Measuring Cross-Country Differences in Misallocation

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

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

November 29, 2019

Abstract

We describe differences between the commonly used version of the U.S. Census of Manufactures available at the RDCs and what establishments themselves report. We find substantially more extreme values of productivity (both measured TFPQ and TFPR) in the originally reported data. Furthermore, there is substantially more dispersion of commonly used measures such as the standard deviation or the interquartile range. To capture covariance, we follow the methodology of Hsieh and Klenow (2009). Measured allocative efficiency is subsantially higher in the cleaned data than the raw data - 4x higher in 2002, 20x in 2007, and 80x in 2012. Many of the important editing strategies at the Census, including industry analysts' manual edits and edits using tax records, are infeasible in non-U.S. datasets. We reanalyze cross-country comparisons starting from unprocessed U.S. and Indian data and using common data cleaning strategies: a simple trimming-outliers approach and a new Bayesian approach for editing and imputation. Under both methods there is little evidence that allocative efficiency is significantly worse for formal firms in India than in the United States. If anything, we find the opposite.

Some of the research in this paper was conducted while the second author was an employee of 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 number DRB-B0115-CDAR-20181016 and DRB Bypass number CBDRB-FY19-CMS-7909. 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, for Hang Kim for his helpful comments, to Allan Collard-Wexler for a thougtful discussion, and to Lee Tanenbaum and Weitao Lin for their expert research assistance.

I Introduction

In the past twenty years, many economists (including us) have written papers and lecture notes highlighting that 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 firm 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 firm behavior - either the "true" values for a particular country, or of cross-country differences - depends on how worried we are about measurement error. In principle, measurement error has an ambiguous effect on measured dispersion, since non-classical noise can either push firms towards or away from what is typical in their sector.

Firms that report inaccurate information may only spuriously appear to be using a socially inefficient quantity of resources. The converse may be true as well: firms may report values which seem optimal but do not reflect reality on the ground. As a result, the confidence we have in our measures of misallocation - either measurements of - depends on the extent of measurement error. In this paper, we discuss two potential sources of measurement error: firms potentially misreporting their own characteristics, and subsequent data processing potentially introducing new errors.

Most statistics agencies initially ask firms 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.[1] Over half of estab-

[1] Less developed countries are less likely to directly edit the raw data. For instance, for India and China we have confirmed that there was no editing or imputation of the data (both in the data documentation, and in email communications with the relevant national statistics agencies). There is some imputation of prices in the Indian data post-2006, but we do not use the price information in our results.

lishments have at least one characteristic which is affected by the cleaning process[2]. The exact procedures vary across industries and time (White et al., 2018), 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 in previous years) and other plants in the same industry.[3] For 2002, 2007, and 2012 we have access to the original and cleaned values reported by firms for plants in the Census of Manufactures,[4] as well as the relevant edit flags.

We have three mains goals in this paper. Our first goal is to describe the extent of data cleaning efforts undertaken by the U.S. Census Bureau. Both editing clear reporting errors (such as when firms report distinct values for the same outcome), as well as more subjective edits (such as manual edits by industry experts) have large effects on measured dispersion. We compare the original responses to the cleaned responses in a variety of way, focusing on two main outcomes: measured TFPR and TFPQ.[5] First, we show visually that cleaning has large effects on the distributions, dramatically lowering the mass in the tails while also shifting the distributions up for TFPR.

Most quantitative measures of misallocation use statistics that are broadly similar to the standard deviation in revenue productivity (TFPR) (Asker et al., 2019). These types

[2] Throughout the paper, we set aside edits to capital given well-known conceptual difficulties with measuring it correctly (Hicks, 1981; Hulten, 1991; Pritchett, 2000; Collard-Wexler and De Loecker, 2016).

[3] Firms that have a variable edited have that variable imputed as if the firm had not reported anything for that variable. Note that the imputed data must also pass the editing rules. For most establishments, at least one of the variables needed to calculate TFP is imputed (White et al., 2018). For payroll and number of employees, the Census Bureau uses administrative records (mainly IRS payroll data) to replace reported data that fails edit rules. The Census Bureau classifies these changes from the reported data as "non-imputes". However, these non-imputes still change plants' measured TFP.

[4] It is worth noting that when Hsieh and Klenow (2009) was written, neither imputation flags nor the original firm responses were available for the Census of Manufacturers.

[5] There are well-known difficulties in measuring TFPQ and TFPR, both because estimating production functions is difficult even with data on quantities, and because in many datasets the only measure of output is revenues. We use simple methods to estimate TFPQ and TFPR. In ongoing work we are collecting longer panel data, which will allow us to say more about measuring production function elasticities.

of measures are 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 also lowers the measured interquartile range of revenue productivity by almost as much. Both the effect of data processing and measured dispersion (in the raw & cleaned data) are increasing over time.

The data cleaning matters most for smaller and younger plants: 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 TFPR even for the largest and oldest plants in the data.

In many models (e.g. Hsieh and Klenow 2009) the covariance of TFPR and TFPQ is important for measuring misallocation, with the logic that distortions matter more for aggregate activity if they hit firms who would otherwise be large. If instead of using the U.S. Census Bureau's cleaned data we simply trim the 1% extremes in the data originally reported by each establishment, we find that moving to the new measured U.S. efficiency would *decrease* measured (Hsieh and Klenow, 2009) TFP by a factor of 8 for the Indian formal sector. We do not take this result literally - we do not think that we have compelling evidence that the U.S. manufacturing sector is characterized by more misallocation than India. Instead, we consider our results a "smoking gun" that measurement (and data processing in particular) is deeply important to the study of misallocation.

Given that the overall cleaning effort has a large effect on measured misallocation, we turn to describing the specific process of cleaning the data. After qualitatively describing the different sets of edit rules, we assess their effects by measuring how much each edit rule affects measured dispersion. We do so in two ways: first by describing how much measured dispersion changes when we *only* use each particular edit rule, and second by each edit's Shapley (1953) value.

The most important edits are "logical" edits and analyst corrections. Logical edits are possible because the Census implicitly asks for the same information in multiple ways, for instance by asking for the plant's total value of shipments as well as the total value of shipments for each product. If the two values for the same outcome diverge, the Census may edit the reported total with the sum of the disaggregated components. Analyst corrections rely on the expertise of full-time industry specialists employed by the U.S. Census Bureau.

In many international datasets with microdata on firms, many types of processing done by the U.S. Census Bureau are infeasible.[6] For instance, logical edits can only be implemented when there are redundant questions. Given the importance of these edits, it is difficult to interpret measured cross-country differences that use different processing methods. Since we are not nihilists, we describe methods for commonly cleaning data across contexts. Trimming is a popular data cleaning strategy which can easily be used across contexts. However, trimming is a blunt instrument, and varies paper to paper: for instance while Hsieh and Klenow (2009) trim static measured distortions, other papers using the exact same data have trimmed other characteristics such as growth rates (Allcott et al., 2016). We describe and implement a more reproducible approach: a theoretically-motivated data cleaning exercise which could then be used across firm-level datasets without further need for data processing (Kim et al., 2015).

Unlike trimming, which drops outliers and leaves missing values blank, the Kim et al. (2015) method simultaneously edits and imputes 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). We then impute entries in order for the cleaned data to pass the edit checks. Unlike the imputation methods the Census

[6] There are also differences in enumeration, for instance in the United States the Census is a web portal whereas few Indian manufacturing plants report having computers.

Bureau uses most frequently in the Census of Manufactures, the Kim et al. (2015) method tries to preserve the joint distribution of the covariates.[7] Given the ratio outlier flags, we favor making edits that are likely given our 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.

Economists have a long tradition of studying (mis)measurement. Our paper is firmly in the spirit of (Romer, 1986a,b, 1989), who study how differences in data quality matter for understanding the historical incidence of business cycles and unemployment. In the misallocation (Banerjee and Duflo, 2005; Restuccia and Rogerson, 2008; Hopenhayn, 2014) literature in particular, Bils et al. (2017); Gollin and Udry (2019) and Esfahani (2019) all 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. One value of our approach is that we leverage expert editing done by those with their own goals for creating reliable data. Furthermore, researchers can use our proposed data cleaning procedure regardless of their question, and even in settings with less rich data collection and cleaning efforts than the US.

In the next section, we recap the theory of distortions underlying much of our analysis. Section 3 discusses data collection and cleaning procedures for manufacturing data in the United States. Section 4 describes the Kim et al. (2015) method for cleaning plant-level data. In Section 5, we compare commonly-cleaned data for the U.S. and India. Section 6 concludes.

[7] The Census methods of imputation of missing data also lower measured misallocation relative to more flexible approaches (White et al., 2018).

II A Theory of Misallocation

In order to keep our results parsimonious, we focus on describing how TFPR and TFPQ are affected by data processing. Since we present aggregate statistics for the whole US economy, we follow methods in Bils et al. (2017) and Hsieh and Klenow (2009) for normalizing plant values. We briefly describe their approach. First, we start from the firm's problem, showing how firm behavior is affected by idiosyncratic distortions on capital and output.[8] In the model, variation in those distortions is captured by variation in firm-level revenue productivity. We then turn to the (static) equilibrium, and derive how aggregate productivity would be affected in a counterfactual where the variation in revenue productivity is removed, which is our measure of misallocation. For our results, we show effects mostly for gross-output (which is a function of capital, labor, and materials). For some results, we also show value added (produced using only capital and labor) specifications, when the differences are potentially informative. We derive only the gross-output case; the value added version is almost identical.

II.A Firm-level Distortions

Final good production Q is a Cobb-Douglas aggregate over sectoral gross output Q_s,

$$Q = \prod Q_s^{\theta_s},$$

so normalizing the price of the final good, P, to 1, expenditure for each sector is a fixed proportion

$$P_s Q_s = \theta_s Q$$

[8] For our descriptions of the theory, we refer to firms. For single-plant firms, this is equivalent to the plant. Following Bils et al. (2017) and Hsieh and Klenow (2009), all of our measurements of TFPR or TFPQ are at the plant-level, even for multi-plant firms.

where P_s is the price index for sector s. The final good can either be consumed or used as an intermediate input: $Q = C + M$.

All firms use the same intermediate input, denoted M_{si}, so $M = \sum_{s=1}^{S} \sum_{i=1}^{N_s} M_{si}$. Within each sector, output takes a CES form over output of each variety Q_{si}:

$$Q_s = \left(\sum_{i=1}^{N_s} Q_{si}^{\frac{\sigma-1}{\sigma}} \right)^{\frac{\sigma}{\sigma-1}}$$

and each firms produces gross output using capital, labor, and the intermediate input with Cobb-Douglas production-function elasticities which vary across sectors:

$$Q_{si} = A_{si} (K_{si}^{\alpha_s} L_{si}^{1-\alpha_s})^{\gamma_s} M_{si}^{1-\gamma_s}.$$

The wage and rental rate are constant in the economy, but firms face idiosyncratic distortions on labor, capital, and intermediate inputs. As a result, each firm's profits are:

$$\pi_{si} = P_{si} Q_{si} - \left(1 + \tau_{L_{si}}\right) w L_{si} - \left(1 + \tau_{K_{si}}\right) R K_{si} - (1 + \tau_{M_{si}}) P M_{si}.$$

II.A.1 Optimization Behavior

Firms are monopolistically competitive and face downward-sloping demand curves given by $Q_{si} = Q_s \left(\frac{P_{si}}{P_s}\right)^{-\sigma}$. Profit maximization implies that the firm's output price is a fixed markup over its marginal cost and so there is complete pass-through of improvements in TFPQ (A_{si}). As a result, revenue productivity, $TFPR_{si} = \frac{P_{si} Q_{si}}{\left(K_{si}^{\alpha_s} L_{si}^{1-\alpha_s}\right)^{\gamma_s} M_{si}^{1-\gamma_s}}$, only varies due to the distortions. In particular, Hsieh and Klenow (2009) demonstrate that

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

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

$$TFPQ_{si} \equiv A_{si} \propto \frac{\left(P_{si}Q_{si}\right)^{\frac{\sigma}{\sigma-1}}}{\left(K_{si}^{\alpha_s} L_{si}^{1-\alpha_s}\right)^{\gamma_s} M^{1-\gamma_s}}$$

In the next subsection, we show how variation in TFPR affects aggregate productivity.

II.B Aggregate Distortions

Aggregate productivity in each sector is

$$TFPQ_s = \frac{Q_s}{(K_s^{\alpha_s} L_s^{1-\alpha_s})^{\gamma_s} M_s^{1-\gamma_s}} \equiv \frac{\overline{TFPR}_s}{P_s}, \tag{2}$$

where, given cost-minimization, the price index for sector s is:

$$P_s = \left(\sum_{i=1}^{M} P_{si}^{1-\sigma}\right)^{\frac{1}{1-\sigma}} \left(\sum_{i=1}^{M} \left(\frac{A_{si}}{TFPR_{si}}\right)^{\sigma-1}\right)^{\frac{1}{1-\sigma}} .$$

Plugging back in to Equation 2 gives the core Hsieh and Klenow (2009) expression for productivity

$$TFP_s = \left(\sum_{i=1}^{M} \left(A_{si}\widetilde{TFPR}_{si}\right)^{\sigma-1}\right)^{\frac{1}{\sigma-1}}, \tag{3}$$

where $\widetilde{TFPR}_{si} = \frac{\overline{TFPR}_s}{TFPR_{si}}$. Since we know from Equation 1 that $TFPR_{si}$ would only be different from $\overline{TFPR}_s$ in the presence of distortions, the "efficient" counterfactual TFP is $\overline{A}_s = \left(\sum_{i=1}^{M} A_{si}^{\sigma-1}\right)^{\frac{1}{1-\sigma}}$. Aggregating over all sectors,

$$\widetilde{TFP} = \frac{TFP}{TFP_{(efficient)}} = \prod_{s=1}^{S} \left[\sum_{i=1}^{M_s} \left(\widetilde{TFPQ}_{si} \times \widetilde{TFPR}_{si}\right)^{\sigma-1}\right]^{\frac{\theta_s}{\sigma-1}}, \tag{4}$$

where $\widetilde{TFPQ}_{si} = \frac{A_{si}}{\overline{A}_s}$. The index in Equation 4 ranges from 0 to 1 and can be calculated from observed data. The three main outcomes that we describe in the paper are $\widetilde{TFPQ}_{si}$, $\widetilde{TFPR}_{si}$, and $\widetilde{TFP}$.

Instead of measuring how sensitive our three measures of productivity are to different underlying assumptions, which has been the primary focus of much of the recent methodological literature on misallocation (Haltiwanger et al., 2018; Asker et al., 2014), we instead calculate them using different cuts of the data, which we describe in the next section.

We also 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 firm age and the (log) number of employees, also at the firm level. For both, we run a local polynomial regression comparing the difference (either regular difference or the absolute difference) between measured and captured $\widetilde{TFPR}_{si}$ and $\widetilde{TFPQ}_{si}$.

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. The quinquennial survey covers roughly 300,000 manufacturing plants.[9] As in most surveys, not all respondents answer all of the questions, and some responses are inconsistent with each other or inconsistent with administrative records data (primarily IRS 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.[10] We go beyond the imputation flags and use the newly available actual responses from the

[9] Information for plants with fewer than 5 employees - roughly one third of the sample - are almost entirely imputed. The standard is to exclude these so-called administrative records plants (Foster et al., 2016), and we follow that standard throughout our analysis.

[10] Item-level edit/impute flags for the 2002 and later censuses became available to researchers a few years ago, and item-level flags for the 1987, 1992, and 1997 censuses became available to researchers in the Federal Statistical Research Data Centers (FSRDCs) in November 2018.

establishments themselves (the "raw" data).

The raw data differs from the final ("cleaned") data in two respects.[11] First, missing values due to non-response in the reported data are imputed in the cleaned data, using a variety of industry-specific regression-based and other imputation strategies. Second, actual responses which fail edit rules in the reported data are also imputed or changed in some way in the final data. The most important edit rules are balance edit rules (certain variables have to add up) and ratio edit rules (ratios of certain variables must be within certain bounds). Edit-rule-failing responses are replaced using a variety of methods, described in Table 1, and the ones that affect measured misallocation the most are described more fully in Subsection IV.C. In the next subsections, we compare two cuts of the data: one where we take all of the edits, and the other where we reject all of them. First we show the differences in the distribution of productivity, then in measures of dispersion, and finally on measures of misallocation.

III.A The distribution of TFPR

We start by describing how data processing affects the measured density of TFPR. We do this in the following way. First, we create bins for each 2% of the distribution of TFPR in the final data. Then, in each of these bins, we plot the difference in density between the raw and clean data (so positive values means relatively more density in the bin in the raw data).[12] The differences-in-densities are plotted in Figure 1. In all three years, there is more mass in the raw data in the tails, which is presumably consistent with most forms of data cleaning that try to address outliers. However, at the lower end this is true only for the lowest (two) bins. All three years exhibit a check-mark shape, where the lower-

[11] Another convention for naming the data is "captured" data for the reported information, and "completed" for the cleaned data.

[12] In order to limit the disclosure of the data, we slightly jitter the exact end points of each bin, so readers cannot look with a magnifying class to identify the TFPR of, e.g. the 22nd percentile firm. Similarly, we move the density in the top and bottom bins into their respective neighbors, so we are only plotting 48 bins.

middle of the distribution is over-represented in the clean data, and especially the top of the distribution is over-represented in the raw data.

IV Who gets edited?

In this section, we describe which types of firms see large differences between their captured and final data, focusing on age and employment (we use measures of firm characteristics, the patterns are similar using establishment characteristics). In Figure 2, we plot the relationship between the absolute difference in measured TFPR (using gross output production functions at 1% trimming, and scaling using industry means) between the captured and final data. The pattern is similar for the three years for which we have data; the absolute difference is consistently falling both in firm age and firm size. The average absolute difference for the largest firms is still not trivial - and is often around 50% - but is larger for the younger and smaller firms

The absolute gap could be positive even if the cleaning were mean zero. Figure 3 shows that the edited data consistently has smaller measured TFPR than the captured data. Again, the size of the average gap is decreasing in firm size and age, although the pattern for firm age is somewhat flatter.

IV.A Dispersion of TFPR

In this section, we quantify the results from the previous subsections. Specifically, we describe the effects of data cleaning on alternative measures of dispersion, specifically the 90/10 ratio, the interquartile range, and the standard deviation. For simplicity, we report values with no trimming. The first column in Table 2 reports the standard deviation. Across all of the years, and for both gross output and value added production specifications, the standard deviation in the cleaned data is around half of in the original data. Measured dispersion increases over time, although by relatively more in the captured data: in the captured data, the standard deviation increased by over 20% from 2002 to 2012. In the final data, it increased by under 10%.

We see a similar pattern for the 90/10 ratio and the interquartile range. The captured data has 2-3 times more spread than the final data does. Even the interquartile range is dramatically affected by the data cleaning undertaken by the U.S. Census: the gross output range in 2012 fell by almost a factor of three. Similar to the standard deviation, there have been a larger increases over time in the captured data than in the final data. However, the ratio of the 90/10 to the interquartile range - the (90/10) / (75/25) ratio of ratios - is slightly larger in the final data than the captured data.

Taken together, these results imply that the cleaning undertaken by the U.S. census is more nuanced than trimming, since it dramatically effects ratios of interior quantiles.

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

For our final set of results comparing the cleaned and captured data, we use the Hsieh and Klenow (2009) model described in Section 2. While further removed from the raw data, 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 firms' firm size ranking (Hopenhayn, 2014).

First, we consider the effects on measured misallocation of replacing cleaned with raw data in the U.S. manufacturing sector in 2002, 2007 and 2012. Table 3 shows the results of calculations across a range of similar datasets.[13] In the raw data with no trimming, allocative efficiency is nearly 0 and gets worse (lower) in 2007 and 2012. Trimming the 1% tails raises measured allocative efficiency to 0.109 in 2002, implying that the US manufacturing sector was about 11% as productive as it would be if there were no misallocation. With this level of trimming, allocative efficiency still falls to 0.012 in 2007 and 0.004 in 2012. We see a similar pattern with 2% trimming except that allocative efficiency levels off in

[13] For all the results reported in the paper we set the elasticity of substitution $\sigma = 3$ when considering value added production functions, and $\sigma = 4$ for gross output (Hsieh and Klenow, 2009; Bils et al., 2017).

2012. When we use the Census-cleaned data, we find a similar pattern over time for all 3 levels of trimming–allocative efficiency falls from 2002 to 2007 and levels off or increases slightly from 2007 to 2012. The data cleaning done by the Census Bureau has an enormous effect on measured allocative efficiency. With 1% trimming, measured allocative efficiency ranges from 4 to 87 times higher in the cleaned data versus the raw data.

The Census of Manufactures is only undertaken quinquenially. In other years the detailed manufacturing data only covers a sample of plants.[14] In Table 4, we report the measured misallocation numbers from the Annual Survey of Manufactures sample within the census year, where each plant is weighted by its ASM sample weight.[15] Both for value added and for gross output production functions, the switch from census to sample has little effect on measured dispersion in the Census-Cleaned data. However, in the raw data there tends to be somewhat less measured misallocation in the sample relative to the census.

IV.B.1 Measured Cross-Country Differences in Misallocation

We now turn to discussing cross-country differences in measured misallocation. While we have measured misallocation in the United Sates for a large set of data choices, in order to avoid tedium we only describe cross-country differences in misallocation for two extremes: the average of 2002, 2007, and 2012 for Census-cleaned data, and the corresponding average in the Census raw data.[16] We compare these averages to (Gross Output) estimates from India's Annual Survey of Industries in 2002 and 2007.[17] The results are shown in Table 5. While estimated allocative efficiency in India in 2002 and 2007 is

[14] 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.

[15] Due to small differences between samples, we could not disclose the 2012 value added number with 2% trimming.

[16] For all values, when possible, we use the reported values from trimming the 1% extremes for TFPQ and the distortions for capital, labor and intermediate inputs.

[17] The ASI is described in Appendix A.I.

slightly higher than the average for the cleaned U.S. data, it is over 8 times higher than average measured allocative efficiency in raw U.S. data. Taken literally—which we do not—the latter result would mean that the Indian manufacturing sector would be only an eighth as productive as it is if it had the same allocative efficiency as the U.S. manufacturing sector.

IV.C The effect of edits on measured misallocation in the raw U.S. data

There are many different data cleaning steps on the road from measured allocative efficiency of 0.004 in the 2012 raw data to 0.349 in the corresponding cleaned data. In order to determine if changes to the raw data are needed, the Census Bureau primarily uses balance edit rules and ratio edit rules. An example of a balance edit rule is that the number of production workers plus the number of non-production workers must equal the total number of employees. For ratio edit rules, the Bureau uses ratios of reported values (for instance, one of the ratios is the total value of shipments over annual payroll). The Census determines industry-specific upper and lower bounds (either by looking at the percentiles of reported outcomes or by relying on additional industry-specific knowledge). If a plant's reported values violate one or more of the balance or ratio edit rules, one or more of the reported values is replaced using one of the types of edits or imputations described in Table 1.

In this subsection, we characterize the effect of each edit. For eight edit supercategories, we measure misallocation in the U.S. using raw data for everything but those edits (and using the cleaned data for the plants affected by the given edit). The eight supercategories are: logical edits (separately for shipments, materials, and payroll), administrative edits, regression imputes, rounding imputes, analyst corrections, and the rest. We describe each supercategory in turn.

Logical edits are done when there are many survey questions which ask for the same information. For instance, total value of shipments shows up in three different parts of the survey: (1) there is a question that asks for the total value of the plants shipments; (2) there are many questions about the value of shipments for specific products that a given industry produces (these values can be summed by the U.S. Census Bureau); and (3) there is a question – separate from (1) – that asks the respondent to total the values of the products in (2). If these values differ by more than a certain amount (the same tolerance is used for all industries within a year, but has varied over time), then the Census compares each of them to annual payroll for the same plant and then takes the "best" one. The "best" one is selected in a form similar to the regression imputes, described below.

Administrative edits are similar to logical edits, but differ in that the alternative source of information is from administrative records. For instance, for payroll the administrative records come from IRS payroll tax records. Again, if the administrative data differ from the reported values, the reported values may be replaced.

Regression imputes 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 also used to impute missing values when no other information is available for a given variable. In general, regression imputes create predictions using one other variable, and (for plants surveyed in the Annual Survey of Manufactures), one-year lags of the imputed variable as well. Unlike administrative and logical edits, there is not necessarily any direct evidence that the reported firm value may be incorrect.[18]

In order to measure the importance of each edit, we undertake the following exercise. For each type of edit, we replace *all* of an establishment's information with the clean data if it was affected by the edit. For example, to understand the importance of logical edits

[18] There may be strong *indirect* evidence that a reported value is incorrect, such as, hypothetically, if a plant reported production worker wages of $5 billion and only 1 production worker.

for total value of shipments, we use the cleaned outcomes for plants whose total value of shipments has a logical edit flag, and the raw outcomes for the other plants. We show how much measured misallocation in the U.S. is affected by each edit by showing (a) the change in misallocation for each edit if it is the only one applied (this always decreases measured misallocation), and (b) its Shapley value. For this context, the Shapley value is the value of the following thought exercise: we first consider all possible combinations of edits, done one at a time. We credit each edit for its marginal contribution within each (ordered) combination, and calculate the Shapley Value as the average of those marginal contributions. The results are in Table 6, with the third column reporting the share of the total decrease from all the edits that the Shapley procedure credits to each edit type.

The most important edits are logical imputes for total value of shipments, which are responsible for a fifth of the decrease in measured misallocation, as well as imputes for the missing values, which are collectively responsible for another fifth. Analyst corrections (for any TFPR variable) and logical imputes for payroll are also each responsible for over 10% of the decline.

The results of Table 5 are unsatisfying - cross-country comparisons of measured misallocation in datasets which have been cleaned differently will pick up differences due to both underlying cross-country differences and cross-country differences in data cleaning. However, while comparing raw data solves the latter problem, it does so at the expense of introducing new errors. The natural solution is to compare datasets which have been commonly cleaned. While one strategy may be to use the approach of the U.S. Census Bureau everywhere, Table 6 shows that around $^{2}/_{3}$ of the changes - the logical imputes, analyst corrections, and administrative record edits - are difficult if not impossible to replicate in other contexts (depending on the availability of alternative reports for the same outcome and industry specialists). As an alternative, in section V we describe and then implement an algorithm for editing and imputing raw firm-level (or plant-level) data.

V An approach to Cleaning Firm-level Data

Establishment i reports p characteristics, $\boldsymbol{y}_i = \left\{y_{i1}, y_{i2} \ldots y_{ip}\right\}$ (where items could be missing). The corresponding true values are $\boldsymbol{x}_i = \left\{x_{i1}, x_{i2} \ldots x_{ip}\right\}$, with s_{ij} indicating 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 is to replace the faulty values so that (i) the cleaned data for each record i is plausible and internally consistent and (ii) the cleaned dataset is drawn from the same joint distribution as the true values $\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots \boldsymbol{x}_n\}$. In order to do this, we use the approach of Kim et al. (2015), which we now describe.

First, we define the feasible region $\mathcal{D}$ of plausible reports. This region limits possible values by two types of rules: balance rules which require entries to add up[19] and a set of ratio edit rules which bound the ratios of any two variables. While the balance rules are *a priori* clear, 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.[20] While $\boldsymbol{s_i}$ is not directly observed, A_i indexes the failed ratio & 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}.

After cleaning the data, we want our 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,

$$f\left(\boldsymbol{x_i}, \boldsymbol{s_i} | \boldsymbol{y_i}, A_i\right) \propto f\left(\boldsymbol{y_i} | \boldsymbol{x_i}, \boldsymbol{s_i}, A_i\right) f\left(\boldsymbol{s_i}, A_i | \boldsymbol{x_i}\right) f\left(\boldsymbol{x_i}\right). \tag{5}$$

For the model of reporting error, we maintain the U.S. Census Bureau's (implicit) ap-

[19] For instance, non-production wages + production wages = total 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.

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

proach to ratio and balance edits: data reported with error provides no information on the true value.[21] Therefore, $f\left(\boldsymbol{y_i}|\boldsymbol{x_i},\boldsymbol{s_i},A_i\right)$ is uniform over the support of feasible values if $y_{ij} \neq x_{ij}$.

However, unlike the Census Bureau, we also assume a uniform distribution for the errors. That is to say, we do not use weights on which variables are more likely to be reported with error, so all candidates $\boldsymbol{s_i}$ that result in feasible solutions are 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,[22] we need to estimate the probability of membership in each component (π), 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 both the balance and ratio edits, we impose that the distribution of $\boldsymbol{x_i}$ conditional on μ, Σ, z_i, given feasible region $\mathcal{D}$ is

$$f\left(\boldsymbol{x}_i|\theta_i\right) = \mathcal{N}\left(\boldsymbol{x}_{i,NT}|\boldsymbol{\mu}_{z_i},\boldsymbol{\Sigma}_{z_i}\right)\prod_{l=1}^{q}\delta\left(x_{iT_\ell}-\sum_{j\in\beta_\ell}x_{ij}\right)\mathbb{1}\left[\boldsymbol{x}_i\in\mathcal{D}\right]$$

where $\delta\left(\cdot\right)$ is that Dirac delta function with the point mass at zero and $\boldsymbol{x}_{i,NT}$ is the set of reported values which are themselves totals of other reported values.

For each 6-digit NAICS industry-year, we run a single chain of Markov Chain Monte Carlo with a burn-in of 2000 iterations, and then 5000 additional iterations, keep the data from each 1000th iteration (for a total of 5 completed datasets for each industry-year in each country). Each iteration consists of first proposing $\boldsymbol{s_i}$ which are consistent with $\boldsymbol{A_i}$,

[21] One important exception to this rule is units errors (a.k.a. “rounded” edits), where the original reported value is divided by 1000 and then rounded to the nearest unit. These are cases where Census suspects the respondent reported in dollars instead of thousands of dollars (as is requested). These cases are identified by comparing the plant’s ratio of a dollar-valued variable to its employment. For example, if a plant hypothetically reports annual payroll of \$1 billion and employment of 50, then its payroll/employment ratio of \$20 million per employee is implausible for any manufacturing plant. However, annual payroll of \$1 million for the same plant implies a ratio of \$20,000 per employee, which might fall within the ratio bounds for some industries. In this case the reported value for annual payroll would be replaced with \$1 million.

[22] In practice we set K=50, which is large enough that no data are in the lowest probability components.

and then editing values $\boldsymbol{y_i}$ given the draw of $\boldsymbol{s_i}$ and the underlying probability distributions for the responses which were not reported with error.

VI Cross Country Differences in Measured Misallocation

The main choice associated with the model in Section V is defining the feasible region $\mathcal{D}$. For the U.S. data, we use the industry-year-specific ratio edit rules (upper and lower bounds) used by the Census Bureau for the 2002, 2007, and 2012 Censuses of Manufactures.

For the Indian data, 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, for each log ratio $r_{jk} = \ln\left(\frac{y_j}{y_k}\right)$, we calculate its 25th and 75th percentiles, Q^{25}_{jk} and Q^{75}_{jk}, and therefore the interquartile range IQR_{jk} . We then flag all ratios that are either smaller than $Q^{25}_{jk} - C \times IQR_{jk}$ or larger than $Q^{75}_{jk} + C \times IQR_{jk}$ where C is a pre-specified threshold. The variables we use are the total cost of materials, total value of shipments, number of blue and white collar workers, blue and white collar wages and (in India) the sampling weight, and we run the estimation separately by industry (6 digit in the U.S. and 2 digit in India) and year. We calculate allocative efficiency separately for each year. In Table 7, we report the average of the 2000-2011 results for India with $C = 3$ and for 2002, 2007, and 2012 in the US. For comparison, we also report results from the Captured Data with 1% and 2% trimming.

The gap between the raw U.S. and raw Indian data shrinks after applying the common data editing procedure. In the commonly-cleaned data with 1% tail trimming, allocative efficiency in the U.S. in 2002 is slightly lower than the Indian time series average (0.499 vs 0.521). In 2007 and 2012, allocative efficiency in the US is, respectively, 0.161 and 0.231—significantly lower than the Indian average. With 2% trimming, allocative efficiency is slightly higher in 2002 in the US vs India, and about 30% lower in 2007 and

2012.

VII 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. We show that in the data that is reported directly to the Census Bureau by U.S. establishments, measured TFP in the United States manufacturing sector is less than an eighth of what it would be if it had the same level of allocative efficiency as the Indian manufacturing sector. We do not take this result literally: there are many reasons to believe that comparing the raw U.S. data to its counterparts in India or other countries is not like-for-like.[23] We also see large differences for other measures of dispersion, such as the interquartile range. While editing matters more for younger and smaller firms, even the oldest and largest ones have their responses relatively heavily edited.

Many of the important edits undertaken in the U.S. are infeasible for researchers using other datasets, because they either use multiple responses for the same information or because they rely on U.S. Census Bureau industry experts. When we use common data cleaning strategies, we find little or no evidence that allocative efficiency is significantly lower in India than in the United States, and for 2007 and 2012 we find that allocative efficiency is significantly *higher* in India than in the US.

There is a large scope for different measurement choices to affect the estimation of misallocation and dispersion in manufacturing, and researchers should not use ad-hoc approaches under the theory that the data cleaning doesn't matter. Our message is not

[23] For no country do we know if measured misallocation in the Captured Data is larger or smaller than it is in reality, nor do we have a way of comparing the relative precision of self-reported information across countries. We do know that the vast majority of Indian firms are unable to fill out their survey forms on a computer.

nihilistic: we see opportunity in working with experts in data cleaning (who currently associate with other disciplines). To that end, we suggest an alternative approach for cleaning firm-level data using a hierarchical Bayesian approach. In addition to generating reasonable answers when we comparing cross-country differences in outcomes, the method is more broadly useful for data cleaning instead of more traditional ad-hoc approaches such as winsorizing.

References

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.

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.

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

Bartelsman, E. J. and W. Gray (1996). The NBER Manufacturing Productivity Database.

Bils, M., P. J. Klenow, and C. Ruane (2017). Misallocation or mismeasurement? Technical report, Working Paper.

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

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

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

Grim, C. (2011). User notes for 2002 census of manufactures. *Unpublished Technical Note*.

Haltiwanger, J., R. Kulick, and C. Syverson (2018). Misallocation measures: The distortion that ate the residual. Technical report, National Bureau of Economic Research.

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

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. (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.

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.

Pritchett, L. (2000). The tyranny of concepts: Cudie (cumulated, depreciated, investment effort) is not capital. *Journal of Economic Growth 5*(4), 361–384.

Randy Becker , Wayne Gray, J. M. (2016). NBER-CES Manufacturing Industry Database: Technical Notes. *National Bureau of Economic Research Technical Working Paper Series*.

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

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

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.

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.

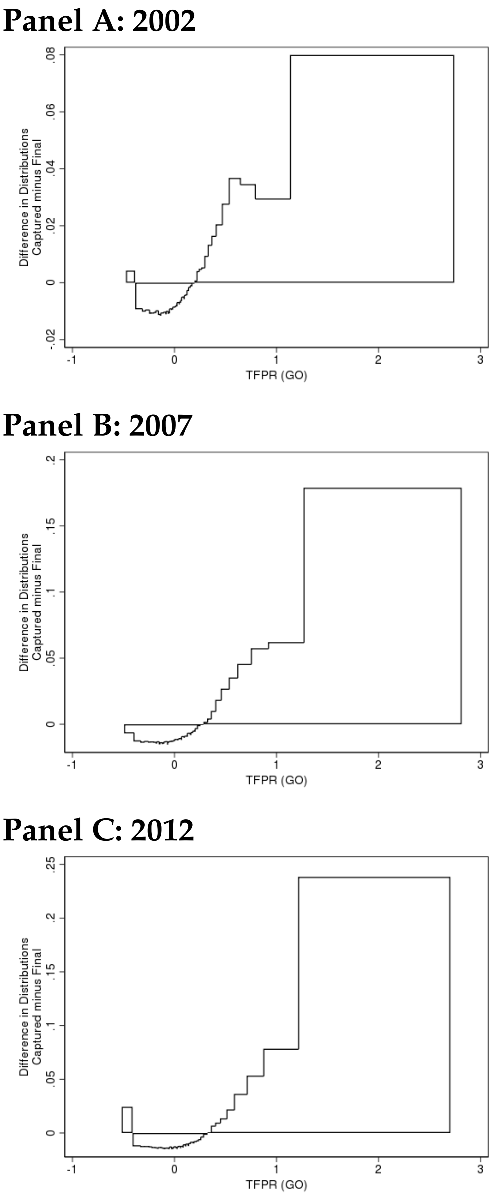
White, T. K., J. P. Reiter, and A. Petrin (2018). Imputation in U.S. Manufacturing Data and Its Implications for Productivity Dispersion. *The Review of Economics and Statistics 100*(3), 502–509.

A Appendix

A.I Cross-Country Estimates of Misallocation

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 firms without local competitors are more likely to be surveyed). Hsieh and Klenow (2009) use the same dataset, and we follow standard practice in generating measures of gross output, intermediate inputs, capital, and payroll. Industries are grouped using India's NIC (National Industrial Classification) codes, and we report the value of reallocation for 2009. For the U.S. and India, we use cost-shares from the NBER-CES Manufacturing Industry Database as our measures of industry production elasticities, and multiply the book value of capital by 10% in order to impute the cost of capital.

Figure 1: Relationship between firm size/age & editing

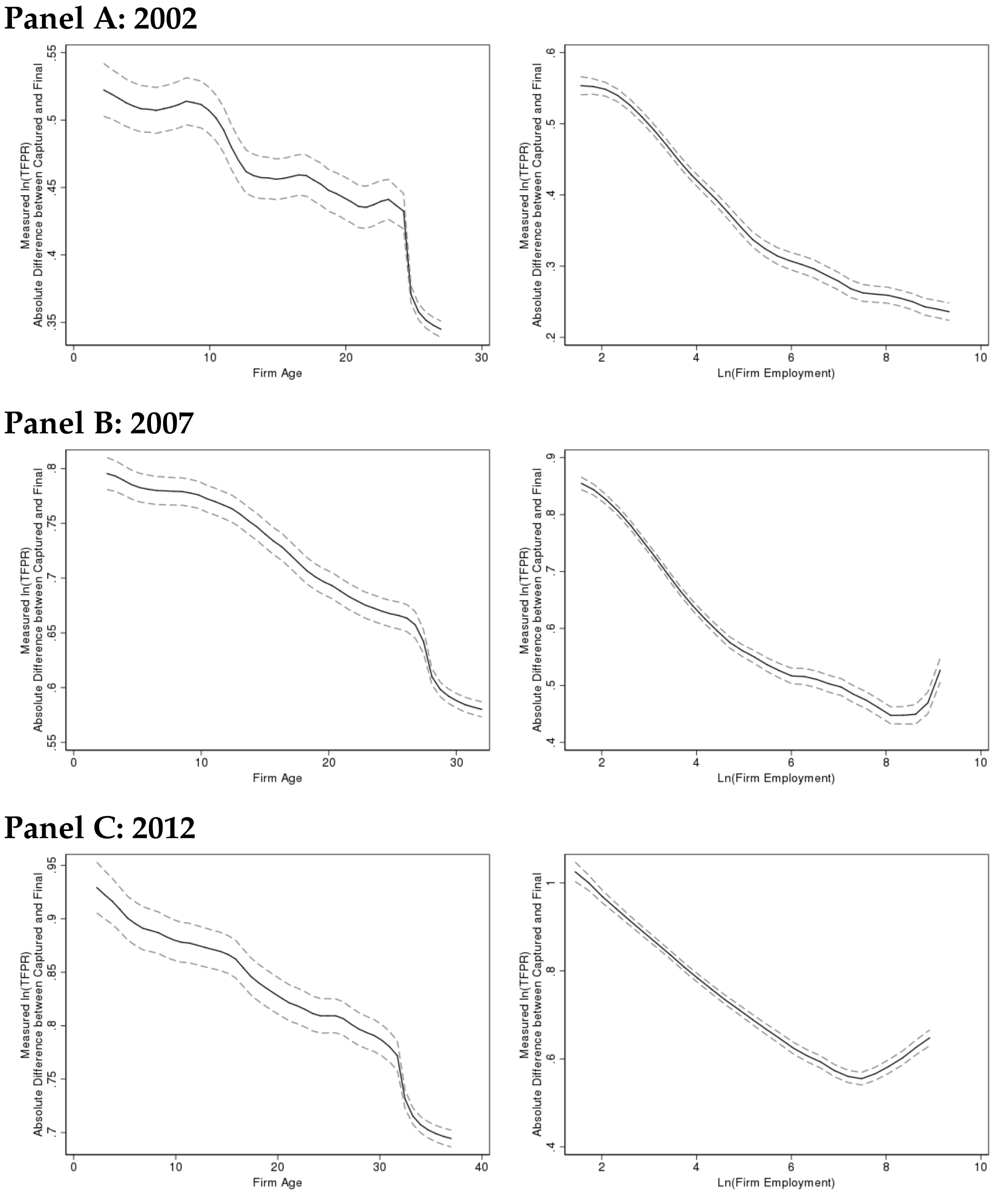


Panel A: 2002

Panel B: 2007

Panel C: 2012

Figure 2: Relationship between firm size/age & editing

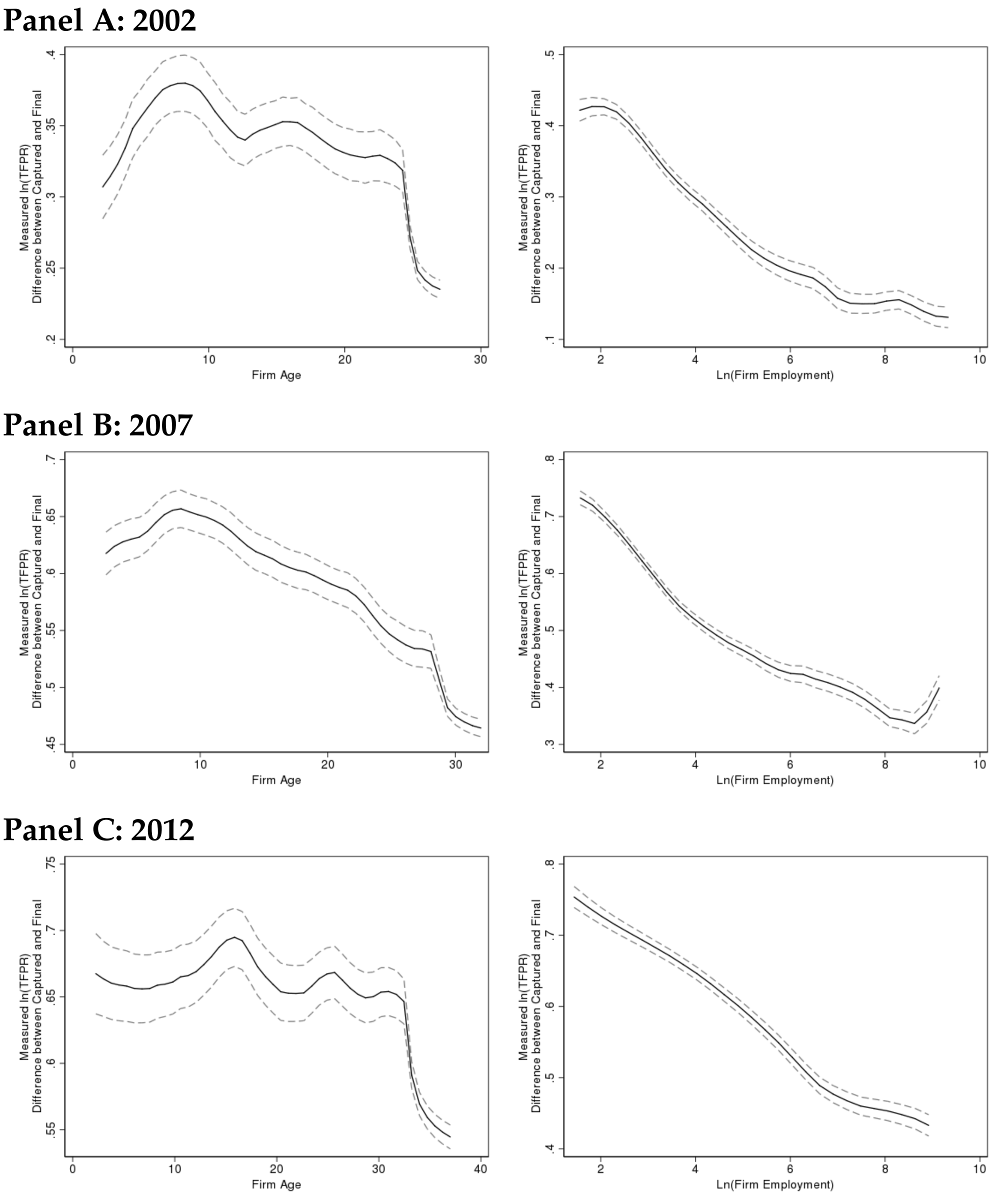


Panel A: 2002

Panel B: 2007

Panel C: 2012

Figure 3: Relationship between firm size/age & editing



Panel A: 2002

Panel B: 2007

Panel C: 2012

Table 1: Edits Made to the U.S. Census of Manufacturers

Edit/Impute Action	Occurs when...
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 (reported or imputed) value.
Model (Donor) Record (D)	the item is imputed using hot deck methods.
High/Low (E)	the item is imputed by direct substitution of value near (high or low) endpoints of imputation range.
Goldplated (G)	the reported value for the item is protected from any changes by the edit. The value of a goldplated item is not changed by the editing system, even if the item fails one or more edits. In general, the goldplate flag is set by an analyst.
Historic (H)	the item is imputed by ratio imputation using historic data for the same establishment (for example, prior year data imputation in Manufacturing)
Subject Matter Rule (J)	the item is imputed using a subject matter defined rule (e.g. y=1/2x).
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. This preserves the basic distribution of the details.
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.
Restore Reported Data (O)	the reported value fails an edit. Either an analyst interactively restores the originally reported value of an edit (set by the interactive update system) or the ratio module later imputes originally reported data for an item which was imputed in the previous edit pass.
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 (balance module default).
Unable to Impute (U)	the reported item is blank or fails an edit, and the system cannot successfully substitute 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.
Acceptable Zero (Z)	the reported value for an item is zero, and the item has passed a presence (zero/blank) test. This often occurs with part time reporters (e.g., births, deaths, idles). The zero value will not be changed, even if it fails one or more edits.

Notes: Edit rule descriptions are from Grim (2011) and White et al. (2018).

Table 2: Dispersion of TFPR in the U.S. Census of Manufactures

Panel A: Gross Output

	Captured Data			Census-Cleaned Data		
	Outcome			Outcome		
Year	St. Dev	90/10	75/25	St. Dev	90/10	75/25
2002	0.889	1.337	0.577	0.401	0.783	0.331
2007	0.955	1.716	0.902	0.442	0.87	0.356
2012	1.089	1.888	1.031	0.421	0.831	0.346

Notes: The TFPR calculation follow Bils, Klenow and Ruane (2017).

Panel B: Value Added

	Captured Data			Census-Cleaned Data		
	Outcome			Outcome		
Year	St. Dev	90/10	75/25	St. Dev	90/10	75/25
2002	0.981	1.779	0.895	0.575	1.238	0.554
2007	1.1	2.227	1.172	0.616	1.338	0.597
2012	1.256	2.487	1.291	0.626	1.304	0.58

Notes: The TFPR calculation follow Bils, Klenow and Ruane (2017).

Table 3: Measured Allocative Efficiency in the 2002, 2007, and 2012 U.S. Census of Manufactures

Panel A: Gross Output

	Captured Data			Census-Cleaned Data		
	Trimming %			Trimming %		
Year	0%	1%	2%	0%	1%	2%
2002	0.00005	0.109	0.176	0.14	0.461	0.554
2007	0.000005	0.012	0.024	0.042	0.302	0.425
2012	0.00000038	0.004	0.024	0.059	0.349	0.455

Notes: The values follow Bils, Klenow and Ruane (2017). Each cell represents a different starting point: either the Census-cleaned or raw data, and trimming the 0, 1, or 2% extremes for TFPQ and the input wedges.

Panel B: Value Added

	Captured Data			Census-Cleaned Data		
	Trimming %			Trimming %		
Year	0%	1%	2%	0%	1%	2%
2002	0.021	0.49	0.55	0.452	0.649	0.714
2007	0.026	0.316	0.435	0.395	0.62	0.673
2012	0.005	0.274	0.392	0.452	0.652	0.696

Notes: The values follow Bils, Klenow and Ruane (2017). Each cell represents a different starting point: either the Census-cleaned or raw data, and trimming the 0, 1, or 2% extremes for TFPQ and the input wedges.

Table 4: Measured Allocative Efficiency in the
2002, 2007, and 2012 U.S. Annual Survey of Manufactures

Panel A: Gross Output

	Captured Data			Census-Cleaned Data		
	Trimming %			Trimming %		
Year	0%	1%	2%	0%	1%	2%
2002	0.003	0.209	0.415	0.160	0.458	0.555
2007	0.00004	0.026	0.058	0.085	0.294	0.416
2012	0.00007	0.004	0.074	0.077	0.340	0.457

Notes: The values follow Bils, Klenow and Ruane (2017). Each cell represents a different starting point: either the Census-cleaned or raw data, and trimming the 0, 1, or 2% extremes for TFPQ and the input wedges.

Panel B: Value Added

	Captured Data			Census-Cleaned Data		
	Trimming %			Trimming %		
Year	0%	1%	2%	0%	1%	2%
2002	0.061	0.470	0.624	0.469	0.620	0.669
2007	0.101	0.442	0.503	0.431	0.585	0.637
2012	0.071	0.396	D	0.461	0.623	0.666

Notes: The values follow Bils, Klenow and Ruane (2017). Each cell represents a different starting point: either the Census-cleaned or raw data, and trimming the 0, 1, or 2% extremes for TFPQ and the input wedges.

Table 5: Cross-Country Comparisons of Measured Allocative Efficiency

	Allocative Efficiency	Indian AE relative to 2002-2012 Average in:	
Year	India	Census-cleaned US	Captured US
2002	0.387	1.05	9.29
2007	0.385	1.04	9.24

Notes: The first column shows estimates of allocative efficiency for India using Bils, Klenow and Ruane (2017). The second and third columns show Indian allocative efficiency relative to the US cleaned and raw data. Indian data sources are discussed in Appendix A.I

Table 6: Changes in Measured Misallocation due to Edits

Edit	Direct Effect on Measured Misallocation	Shapley Value	Shapley Share
TVS logical impute	-166%	-64%	0.21
Impute for Missing	-155%	-58%	0.19
Analyst correction	-130%	-48%	0.16
Payroll logical impute	-140%	-40%	0.13
Divide by 1000 edit	-89%	-24%	0.08
Other imputes	-85%	-27%	0.09
Regression imputes for Materials	-73%	-19%	0.06
Logical imputes for Materials	-59%	-21%	0.07
Administrative Record Impute	-6%	-5%	0.02

Notes. This table shows the effect of each type of edit on measured misallocation in the United States. Column 1 reports the difference in measured misallocation between the original raw data and the raw data, but with the cleaned final data for the firms affected by that row's edit. Since the sum of the changes is not equal to the difference in measured misallocation between the raw and clean data, Column 2 reports each row's Shapley Value and Column 3 the share of the change.

Table 7: Measured Allocative Efficiency After Common Data Cleaning

		Captured Data		Data Cleaned using Kim et al.		
		Trimming %		Trimming %		
Country	Year(s)	1%	2%	0%	1%	2%
U.S.	2002	0.116	0.189	0.257	0.499	0.581
U.S	2007	0.015	0.026	0.038	0.161	0.369
U.S.	2012	0.004	0.031	0.029	0.231	0.384
India	2000-2011	0.393	0.441	0.465	0.521	0.554

Notes: The values follow Bils, Klenow and Ruane (2017). Each cell represents a different starting point: either the raw data or data cleaned using Kim et al. (2015), and trimming 0, 1, or 2% extremes for TFPQ and the input wedges.