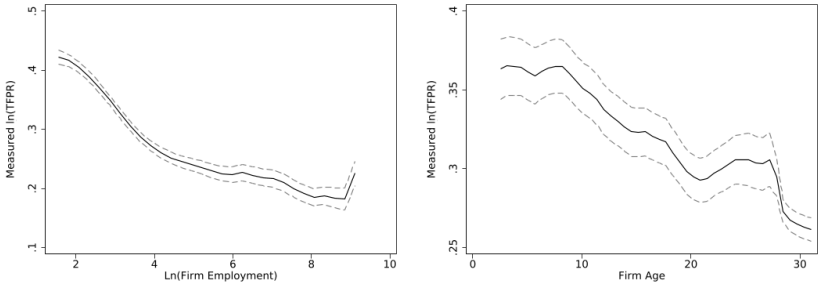
This figure plots a local polynomial regression, predicting the (absolute) difference between ln(TFPR) in the Bayesian edited vs captured data, over firm age and (ln) firm employment, as described in the text. To comply with Census disclosure rules, the 5 percent tails of each graph have been trimmed.

Figure A. 3: Difference between Captured and Bayesian edited Productivity, by Firm Employment and Age

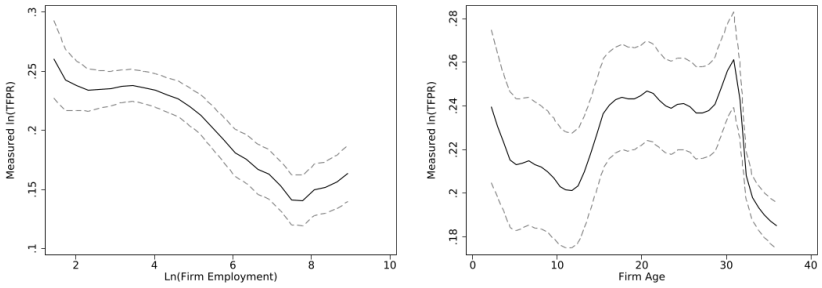
(a) **2002**



(b) **2007**



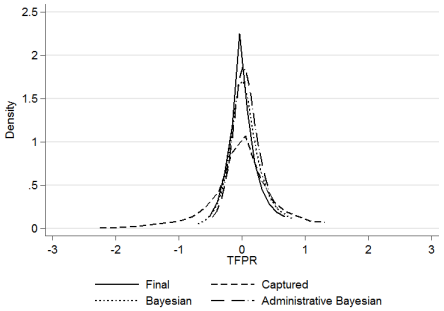
(c) **2012**



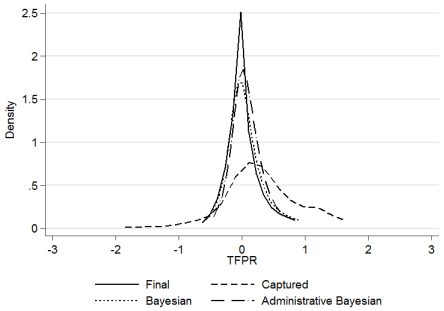
This figure plots a local polynomial regression, predicting the value of ln(TFPR) in the captured data minus the value in the Bayesian-edited data, over firm age and (ln) firm employment, as described in the text. To comply with Census disclosure rules, the 5 percent tails of each graph have been trimmed.

Figure A. 4: Productivity Density, Captured vs Final vs Bayesian-edited (with and without administrative data)

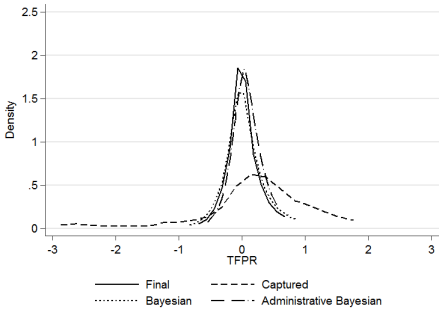
(a) **2002**



(b) **2007**



(c) **2012**



This figure plots the kernel densities for (scaled) TFPR in the captured and final data. For comparability across graphs, and to comply with Census disclosure rules, we trim one percent of the left tail and five percent of the right tail.

Appendix Table 1. Changes Made to the U.S. Census of Manufacturers

Edit/Impute	Occurs When…
(1)	(2)
Administrative (A)	The item is imputed by direct substitution of corresponding administrative data (for the same establishment/record).
Cold Deck Statistical (B)	The item is imputed from a statistical (regression/beta) model based on historic data.
Analyst Corrected (C)	The reported value fails an edit, and an analyst directly corrects the value.
Model (Donor) Record (D)	The item is imputed using hot deck methods.
Receipts (F)	EIN/SSN cross-reference match or primary SSN (one-for-one match).
High/Low (E)	The item is imputed with a value near the endpoints of the imputation range.
Historic (H)	The item is imputed by using historic ratio data for the same establishment.
Subject Matter Rule (J)	The item is imputed using a subject matter defined rule (e.g. y=1/2x).
Prior Year Ratio (HQ)	The item is imputed using a ratio of historic data times current reported values.
Raked (K)	The sum of a set of detail items do not balance to the total. The details are then changed proportionally to correct the imbalance.
Logical (L)	The item's imputation value is defined by an additive mathematical relationship (e.g., obtaining a missing detail item by subtraction).
Midpoint (M)	The item is imputed by direct substitution of midpoint of imputation range.
Rounded (N)	The reported value is replaced by its original value divided by 1000.
Prior Year Administrative (P)	The item is imputed by ratio imputation using corresponding administrative data from prior year (for same establishment).
Direct Substitution (S)	The item is imputed by direct substitution of another item's value (from within the same questionnaire.)
Trim-and-Adjusted (T)	The item was imputed using the Trim-and Adjust balancing algorithm.
Unable to Impute (U)	a statistically reasonable value for the original data.
Industry Average (V)	The item is imputed by ratio imputation using an industry average.
Warm Deck Statistical (W)	The item is imputed from a statistical (regression/beta) model based on current data.
Unusable (X)	The sum of a set of detail items cannot be balanced to the total because none of the scripted solutions achieved a balance.
Data Impute (2)	Data item imputed from a reported data item
Payroll Quarterly (4)	First quarter payroll reported higher than annual payroll; data adjusted during general data prep or legacy edits

Notes: Edit/imputation descriptions for the U.S. Census of Manufacturers, from Grim (2011) and White et al. (2018), with a few additional categories. These flags can be requested at the FSRDCs for the manufacturing data in this paper, and explain what changed when the final data disagree with the captured data. There is an additional flag used in the paper, (G), which corresponds to "goldplated" data that an analyst ensures is not changed.

Appendix Table 2. Annual Changes Made to the U.S. Census of Manufacturers

Edit/Impute Label	Share of Plants (2002)	Share of Plants (2007)	Share of Plants (2012)
(1)	(2)	(2)	(3)
Impute Missing Value	0.464	0.443	.546
Logical Edit for Payroll	0.133	0.079	.055
Analyst Edits	0.023	0.057	0.085
Logical Edit for Shipments	0.045	0.044	0.03
Logical Edit for Materials	0.035	0.038	0.036
Regression Edit for Materials	0.026	0.033	0.02
Edit from Administrative Records	0.002	0.019	0.033
Divide by 1000	0.003	0.006	0.008
Other Capital Edits	0.036	0.336	0.238
Other Replicable Change	0.018	0.02	0.017
Any Change, Not Elsewhere Classified	0.063	-	0.0003

Notes: Lables are defined in Table 1. The "Share of Plants" is the share of plants who have been affected by the corresponding label in each year. The value for "Any Change, Not Elsewhere Classified" could not be disclosed in 2007 because fewer than 10 firms are affected. The sample is plants for whom it is possible to calculate productivity in the final data.

Appendix Table 3. Dispersion of Plant Characteristics in the Captured and Final Data

	Captured Data			Final Data		
	90/10 ratio	75/25 ratio	Standard Deviation	90/10 ratio	75/25 ratio	Standard Deviation
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A. Capital						
2002	4.784	2.546	1.913	4.262	2.161	1.721
2007	5.142	2.702	2.08	3.777	1.866	1.524
2012	5.303	2.686	2.155	3.684	1.863	1.49
Panel B. Materials						
2002	3.39	1.84	1.383	3.129	1.639	1.272
2007	3.946	2.087	1.634	3.172	1.653	1.282
2012	4.652	2.369	1.902	3.119	1.669	1.274
Panel C. Payroll						
2002	4.602	2.373	1.882	3.939	1.994	1.589
2007	4.717	2.441	1.913	4.124	2.103	1.642
2012	4.986	2.644	2.002	4.122	2.107	1.644
Panel D. Shipments						
2002	3.679	2.007	1.501	3.458	1.795	1.412
2007	3.798	2.034	1.577	3.586	1.854	1.446
2012	4.575	2.364	1.852	3.577	1.873	1.445

Notes: Tthis table reports moments of the distribution of the inputs and shipments. The variables are logged and demeaned within each industry. Within each dataset, the samples are balanced (it must be possible to calculate productivity), but the samples in the captured and final data are different.

Appendix Table 4. Dispersion of Plant Input Shares, Captured vs. Final

	Captured Data / Final Data		
	90/10 ratio	75/25 ratio	Standard Deviation
	(1)	(2)	(3)
Panel A. Capital/Shipments			
2002	1.452	1.709	1.41
2007	1.888	2.19	1.852
2012	2.038	2.163	2.159
Panel B. Materials/Shipments			
2002	1.596	1.718	1.85
2007	1.996	2.146	2.121
2012	2.401	2.644	2.616
Panel C. Payroll/Shipments			
2002	1.951	2.02	1.977
2007	1.771	2.073	1.833
2012	2.216	2.268	2.171

Notes: This table reports moments of the distribution of inputs/shipments. The input shares are logged and demeaned within each industry. We calculate the relevant value in the captured data divided by its final counterpart, and report the ratio. The underlying values are in Appendix Table 3. Column 4 reports the within R^2 of a regression of the corresponding (logged) variable in the captured data on its value in the final data, with 6-digit NAICS fixed effects. Within each dataset, the samples are balanced (it must be possible to calculate productivity), but the samples in the captured and final data are different.

Appendix Table 5. Characteristics of Changes Between Final and Captured Data, Productivi

Type of Change	Share of Plants (2002)	Share of Plants (2007)	Share of Plants (2012)
(1)	(2)	(2)	(3)
TFPQ:			
… Final < Captured	0.312	0.368	0.488
… Unchanged	0.544	0.196	0.391
… Final > Captured	0.144	0.436	0.121
TFPR:			
… Final < Captured	0.313	0.369	0.495
… Unchanged	0.544	0.196	0.39
… Final > Captured	0.144	0.435	0.115

Notes: TFPR and TFPQ are calculated using sectoral cost shares (and TFPQ uses the model to convert from revenues to quantities). The variables are logged but not demeaned. The samples is plants for whom it is possible to calculate productivity in the captured data

Appendix Table 6. Productivity Dispersion in the Final and Captured Data

	Captured Data			Final Data		
	90/10	75/25	Standard Deviation	90/10	75/25	Standard Deviation
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A. TFPQ						
2002	7.215	5.606	3.028	1.944	0.951	0.802
2007	6.546	4.651	2.64	2.348	1.113	0.946
2012	7.42	5.165	3.243	1.792	0.904	0.738
Panel B. TFPR						
2002	1.564	0.579	0.906	0.731	0.286	0.402
2007	1.684	0.795	0.947	0.796	0.31	0.432
2012	2.36	0.94	1.197	0.754	0.305	0.407

Notes: This Table reports moments of the distribution of TFPQ and TFPR, logged and demeaned within each industry (as described in the text). Within each dataset, the samples are balanced (it must be possible to calculate productivity), but the samples in the captured and final data are different.

Appendix Table 7: Alternative Measures of Factor Allocation

	Captured Data	Final Data	Ratio
	(1)	(2)	(3)
Panel A. OP "within" share			
2002	0.003	0.935	0.003
2007	0.011	0.923	0.012
2012	0.001	1.083	0.001
Panel B. Payroll and Shipments, R^2			
2002	0.711	0.855	0.832
2007	0.634	0.836	0.758
2012	0.563	0.846	0.665
Panel C. All Inputs and Shipments, R^2			
2002	0.826	0.899	0.919
2007	0.816	0.897	0.910
2012	0.803	0.921	0.872
Panel D. TFPQ and TFPR, R^2			
2002	0.719	0.458	1.570
2007	0.766	0.488	1.570
2012	0.791	0.446	1.774

Notes: This table reports the distribution of alternative measures of (mis)allocation. We calculate the relevant value in the captured data divided by its final counterpart, and report the ratio. Panel A reports the ratio of the unweighted-weighted average TFPR over the by the shipments-weighted average, inspired by Olley and Pakes (1996). Panel B reports the within R2 of a regression of (log) payroll on (log) output, with 6-digit NAICS fixed effects. Panel C reports the within R2 of a regression of all three (log) inputs on (log) output, with 6-digit NAICS fixed effects. Panel D reports the within R^2 of a regression of (log) TFPQ on (log) TFPR, with 6-digit NAICS fixed effects. Within each dataset, the samples are balanced (it must be possible to calculate productivity), but the samples in the captured and final data are different.

Appendix Table 8. Measured Allocative Efficiency in the Final and Captured Data

	Captured Data			Final Data		
	No Trimming	1 Percent Trimming	2 Percent Trimming	No Trimming	1 Percent Trimming	2 Percent Trimming
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A. Gross Output						
2002	0.00001	0.023	0.047	0.027	0.147	0.197
2007	0.00001	0.017	0.031	0.01	0.092	0.11
2012	0.000002	0.002	0.005	0.025	0.116	0.14
Panel B. Value Added						
2002	0.00002	0.04	0.08	0.005	0.179	0.262
2007	0.00003	0.022	0.06	0.005	0.124	0.192
2012	0.0000003	0.004	0.007	0.001	0.137	0.214
Panel C. Annual Survey of Manufacturers						
2002	-	0.04	0.058	0.035	0.141	0.197
2007	0.0003	0.036	0.071	0.034	0.094	0.125
2012	0.0007	0.035	0.062	0.033	0.138	0.158

Notes: The equations for calculating allocative efficiency are as described in the text. For the trimming, we drop the (upper and lower) extremes for TFPQ and TFPR. The values are balanced within each dataset (it must be possible to calculate productivity), but unbalanced across. Panel A uses a gross output specification as in Bils et al. (2021) and Blackwood et al. (2021). Panel B uses a value added specification as in Hsieh and Klenow (2009). Panel C uses only the ASM plants (and the corresponding weights), the ASM value for the captured data with no trimming could not be disclosed in 2002. The samples are different in the captured and final data.

Appendix Table 9. Dispersion and Measured Misallocation in the Final Data, Only for Plants in the Captured Sample

	Final (Captured Sample) / Final (Full Sample)		
	2002	2007	2012
	(1)	(2)	(3)
Materials SD	1.076	1.039	1.025
Materials 90/10	1.033	1.004	.993
Materials 75/25	1.499	1.417	1.285
Materials Revenue Share SD	1.23	1.177	1.056
Materials Revenue Share 90/10	1.092	1.065	.976
Materials Revenue Share 75/25	1.024	1.004	.988
Payroll SD	0.998	0.986	0.942
Payroll 90/10	0.961	0.952	0.954
Payroll 75/25	1.238	1.153	1.264
Payroll Revenue Share SD	1.07	1.051	1.133
Payroll Revenue Share 90/10	1.039	1.027	1.094
Payroll Revenue Share 75/25	0.947	0.941	0.93
Shipments SD	0.977	0.975	0.945
Shipments 90/10	1.022	0.998	0.988
Shipments 75/25	0.986	0.967	0.978
TFPQ SD	0.97	0.96	0.961
TFPQ 90/10	0.962	0.921	0.957
TFPQ 75/25	0.922	0.911	0.973
TFPR SD	0.936	0.977	0.974
TFPR 90/10	1.001	1.005	0.963
TFPR 75/25	1.239	1.19	1.122
Weighted OP Productivity	1.153	1.123	1.05
Unweighted OP Productivity	1.065	1.046	0.992
Payroll and Shipments, R^2	1.038	1.033	0.985
TFPQ and TFPR, R^2	0.958	0.982	0.964
Gross Output Allocative Efficiency	1.463	2.636	1.753

Notes: We calculate the relevant value in the captured data divided by its final counterpart, and report the ratio. The sample is balanced, the plants for whom it is possible to calculate productivity in the captured data. The values for the ratios from the unbalanced samples are in Tables 4, 5, 6, and Appendix Table 7.

Appendix Table 10. Measured Allocative Efficiency, Captured vs. Final under alternative assumptions

Edit/Impute Label	2002	2007	2012
(1)	(2)	(2)	(3)
Baseline	0.0004	0.001	0.00007
Reported Capital Values	0.0003	0.0008	0.00005
Ahmad and Riker (2019) demand elasticities	0.00006	0.0009	0.00005
Demirer (2021) pf elasticities	0.0006	0.004	0.008
Blackwood et al. (2021) "OP" pf elasticities	0.001	0.00004	0.002
Blackwood et al. (2021) "OPD" pf elasticities	0.006	0.00006	0.0001
Baseline, Blackwood et al. (2021) sample	0.0003	0.00003	0.00003
Returns to Scale .9	0.0003	0.001	0.00008
Returns to Scale 1.1	0.039	0.028	0.105
Returns to Scale .9, roundabout	<0.0000001	<0.0000001	<0.0000001
Roundabout (Constant Returns to Scale)	<0.0000001	<0.0000001	<0.0000001
Returns to Scale 1.1, roundabout	0.00005	0.00003	0.01

Notes: We calculate the relevant value in the captured data divided by its final counterpart, and report the ratio. The values for calculating allocative efficiency are as described in the text, using a gross-output specification and all of the plants in the Census of Manufacturers. The values are balanced within each dataset (it must be possible to calculate productivity), but unbalanced across. The "reported capital values" are the ones directly reported in the data (multiplied by a 10% rental rate), instead of the values from the BLS-Census Multifactor Productivity project. Ahmad and Riker (2019) use the revenue share of the flexible inputs to back out demand elasticities, we use the average of their estimates. Demirer (2021) and Blackwood et al. (2021) present alternative production function elasticities. Blackwood et al. (2021) only calculate elasticities for 50 sectors, so we also show how our results change when we use a cost shares but only within their sample of plants. The final models estimate the gains under alternative returns to scale, and under roundabout production, following Blackwood et al. (2021).

Appendix Table 11. Effect of Flags on Measured Productivity Dispersion, Shapley Value

Edit/Impute Label	Standard Deviation of TFPR	Standard Deviation of TFPQ	R^2 TFPQ and TFPR
(1)	(2)	(2)	(3)
Impute Missing Value, 2002	0.163	0.012	.167
… 2007	0.169	0.019	.144
… 2012	0.127	0.024	.134
Logical Impute for Payroll, 2002	0.085	0.068	.085
… 2007	0.086	0.021	.053
… 2012	0.035	0.023	.05
Analyst Edits, 2002	0.118	0.048	0.077
… 2007	0.13	0.106	0.116
… 2012	0.592	0.79	0.406
Logical Impute for Shipments, 2002	0.079	0.264	0.119
… 2007	0.097	0.464	0.13
… 2012	0.015	0.031	0.026
Logical Impute for Materials, 2002	0.117	-0.008	0.077
… 2007	0.035	-0.011	0.049
… 2012	0.01	0.006	0.03
Regression Impute for Materials, 2002	0.147	0.012	0.166
… 2007	0.122	-0.009	0.161
… 2012	0.045	0.012	0.085
Impute from Administrative Records, 2002	0.002	0.018	0.007
… 2007	0.012	0.006	0.012
… 2012	0.025	0.0005	0.015
Divide by 1000, 2002	0.084	0.121	0.038
… 2007	0.109	0.243	0.043
… 2012	0.022	-0.0007	0.02
Other Capital Edits, 2002	0.032	0.028	0.029
… 2007	0.177	0.096	0.2
… 2012	0.119	0.08	0.183
Other Replicable Change, 2002	0.135	0.41	0.129
… 2007	0.0002	-0.001	0.0002
… 2012	0.0002	-0.0002	0.0001
Any Change, Not Elsewhere Classified, 2002	0.037	0.027	0.106
… 2007	0.062	0.067	0.092
… 2012	0.011	0.034	0.052

Notes: Lables are defined in Table 1. We calculate Gross Output TFPR and TFPQ (as in Table 5) with no trimming for every possible combination of flags (if a flag is turned on, we use only final values for the plant, not just for the particular characteristic affected by the flag). We then calculate the Shapley (1953) value for each flag for each outcome, and report the share of the change from captured to final measured misallocation attributable to each flag (so the columns would sum to one if not for rounding). Column 3 reports the within R2 of a regression of (log) TFPQ on (log) TFPR, with 6-digit NAICS fixed effects.

Appendix Table 12. Characteristics of Changes Between Bayesian Edited, Captured, and Final Data

Type of Change	Share of Plants (2002)	Share of Plants (2007)	Share of Plants (2012)
(1)	(2)	(2)	(3)
Panel A: Bayesian vs. Captured			
Materials			
… Bayesian < Captured	0.006	0.025	0.02
… Unchanged	0.956	0.931	0.938
… Bayesian > Captured	0.038	0.045	0.042
Payroll:			
… Bayesian < Captured	0.01	0.018	0.026
… Unchanged	0.982	0.968	0.962
… Bayesian > Captured	0.008	0.014	0.013
Shipments			
… Bayesian < Captured	0.009	0.016	0.021
… Unchanged	0.973	0.956	0.959
… Bayesian > Captured	0.018	0.028	0.02
Panel B: Bayesian vs. Final			
Materials			
… Bayesian < Final	0.176	0.171	0.174
… Unchanged	0.697	0.697	0.725
… Bayesian > Final	0.127	0.132	0.101
Payroll:			
… Bayesian < Final	0.212	0.115	0.119
… Unchanged	0.721	0.827	0.823
… Bayesian > Final	0.067	0.058	0.058
Shipments			
… Bayesian < Final	0.121	0.134	0.104
… Unchanged	0.832	0.799	0.839
… Bayesian > Final	0.046	0.067	0.057

Notes: The "Share of Plants" is the share of plants who have been affected by the corresponding label in each year. Any Regression Flag includes both imputed and edited values. The samples are balanced within each panel, to plants for which it is possible to calculate productivity in both relevant datasets. The Bayesian edited data is constructed following Kim et al. (2015), as described in the text.

Appendix Table 13. Distribution of the Captured/Bayesian Ratios

	First Percentile (1)	Fifth Percentile (2)	Tenth Percentile (3)	Ninetieth Percentile (4)	Ninety-Fifth Percentile (5)	Ninety-Ninth Percentile (6)
Panel A. Materials						
2002	0.011	1	1	1	1	1
2007	0.008	0.998	1	1	1	2.868
2012	0.009	0.999	1	1	1	36.42
Panel B. Payroll						
2002	0.994	1	1	1	1	1
2007	0.155	1	1	1	1	2.166
2012	0.185	1	1	1	1	542.2
Panel C. Shipments						
2002	0.095	1	1	1	1	1
2007	0.058	1	1	1	1	1.967
2012	0.07	1	1	1	1	6.337

Notes: We calculate the value in the captured data divided by its counterpart in the Bayesian-edited data. This Table reports the distribution of those ratios. For disclosure purposes, the reported values are not the cutoff at the exact percentile, but the average value of all of the plants within the centile. The sample is plants for whom it is possible to calculate productivity in the captured data. The Bayesian edited data is constructed following Kim et al. (2015), as described in the text.

Appendix Table 14. Distribution of the Final/Bayesian Ratios

	First Percentile (1)	Fifth Percentile (2)	Tenth Percentile (3)	Ninetieth Percentile (4)	Ninety-Fifth Percentile (5)	Ninety-Ninth Percentile (6)
Panel A. Materials						
2002	0.057	0.542	0.878	1.406	2.42	9.299
2007	0.048	0.529	0.857	1.355	2.441	10.44
2012	0.029	0.547	0.964	1.229	2.279	10.56
Panel B. Payroll						
2002	0.216	0.869	1	1.018	1.345	2.594
2007	0.135	0.929	1	1.008	1.278	2.674
2012	0.104	0.868	1	1.022	1.385	2.905
Panel C. Shipments						
2002	0.193	1	1	1.033	1.426	3.154
2007	0.144	0.827	1	1.052	1.572	3.617
2012	0.095	0.916	1	1.001	1.295	3.421

Notes: We calculate the value in the final data divided by its counterpart in the Bayesian-edited data. This Table reports the distribution of those ratios. For disclosure purposes, the reported values are not the cutoff at the exact percentile, but the average value of all of the plants within the centile. The sample is plants for whom it is possible to calculate productivity in both datasets. The Bayesian edited data is constructed following Kim et al. (2015), as described in the text.

Appendix Table 15. Dispersion of Plant Characteristics, Bayesian vs. Captured

	Captured Data / Bayesian Edited Data		
	90/10 ratio	75/25 ratio	Standard Deviation
	(1)	(2)	(3)
Panel A. Materials			
2002	1.089	1.114	1.084
2007	1.274	1.274	1.287
2012	1.279	1.249	1.307
Panel B. Payroll			
2002	1.084	1.077	1.108
2007	1.101	1.099	1.119
2012	1.071	1.068	1.078
Panel C. Shipments			
2002	1.052	1.093	1.053
2007	1.082	1.104	1.102
2012	1.15	1.109	1.162

Notes: This table reports moments of the distribution of the inputs and shipments. The variables are logged and demeaned within each industry. We calculate the relevant value in the captured data divided by its Bayesian-edited counterpart, and report the ratio. Within each dataset, the samples are balanced (it must be possible to calculate productivity), but the samples in the captured and Bayesian-edited data are different. The Bayesian edited data is constructed following Kim et al. (2015), as described in the text.

Appendix Table 16. Dispersion of Plant Characteristics, Final vs. Bayesian

	Final Data / Bayesian Edited Data		
	90/10 ratio	75/25 ratio	Standard Deviation
	(1)	(2)	(3)
Panel A. Materials			
2002	1.005	0.993	0.997
2007	1.024	1.009	1.009
2012	0.858	0.88	0.876
Panel B. Payroll			
2002	0.928	0.905	0.935
2007	0.963	0.946	0.961
2012	0.885	0.851	0.885
Panel C. Shipments			
2002	0.989	0.977	0.991
2007	1.021	1.006	1.01
2012	0.899	0.879	0.907

Notes: This table reports moments of the distribution of the inputs and shipments. The variables are logged and demeaned within each industry. We calculate the relevant value in the final data divided by its Bayesian-edited counterpart, and report the ratio. Within each dataset, the samples are balanced (it must be possible to calculate productivity), but the samples in the final and Bayesian-edited data are different. The Bayesian edited data is constructed following Kim et al. (2015), as described in the text.

Appendix Table 17. Productivity Dispersion, Bayesian vs. Captured

	Captured Data / Bayesian Data		
	90/10	75/25	Standard Deviation
	(1)	(2)	(3)
Panel A. TFPQ			
2002	3.919	5.809	4.07
2007	3.308	4.679	3.325
2012	3.237	4.407	3.556
Panel B. TFPR			
2002	1.773	1.54	2.059
2007	1.857	2.103	2.1
2012	2.384	2.26	2.433

Notes: This table reports moments of the distribution of distribution of TFPQ and TFPR. TFPR and TFPQ are calculated using sectoral cost shares (and TFPQ uses the model to convert from revenues to quantities). The variables are logged and demeaned within each industry, as described in the text. We calculate the relevant value in the captured data divided by its Bayesian-edited counterpart, and report the ratio. Within each dataset, the samples are balanced (it must be possible to calculate productivity), but the samples in the captured and Bayesian-edited data are different. The Bayesian edited data is constructed following Kim et al. (2015), as described in the text.

Appendix Table 18. Productivity Dispersion, Bayesian vs. Final

	Final Data / Bayesian Data		
	90/10	75/25	Standard Deviation
	(1)	(2)	(3)
Panel A. TFPQ			
2002	1.056	0.985	1.078
2007	1.186	1.12	1.191
2012	0.782	0.771	0.809
Panel B. TFPR			
2002	0.829	0.761	0.914
2007	0.878	0.82	0.958
2012	0.762	0.733	0.827

Notes: This table reports moments of the distribution of distribution of TFPQ and TFPR. TFPR and TFPQ are calculated using sectoral cost shares (and TFPQ uses the model to convert from revenues to quantities). The variables are logged and demeaned within each industry, as described in the text. We calculate the relevant value in the final data divided by its Bayesian-edited counterpart, and report the ratio. Within each dataset, the samples are balanced (it must be possible to calculate productivity), but the samples in the final and Bayesian-edited data are different. The Bayesian edited data is constructed following Kim et al. (2015), as described in the text.

Table 8. Measured Allocative Efficiency, Bayesian vs. Captured

	Captured Data / Bayesian Data		
	No Trimming	1 Percent Trimming	2 Percent Trimming
	(1)	(2)	(3)
Panel A. Gross Output			
2002	0.0002	0.186	0.304
2007	0.0003	0.167	0.226
2012	0.00006	0.024	0.043
Panel B. Value Added			
2002	0.004	0.284	0.368
2007	0.001	0.18	0.316
2012	0.0004	0.044	0.047
Panel C. Annual Survey of Manufacturers (Gross Output)			
2002		0.242	0.283
2007	0.007	0.261	0.386
2012	0.017	0.267	0.41

Notes: The equations for calculating allocative efficiency are as described in the text. For the trimming, we drop the (upper and lower) extremes for TFPQ and TFPR. The values are balanced within each dataset (it must be possible to calculate productivity), but unbalanced across. For each of the statistics in the table we calculate the relevant value in the captured data divided by its Bayesian-edited counterpart, and report the ratio. Panel A uses a gross output specification as in Bils et al. (2021) and Blackwood et al. (2021). Panel B uses a value added specification as in Hsieh and Klenow (2009). Panel C uses only the ASM plants (and the corresponding weights), the ASM value for the captured data with no trimming could not be disclosed in 2002. The Bayesian edited data is constructed following Kim et al. (2015), as described in the text.

Appendix Table 20: Alternative Measures of Factors Allocation, Bayesian vs Captured and Final

	Captured /Bayesian Data	Final /Bayesian Data
	(1)	(2)
Panel A. OP "within" share		
2002	0.002	0.881
2007	0.011	0.921
2012	0.001	1.367
Panel B. Payroll and Shipments, R^2		
2002	0.82	0.986
2007	0.742	0.978
2012	0.661	0.993
Panel C. TFPQ and TFPR, R^2		
2002	1.668	1.063
2007	1.672	1.066
2012	1.669	0.941

Notes: This table reports the distribution of alternative measures of (mis)allocation. We calculate the relevant value in the captured data divided by its Bayesian-edited counterpart, and report the ratio in column 1. Column 2 shows the equivilent ratio for the final data divided by its Bayesian-edited counterpart. Panel A reports the ratio of the unweighted-weighted average TFPR over the by the shipments-weighted average, inspired by Olley and Pakes (1996). Panel B reports the within R2 of a regression of (log) payroll on (log) output, with 6-digit NAICS fixed effects. Panel C reports the within R2 of a regression of all three (log) inputs on (log) output, with 6-digit NAICS fixed effects. Panel D reports the within R^2 of a regression of (log) TFPQ on (log) TFPR, with 6-digit NAICS fixed effects. Within each dataset, the samples are balanced (it must be possible to calculate productivity), but the samples in the captured and final data are different. The Bayesian edited data is constructed following Kim et al. (2015), as described in the text.

Appendix Table 21. Coefficient of Variation for Bayesian-Edited outcomes

	2002	2007	2012
	(1)	(2)	(3)
Materials SD	0.001	0.001	.006
Materials 90/10	0.001	0.001	.006
Materials 75/25	0.001	0.001	.006
Materials Revenue Share SD	0.002	0.003	.011
Materials Revenue Share 90/10	0.003	0.004	.009
Materials Revenue Share 75/25	0.002	0.004	.009
Payroll SD	0.0006	0.0007	0.006
Payroll 90/10	0.0007	0.001	0.009
Payroll 75/25	0.0008	0.0009	0.008
Payroll Revenue Share SD	0.002	0.002	0.008
Payroll Revenue Share 90/10	0.002	0.002	0.007
Payroll Revenue Share 75/25	0.002	0.005	0.006
Shipments SD	0.0007	0.001	0.005
Shipments 90/10	0.0007	0.002	0.006
Shipments 75/25	0.001	0.001	0.006
TFPQ SD	0.008	0.02	0.017
TFPQ 90/10	0.009	0.022	0.014
TFPQ 75/25	0.01	0.015	0.014
TFPR SD	0.003	0.004	0.016
TFPR 90/10	0.003	0.006	0.009
TFPR 75/25	0.003	0.006	0.009
Weighted OP Productivity	0.034	0.048	0.19
Unweighted OP Productivity	0.001	0.002	0.001
Payroll and Shipments, R^2	0.0003	0.0004	0.0005
TFPQ and TFPR, R^2	0.004	0.004	0.003
Gross Output Allocative Efficiency	0.05	0.09	0.094

Notes: Within each industry/year, we use a burn-in of 2000 iterations, and the draw 100 implicates with 500 iterations between implicates. We then calculate the coefficient of variation for each of the statistics from Table 4, 5, 6, and Appendix Table 7. The Bayesian edited data is constructed following Kim et al. (2015), as described in the text, and we use a balanced sample of plants for whom it is possible to calculate productivity in the Bayesian-edited data.

Appendix Table 22. Dispersion for Final Data with Regression Flags

	Standard Deviation		
	2002	2007	2012
	(1)	(2)	(3)
Panel A. Captured Reported and Edited			
Materials, Captured	1.209	1.274	1.045
Materials, Final	0.703	0.777	0.849
Materials, Bayesian	0.892	0.933	0.956
Shipments, Captured	1.21	1.434	1.363
Shipments, Final	0.81	1.027	0.861
Shipments, Bayesian	0.953	1.254	1.125
Panel B. Captured Missing and Imputed			
Materials, Final	0.791	0.77	0.721
Materials, Bayesian	1.02	0.994	0.922
Shipments, Final	0.939	0.851	0.801
Shipments, Bayesian	1.029	0.964	0.888

Notes: This table reports the standard deviation of materials and shipments in the captured, final, and Bayesian-edited data for the plants with a regression flag equal to 1 relative to the values for the plants with a regression flag equal to 0 (for the corresponding characteristic). Panel A is balanced for the plants for whom productivity can be calculated in the captured dataset, and Panel B is balanced within each input (to the plants for whom productivity can be calculated in the Bayesian-edited data. Across the three years, the average (rounded) number of plants represented in Panel A is 56500 plants a year, the sample for materials in Panel B averages 26000 plants a year, the sample for shipments in Panel B averages 5000 plants a year.

Appendix Table 23. Dispersion and Measured Misallocation Under Alternative Bayesian Models

	Single-unit administrative records			Big administrative records			Hybrid		
	2002	2007	2012	2002	2007	2012	2002	2007	2012
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Materials SD	0.97	0.966	0.884	0.954	0.959	0.867	0.959	0.964	0.868
Materials 90/10	0.959	0.959	0.898	0.954	0.961	0.882	0.959	0.965	0.884
Materials 75/25	0.911	0.929	0.91	0.944	0.942	0.915	0.95	0.953	0.92
Materials Revenue Share SD	0.842	0.87	0.837	0.866	0.859	0.838	0.877	0.878	0.847
Materials Revenue Share 90/10	0.819	0.855	0.815	0.824	0.837	0.81	0.846	0.873	0.841
Materials Revenue Share 75/25	0.956	0.963	0.893	0.956	0.968	0.888	0.962	0.973	0.889
Payroll SD	0.993	0.985	0.868	0.974	0.981	0.85	0.978	0.984	0.852
Payroll 90/10	0.985	0.979	0.838	0.982	0.987	0.833	0.986	0.988	0.833
Payroll 75/25	0.917	0.927	0.907	0.919	0.93	0.89	0.93	0.954	0.909
Payroll Revenue Share SD	0.905	0.915	0.893	0.9	0.918	0.875	0.92	0.95	0.905
Payroll Revenue Share 90/10	0.903	0.921	0.898	0.896	0.923	0.874	0.934	0.99	0.927
Payroll Revenue Share 75/25	0.98	0.988	0.866	0.989	1.004	0.873	0.992	1.005	0.874
Shipments SD	1.022	1.016	1.012	1.023	1.021	1.018	1.015	1.005	1.002
Shipments 90/10	0.987	0.977	0.864	0.96	0.966	0.841	0.966	0.974	0.846
Shipments 75/25	0.979	0.972	0.876	0.974	0.979	0.865	0.983	0.991	0.872
TFPQ SD	0.973	0.971	0.883	0.976	0.986	0.882	0.986	0.999	0.89
TFPQ 90/10	0.808	0.799	0.686	0.802	0.855	0.665	0.939	0.976	0.917
TFPQ 75/25	0.82	0.788	0.706	0.825	0.872	0.681	1.049	0.982	1.141
TFPR SD	0.82	0.81	0.763	0.847	0.875	0.738	1.17	1.065	1.913
TFPR 90/10	0.64	0.712	0.711	0.67	0.692	0.705	0.773	0.847	0.834
TFPR 75/25	0.874	0.896	0.836	0.908	0.904	0.843	0.921	0.92	0.857
Weighted OP Productivity	0.804	0.831	0.768	0.835	0.825	0.769	0.853	0.852	0.787
Unweighted OP Productivity	0.768	0.802	0.752	0.773	0.79	0.731	0.821	0.867	0.792
Payroll and Shipments, R^2	0.909	0.901	0.897	0.906	0.889	0.886	0.92	0.907	1.557
TFPQ and TFPR, R^2	0.839	0.788	0.614	0.832	0.813	0.522	1.034	1.107	0.739
Gross Output Allocative Efficiency	1.692	1.736	1.965	1.55	1.413	2.604	0.621	0.373	0.665

Notes: In addition to the baseline Bayesian-edited data (which uses captured employment, materials, payroll, and shipments), we created three alternative datasets, which also use administrative information on payroll and shipments. For the "single-unit administrative records" version of the data, we add in administrative records only for the single unit plants. The ``big administrative records" version of the data uses administrative records for the entire manufacturing sector, as described in the text. The "hybrid" data supplements the ``big administrative records" version of the data with the analyst edited values when possible. The values reported in the table are the values in the relevant dataset/year divided by the corresponding value in the baseline Bayesian-edited dataset, the samples are balanced within datasets but unbalanced across.

Appendix Table 24. Comparing Shipments and Labor Productivity in the Various Datasets to the Administrative Records

	2002	2007	2012	2002	2007	2012	2002	2007	2012
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Panel A: Main Data Versions									
	Captured			Final			Baseline Bayesian		
Shipments, mean	4.03	5.35	50.52	1.04	1	0.99	1	0.98	2.59
Shipments, sd	14.79	15.56	216	0.65	0.5	0.5	1.3	0.98	6.66
Shipments, skew	0.58	0.55	0.5	0.44	0.32	0.23	1.85	1.26	0.38
Labor Productivity, mean	4.75	6.88	81.88	0.86	0.88	0.92	0.88	0.84	0.91
Labor Productivity, sd	20.29	49.51	644	0.29	0.27	0.9	0.32	0.22	0.32
Labor Productivity, skew	0.46	2.28	0.59	0.15	0.34	1.8	0.16	0.14	0.13
R^2, Shipments on Employment	0.64	0.53	0.29	1.2	1.2	1.13	1.18	1.19	1.17
R^2, Admin Shipments	0.66	0.56	0.33	0.81	0.81	0.87	0.8	0.81	0.69
R^2, Admin Labor Productivity	0.29	0.25	0.11	0.29	0.28	0.41	0.36	0.39	0.43
Panel B: Alternative Bayesian Datasets									
	Single-unit administrative records			Big administrative records			Hybrid		
Shipments, mean	0.97	0.95	1	0.97	0.94	1	0.97	0.94	0.97
Shipments, sd	0.63	0.49	0.99	0.63	0.5	1	0.62	0.47	0.49
Shipments, skew	0.45	0.47	1.01	0.43	0.5	1	0.43	0.33	0.23
Labor Productivity, mean	0.85	0.84	0.89	0.85	0.83	0.89	0.85	0.84	0.92
Labor Productivity, sd	0.31	0.22	0.31	0.32	0.22	0.3	0.33	0.23	0.9
Labor Productivity, skew	0.29	0.17	0.14	0.34	0.2	0.14	0.33	0.24	1.81
R^2, Shipments on Employment	1.27	1.24	1.17	1.27	1.25	1.17	1.26	1.22	1.15
R^2, Admin Shipments	0.86	0.88	0.89	0.86	0.88	0.89	0.86	0.88	0.89
R^2, Admin Labor Productivity	0.49	0.57	0.58	0.48	0.55	0.59	0.48	0.53	0.57

Notes: This table compares our data to the administrative data from the Business Register for just single-unit plants. Every value compares the respective value in the relevant dataset divided by the corresponding value in the Business Register (other than the last two rows of each panel). In addition to the baseline Bayesian-edited data (which uses captured employment, materials, payroll, and shipments), we created three alternative datasets, which also use administrative information on payroll and shipments. For the "single-unit administrative records" version of the data, we add in administrative records only for the single unit plants. The ``big administrative records" version of the data uses administrative records for the entire manufacturing sector, as described in the text. The "hybrid" data supplements the ``big administrative records" version of the data with the analyst edited values when possible. The sample is balanced across all cells. The first six rows look at the mean, standard deviation, and skew for (ln) shipments and (ln) labor productivity (not normalized by sector). We then look within each dataset at the within R^2 of a regression of shipments on employment, with 6-digit-NAICS fixed effects, again relative to the value in the Business Register). The second-to-last rows in each panel shows the within R^2 of a regression of shipments in the administrative data on shipments in the relevant dataset, with 6-digit-NAICS fixed effects. The final row does the same for labor productivity.

Appendix Table 25. Sample Sizes

	No Trimming	1 Percent Trimming	2 Percent Trimming
	(1)	(2)	(3)
Panel A. Captured Data			
2002	105,000	102,000	98,000
2007	108,000	105,000	101,000
2012	79,500	77,000	74,500
Panel B. Final Data			
2002	196,000	190,000	183,000
2007	194,000	187,000	180,000
2012	175,000	169,000	163,000
Panel C. Bayesian-edited Data			
2002	151,000	146,000	141,000
2007	154,000	148,000	143,000
2012	131,000	127,000	123,000

This table presents sample sizes for the results on allocative efficiency, by dataset and trimming. Values are rounded following U.S. Census disclosure rules.

Appendix Table 26. Measured Allocative Efficiency, Bayesian vs. Captured in the Indian Annual Survey of Industries

	Captured Data / Bayesian Data		
	No Trimming	1 Percent Trimming	2 Percent Trimming
	(1)	(2)	(3)
Panel A. Gross Output			
2002	0.00005	0.00002	0.00004
2010	0.00002	0.00001	0.00001
Panel B. Value Added			
2002	0.00005	0.00005	0.007
2010	0.00006	0.00008	0.011

Notes: The values for calculating allocative efficiency are as described in the text. For the trimming, we drop the extremes for TFPQ and TFPR. The values are balanced within each dataset (it must be possible to calculate productivity), but unbalanced across. The reported values show the ratio of the value in the captured data over the value in the Bayesian-edited data. The underlying samples are plants in the Indian Annual Survey of Industries with no missing values.