Research with Private Sector Business Microdata: The Case of NETS/D&B

Leland D. Crane Ryan A. Decker*

July 12, 2020

Abstract

Business microdata have proven useful in a number of fields, but official sources of comprehensive microdata are subject to high access costs and confidentiality restrictions. A growing number of researchers instead use a private data source seeking to cover the universe of U.S. business establishments, the National Establishment Time Series (NETS) based on Dun & Bradstreet microdata. We study the representativeness of NETS in terms of the distribution of employment and establishment counts across industry, geography, and establishment size. We then evaluate NETS in terms of its ability to corroborate key insights from the business dynamics literature with a particular focus on the behavior of new and young firms. We find that NETS microdata can be made reasonably representative of U.S. business activity in the static cross section, but the data also exhibit patterns of business dynamics that are markedly different from official administrative sources. We make suggestions for researchers on how best to use, or not to use, NETS and Dun & Bradstreet data in economic research.

JEL Classifications: C81, L26, D22, M13, J42

*Both authors are at the Federal Reserve Board of Governors

This paper previously circulated under the title "Business Dynamics in the National Establishment Time Series (NETS)." We thank Keith Barnatchez for early contributions to our study of NETS data. We thank Joonkyu Choi, John Haltiwanger, Sabrina Howell, Ziho Park, Erick Sager, Jim Spletzer, and conference and seminar participants at CAED 2019, SGE 2019, and the Federal Reserve Board for helpful comments. We also thank Gray Kimbrough for making his graphing schema available.

1 Introduction

Research based on business microdata has become increasingly important in economics in recent years. Such research can be difficult, however. The most easily available sources of business microdata, such as Compustat, do not cover the universe of private businesses (Compustat only covers publicly traded firms) and may be subject to significant selection problems (Davis, Haltiwanger, Jarmin and Miranda (2007)). The availability of comprehensive U.S. business microdata has increased significantly during the last decade due to efforts by statistical agencies; but access to these data is still costly, and prudent (and legally mandated) confidentiality restrictions limit the scope of research that can be conducted with such data. A prominent private sector data source has emerged, however, with nominal coverage of a significant fraction of the universe of U.S. business activity and without onerous publication scope restrictions. The National Establishment Time Series (NETS), a product of Walls & Associates, consists of longitudinally linked Dun & Bradstreet (D&B) establishment-level data (with firm linkages) on business employment, sales, industry, and location, as well as other variables of potential interest to researchers and policymakers.[1]

We first explore the representativeness of NETS in the cross section, comparing the data to the Census Bureau's County Business Patterns (CBP) and Nonemployer Statistics (NES).[2] Static distributions of NETS data can be made reasonably comparable to official sources, on average, in terms of establishment size, industry, and geography cells, subject to important limitations arising from divergent counts of small establishments and, to a lesser extent, very large establishments. These differences may be due to imputation in NETS, the mismeasurement of employment at nonemployer businesses, or coverage of informal activities.

We then investigate establishment growth and firm lifecycle patterns in NETS, including higher moments of growth distributions, motivated by key insights from the firm dynamics literature.[3] We find that NETS data cannot replicate key empirical patterns of establishment and firm growth documented in comprehensive, official administrative data—the Census Bureau's Longitudinal Business Database (LBD) and associated public-use product, the Business Dynamics Statistics (BDS).[4]

[1] While we focus on the NETS package in this paper, our conclusions apply broadly to the Dun & Bradstreet data from which NETS is derived. A large literature in entrepreneurship and related fields uses Dun & Bradstreet data directly.

[2] Our cross-sectional analysis draws heavily on our previous work in Barnatchez, Crane and Decker (2017). In that work we also show that comparisons between NETS and CBP are broadly consistent with comparisons between NETS and the BLS Quarterly Census of Employment and Wages (QCEW).

[3] We adopt the Census Bureau definitions of "establishments" and "firms" where an establishment is a single business operating location, and a firm is a collection of one or more establishments under common ownership or control. We describe our method for mapping NETS into Census-comparable concepts below.

[4] We did not access LBD microdata for this project; rather, we rely on published results from the LBD as well as our own analysis of the publicly available BDS. All previously published LBD results we describe have undergone appropriate Census Bureau disclosure avoidance processes to ensure that no confidential information

A significant, though not sole, reason for these limitations appears to be the prevalence of data imputation in NETS, the effects of which are magnified in a dynamic setting. In 2014, employment is imputed for more than two thirds of establishments with fewer than five employees, while the imputation rate is more than one third for establishments of five to nine employees. Even larger establishment classes have employment imputation rates close to 10 percent. Imputation is particularly prevalent among young firms: about 90 percent of firms aged zero or one have imputed employment data. Imputation can be particularly consequential in dynamic settings where multiple years of data must be relied upon for a single observation. We find that in 2014, the employment data for 10 percent of firms had been imputed for seven or more *consecutive* years. Imputation of sales data is even more prevalent, at rates around 80 percent among small firms and 95 percent among large firms. Nearly all of the sales data for establishments of multi-establishment firms are imputed.

More broadly, business-level employment data exhibit surprisingly little volatility in NETS. The distribution of firm employment growth rates is far less dispersed and skewed in NETS than in official data. Young firm growth, which existing literature shows is characterized by substantial dispersion and skewness, is particularly poorly captured in such a setting. NETS appears ill suited for the study of labor market flows, firm entry and exit, and business lifecycle dynamics, though careful use of NETS in case study settings may still be appropriate.

Our findings point to significant limitations of NETS and D&B data for the study of business dynamics, but this does not mean these data cannot be used for research. We conclude by offering concrete suggestions for researchers.

The paper proceeds as follows. In Section 2, we briefly describe the NETS data, related literature, and our data preparation methods. Given the prevalence of data imputation in NETS, in Section 3 we describe patterns of imputation with a particular focus on implications for business dynamics measurement. In Section 4 we compare NETS data with official sources in terms of static, cross-sectional distributions of economic activity. In Section 5 we compare NETS data with official sources in terms of aggregate patterns of business dynamics, the geographic and industrial composition of firm growth, and the lifecycle behavior of firms. Section 6 provides our concrete suggestions for using NETS and D&B data then provides an argument for preferring official data to NETS when discrepancies between the two arise; some readers may prefer to start in that section. Section 7 concludes.

2 Data background and preparation

2.1 NETS background

Barnatchez, Crane and Decker (2017) describe NETS data in detail; we refer the interested

is disclosed.

reader to that paper for more details while we provide a short summary here. For many years, Dun & Bradstreet has actively sought to maintain a database of all business establishments in the U.S., which the firm uses in its business of selling marketing and other information. D&B collects these data from state secretaries of state, Yellow Pages, court records, credit inquiries, and direct telephone contact. Each year, D&B provides a snapshot of the establishment cross section to Walls & Associates, which creates longitudinal links and cleans the data for use by researchers and others. The finished data include annual establishment-level information on detailed industry, employment, sales, and other variables, with longitudinal establishment linkages and firm identifiers to link the establishments of multi-unit firms. The Census Bureau's LBD, widely used by academic and government researchers, is similar in structure and aspiration to NETS, except that NETS seeks to track nonemployer businesses while the LBD is limited to employers (i.e., businesses with at least one paid employee). The NETS product to which we have access covers the years 1990-2014.

A key question about the NETS/D&B data concerns the nature of the business universe covered by the data. NETS marketing materials advertize full coverage of the entire U.S. business universe, including nonemployer businesses (i.e., those businesses without formal employees, omitted from several popular official datasets such as CBP and QCEW). Figure 1, which is taken from Barnatchez, Crane and Decker (2017), compares NETS to the business universe known to the U.S. Census Bureau. The dashed lines refer to the Census universe. The lower black dashed line is the total number of establishments in the Census Bureau employer universe; that is, this line reports the number of business establishments that have formal employees (i.e., those who are on Social Security Administration records and are issued W-2 paperwork). These are given by the County Business Patterns product but arise from the Business Register that also feeds the Statistics of U.S. Businesses (SUSB) program as well as the Longitudinal Business Database (LBD) and Business Dynamics Statistics (BDS) products. The upper red dashed line adds nonemployer businesses as reported in the Census Bureau's Nonemployer Statistics (NES), which rely on IRS data on business income. That is, the dashed red line is the sum of employers and nonemployers and therefore represents the entire universe of business activity as known to the Census Bureau.[5]

The solid lines on Figure 1 refer to the NETS universe. A common view in the literature is that NETS data often include all workers at a location, even those that are not formal employees (such as business owners not on the payroll). Following previous literature, then, we construct a NETS "employer" or "payroll" universe using the assumption that each firm has one non-payroll owner reported to D&B as an employee. That is, we subtract one employee from every firm (at the headquarters establishment) and discard any establishments whose resulting employment is zero. This NETS "payroll" or employer universe is shown by the

[5]There are some specific areas of the economy that are known to be missing from Census data: primarily farms (NAICS 111 and 112), of which there are about 2 million (a quarter of which are employers) according to the Department of Agriculture's Census of Agriculture (2017); and railroads (NAICS 482), of which there are fewer than 1,000 firms employing roughly 200,000 people, according to the Railroad Retirement Association.

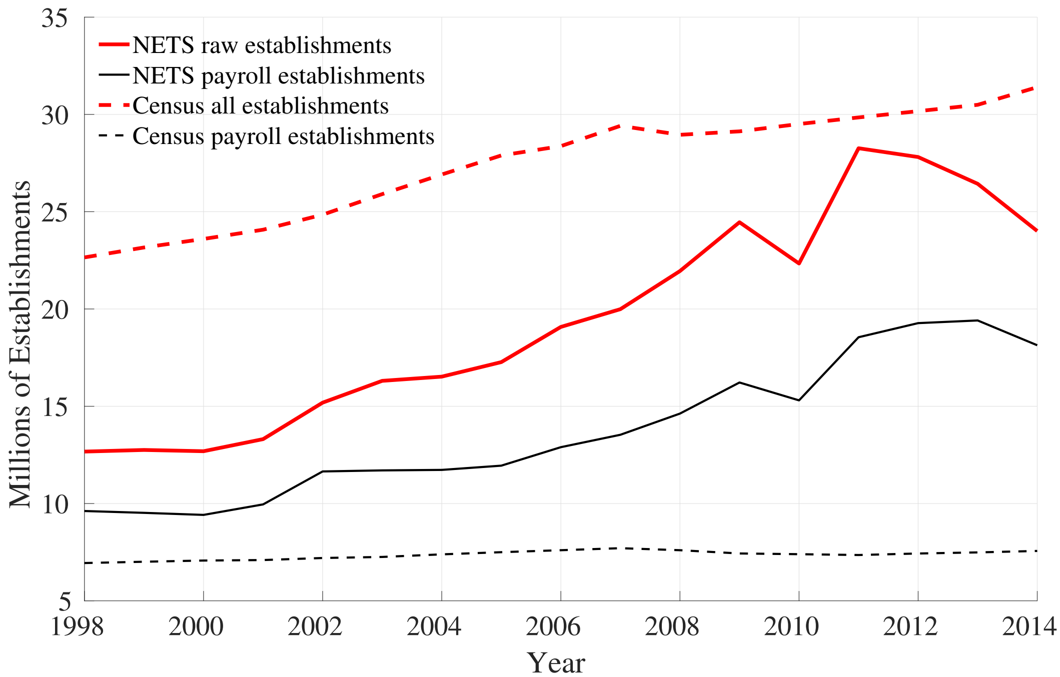


Figure 1: Establishment counts in NETS and Census Bureau data

solid black line. As is evident, NETS has too many employer establishments to be consistent with the Census Bureau employer universe, and the excess coverage is rising over time. This is also true in comparisons against QCEW, the independently constructed Bureau of Labor Statistics (BLS) product based on administrative data from the unemployment insurance system.[6] This likely reflects increased coverage by D&B rather than true business creation; the rise in excess coverage is concentrated among small establishments and features extremely high imputation rates (as we will show below). The solid red line on Figure 1 shows the total raw establishment count in NETS, inclusive of both employers and nonemployers. This line is below the total business universe line for the Census Bureau—substantially below it in earlier years.

In short, it is difficult to define the NETS/D&B universe. NETS has too many establishments to be the U.S. employer universe, and it has too few establishments to be the total (employer plus nonemployer) universe. This raises considerable uncertainty about NETS coverage and suggests caution about using NETS to construct aggregate statistics. Moreover, the apparent improvement in coverage during the 2000s is likely to reflect spurious

[6]Importantly, CBP and QCEW do not exactly match each other—in recent years, QCEW has seen more growth in establishment counts, primarily among smaller establishments—but Barnatchez, Crane and Decker (2017) show that CBP and QCEW are much closer to each other than either is to NETS.

business entry (i.e., preexisting businesses newly added to D&B data) that will create error in the measurement of business formation.

Barnatchez, Crane and Decker (2017) review recent papers using NETS data for a variety of research questions.[7] Here we describe just two key references. First, Neumark, Zhang and Wall (2005) evaluate the California sample of NETS through comparisons to QCEW. The authors recommend dropping establishments with one employee to approximate the employer universe; we adopt a modified (firm-level) version of this rule in our work. Neumark, Zhang and Wall (2005) also highlight the prevalence of imputation in the data and note that frequent imputation causes a low frequency of employment change at the establishment level. Most relevant to our purposes, the authors calculate employment growth at the county-by-industry level and study the correlation of employment growth between NETS and QCEW. Annual employment growth is weakly correlated between the two sources (0.528), so the authors study 3-year employment growth, which shows a correlation of 0.864. These aggregate exercises are useful and suggestive; we differ from Neumark, Zhang and Wall (2005) in focusing on the nationwide NETS sample and on a wider range of measures of business dynamics.

Separately, Echeverri-Carroll and Feldman (2017) evaluate the usefulness of NETS for studying entrepreneurship and firm entry by focusing on two specific case studies: the Austin-Round Rock (Texas) metropolitan statistical area and the North Carolina "Research Triangle." The authors compare NETS to Texas and North Carolina secretary of state (SOS) data (compiled by Guzman and Stern (2016)) and recommend restricting the data as follows: exclude known sole proprietorships (which do not appear in secretary of state data) and firms with nonprofit components, focus on headquarters establishments, and omit single-employee firms (as we do in the present paper and related work). With these restrictions, NETS data match secretary of state data for the two cities reasonably well, though there still exist significant discrepancies particularly in recent years of data. Importantly, the authors show that successive NETS vintages revise heavily for several recent years, so NETS reliability improves over time yet should be expected to be weak for the most recent years in the data (particularly the most recent four years).

A particularly notable contribution of Echeverri-Carroll and Feldman (2017) is that they match NETS microdata with SOS data for software startups in Austin, a painstaking process with large benefits for our questions here. They first exclude recent years of data to avoid vintage problems discussed above. They then seek to match about 3,500 NETS firms to the SOS data, first focusing on name and zip code matches, then relaxing to name matches only, using standard name generalization techniques. They successfully match about 40 percent

[7] Some additional examples of recent work, not reviewed there, are Heider and Ljungqvist (2015), Faccio and Hsu (2017), Chava, Oettl, Singh and Zeng (2018), Rossi-Hansberg, Sarte and Trachter (2018), Ma, Murfin and Pratt (2020), and Denes, Howell, Mezzanotti, Wang and Xu (2020). Cho, Mclaughlin, Zeballos, Kent and Decken (2019) match NETS to other establishment-level databases, as well as public-use Economic Census files.

of NETS firms to SOS firms. Among those matched, only 50 percent report the same entry year in NETS as in SOS data. About 75 percent have NETS and SOS entry years within two years of each other, and about 80 percent are within 3 years. The authors discuss reasons for the low match rate, which include missing legal form of organization data in NETS. The implications of this exercise are mixed, but the SOS data provide a degree of validation of NETS and suggest usefulness in limited exercises, particularly in case study settings similar to Echeverri-Carroll and Feldman (2017).

While Echeverri-Carroll and Feldman (2017) focus on carefully matched microdata comparisons within two specific case studies (Austin and the Research Triangle), we focus more broadly on comparisons using known patterns of firm dynamics across the U.S. We will argue that NETS data are of limited usefulness for studying broad patterns of firm dynamics, leaving the Echeverri-Carroll and Feldman (2017) case study approach as the better (though more tedious and time intensive) use case for NETS.

2.2 LBD and CBP background

County Business Patterns (CBP) and the Longitudinal Business Database (LBD), products of the U.S. Census Bureau, cover the near-universe of private nonfarm employer business establishments in the U.S. starting in the mid-1970s. Both are constructed from the Census Bureau's Business Register, which is drawn from federal business tax records (both IRS and Social Security Administration), a variety of Census Bureau surveys, and the semi-decadal Economic Censuses (conducted in years ending 2 and 7).[8] Importantly, the source data include, by construction, all in-scope employer businesses in the U.S. that are known to tax authorities.[9] CBP data are publicly available in annual tabulations of static employment and establishment counts by industry and narrow geography. LBD data consist of longitudinally linked establishment microdata, available only to sworn researchers in Federal Statistical Research Data Centers; the establishment-level records include employment (as of March 12 of a given year), detailed industry, location, and other establishment characteristics. Establishment records also include firm identifiers that group establishments under common ownership or operational control.

The LBD has become a critical resource for the study of firm dynamics. For example, Davis, Haltiwanger, Jarmin and Miranda (2007) first documented multi-decade declines in measures of firm-level employment volatility and gross job flows in the U.S. private sector using the LBD; the authors also linked the LBD to Compustat, a widely used dataset of publicly traded firms, and documented key differences in the behavior of publicly traded and

[8]Barnatchez, Crane and Decker (2017) describe features of the Business Register that are relevant for comparisons with NETS. The Statistics of U.S. Businesses (SUSB) product is also based on the Business Register. DeSalvo, Limehouse and Klimek (2016) describe the Business Register in exhaustive detail. Jarmin and Miranda (2002) describe the construction of the LBD.

[9]The primary scope restrictions are the omission of farms and railroads, as previously mentioned.

privately held businesses. Haltiwanger, Jarmin and Miranda (2013) used the LBD to show that the job creation contribution that is widely attributed to small firms is more appropriately attributed to young firms. Decker, Haltiwanger, Jarmin and Miranda (2014) described key characteristics of young firms in the LBD, including "up-or-out" dynamics and high growth rate dispersion and skewness. Alon, Berger, Dent and Pugsley (2018) show that cohort productivity growth declines with age and that high productivity growth of young firms is primarily a selection phenomenon. A further large literature exploits the LBD for studies of international trade, labor market flows, and a wide range of other topics.

While the LBD has become the primary resource for research on firm dynamics, it is subject to strong confidentiality requirements and is therefore only accessible to sworn researchers with approved projects working in the Census Bureau or a Federal Statistical Research Data Center (FSRDC). Specially sworn researchers using the LBD in FSRDCs must carefully follow rules to comply with federal law and prudent confidentiality concerns, and publishing results from statistical work on the LBD requires a lengthy process for disclosure avoidance. The process is generally costly and time consuming, consistent with researchers' obligation to ensure the protection of confidential information. Given the importance of the data, therefore, the Census Bureau publishes the publicly available Business Dynamics Statistics (BDS), which consists of aggregates of the LBD designed to track business dynamics at the level of sectors, firm age and size groups, and establishment locations. Research using the BDS has made considerable contributions to the literature. However, there are many questions that cannot be answered with the BDS, particularly questions about higher moments of the firm distribution and firm dynamics, that require microdata.

The limitations of the BDS and CBP and the tradeoffs involved with LBD access and use create demand for a public use source of business microdata like NETS. It is therefore important that researchers understand the strengths and limitations of NETS. The main purpose of this paper is to compare NETS with CBP and the LBD, with the latter serving as the benchmark against which any employer business microdata should be judged given its well-defined and near-universal scope and its wide use in the literature (see Section 6 for more discussion of official versus private data sources). We do not present any original results from LBD microdata; rather, we compare our original NETS calculations to existing LBD and BDS calculations from the literature.

2.3 NETS data preparation

2.3.1 Sample restriction

We first implement sample restrictions described in detail by Barnatchez, Crane and Decker (2017). First, since NETS aspires to include both the nonemployer and employer universe, and since coverage beyond the employer universe is evident in the data, we restrict the sample to our best guess of the employer universe by subtracting one employee from the employment of each firm headquarters establishment then dropping establishments with zero (post-subtraction) employment.[10] This is a modified version of the sample restriction recommended by Neumark, Zhang and Wall (2005) and follows from the notion that owners are likely to be counted as employees in NETS though they may not be in official sources, where employment has a strict definition based on paycheck issuance. We restrict NETS to the employer universe to be comparable with the datasets to which we will make comparisons—CBP, the LBD, and the BDS—which are employer datasets. We then restrict the NETS sample to match the industry scope of these datasets (see Barnatchez, Crane and Decker (2017) for a specific industry scope list).

2.3.2 Establishment identifiers

Studying business dynamics is more complicated than studying cross-sectional snapshots of microdata. In particular, questions of business dynamics require careful attention to longitudinal linkages of business identifiers. Data problems (such as broken linkages) and real-world events like mergers and acquisitions generate challenges to longitudinal concepts and require researchers to make judgments. Given our goal of assessing the NETS data relative to official data, we attempt to treat the NETS data in a way that makes them most comparable to the LBD and the empirical firm dynamics literature based on the LBD.

The basic unit of observation in NETS is the *dunsnumber*. D&B views the *dunsnumber* as a line of business; but with respect to official sources, it is most similar to the concept of an establishment. In the LBD, an establishment is a single business operating location (identified by *lbdnum* in the LBD). In NETS, though, a single business operating location can have multiple *dunsnumbers*. This can be the case, for example, when the production operations and sales operations of a business are co-located but counted separately by D&B. In Barnatchez, Crane and Decker (2017), we aggregate *dunsnumbers* to the establishment level to be consistent with CBP and QCEW establishment definitions; to do this, we identify *dunsnumbers* that have the same reported firm headquarters (*hqduns*), 5-digit zip code, and first five street address characters (i.e., roughly speaking, same street and building number). This approach is designed to identify lines of business operating in the same location and falling under the same firm. We then sum the employment of the matched lines of business and assign the merged establishment a new identifier (termed the *locduns*) and the industry code of the largest line of business (in terms of employment). Since establishments in official data are assigned industry codes to reflect their principal activity, this method of merging D&B lines of business should roughly approximate the official concept. In practice, the line of business vs. establishment distinction seems to matter mostly for a small number of headquarters establishments.

[10]In ongoing work we study the NETS coverage of the nonemployer universe in more detail, but Figure 1 already suggests that NETS coverage is incomplete.

We follow the approach above for constructing establishment microdata, but we introduce additional procedures for ensuring the longitudinal integrity of the resulting merged *locduns* establishment identifiers. A naive application of the Barnatchez, Crane and Decker (2017) method could result in spurious changes in *locduns* establishment identifiers that reflect changes in the composition of establishment employment rather than the death of one establishment and birth of another. We first identify a *locduns* establishment as a continuer (i.e., not a birth or death) if there is a year-to-year overlap in at least one original line-of-business *dunsnumber*; that is, if a *locduns* disappears from the data we only assume the establishment has exited if all its associated line-of-business *dunsnumbers* cease to exist. We create a new identifier, the *netsnum*, that does not change from year to year even if a merged establishment's *locduns* changes due to changing employment composition of lines of business. In the (rare) case that lines of business that exist in the same location but belong to different firms (i.e., have different *hqduns*) in year *t-1* move into the same firm (i.e., take on the same *hqduns*) in year *t*, we assign the year-*t* combined entity the *netsnum* of the year-*t-1* *locduns* establishment that contributed the most employment (in terms of lines of business) to the new entity.[11]

The resulting *netsnum* is a longitudinal identifier that is close in concept and spirit to the longitudinal establishment identifier in the LBD (*lbdnum*). We next focus on longitudinal firm identifiers.

2.3.3 Firm identifiers

A firm is a collection of establishments. The LBD defines the firm based on common ownership or operational control. The NETS firm concept is based on a common headquarters establishment (*hqduns*), where the *hqduns* refers to the *dunsnumber* of the headquarters establishment. NETS apparently allows for multiple levels of headquarters—perhaps including both regional and national headquarters—because we observe some cases in which an establishment record has a *dunsnumber* that is equal to other establishments' *hqduns*, but that itself refers to a different *hqduns*.[12] That is, there are cases in which an establishment appears to be a headquarters for other establishments but does not refer to itself as its own headquarters. We attempt to unite all establishments that are related through headquarters, either directly or indirectly, under a single firm identifier by "rolling up" *hqduns* identifiers. That is, we assign all related establishments the *hqduns* of the highest level headquarters, which necessarily reports itself as its own *hqduns* (or, in rare cases, reports an *hqduns* that does not appear as a *dunsnumber* anywhere else).[13]

[11] This is a rare occurrence because it suggests that two separate firms with establishments in the same building engaged in a merger or acquisition.

[12] There is some discussion of this in NETS marketing materials.

[13] In extremely rare cases, we observe headquarters linkages that are "cyclical;" for example, *dunsnumber* A reports *dunsnumber* B as its headquarters, while B reports A as its headquarters. In those cases, we arbitrarily

The firm identifier setup in NETS also presents the longitudinal challenge of determining which groups of establishments are successors to each other over time. As with the LBD's firm identifier (*firmid*), the *hqduns* can change for many reasons, including merger and acquisition activity but also simple data problems. Unlike in the LBD, in NETS the firm identifier automatically changes if the firm moves its headquarters from one establishment to another. We reassign *hqduns* firm identifiers as follows. For a given firm x in year $t-1$, we identify all firms in year t that control at least some of firm x's $t-1$ establishments. We select x's *candidate successor* as the firm which controls the plurality of employment at these continuing establishments. Very often, this firm has the same $hqduns$ and essentially the same establishments as x, and there is no ambiguity. But when a firm "fractures" into several seperate entities, it is sensible to match the source firm to the largest continuing fragment.

One more step is necessary to have consistent firm linkages. According to the rule above, it is possible for a single period t firm to be the candidate successor for two distinct period $t-1$ firms. For example, a firm z in year t could include the largest continuing fragments of both firms x and y from year $t-1$. This would be the case for an acquisition or a merger. In such a case we assume that z is the successor to whichever of x and y accounts for the largest share of employment at the new firm. The successor firm is assigned the same $hqduns$ number as the source firm. Firms which lack a successor are assumed to have ceased to exist. This process is repeated year by year for the whole sample. This treatment of mergers has a number of limitations, though LBD firm identifiers are also not immune to merger problems and we accordingly follow best practice from the literature when we define firm age and growth rates.

We construct firm age to be consistent with the widely used convention from the literature (e.g., Haltiwanger, Jarmin and Miranda (2013)). At the first appearance of a new firm identifier (*hqduns*) in the data, we assign the firm the age of its oldest establishment (where establishment age is given by years since the first appearance of the establishment's *netsnum*, which is described above). Thereafter, the firm ages naturally. This approach abstracts from problems associated with spurious changes in the firm's headquarters identifier and is consistent with the convention used in the LBD-based literature to which we will compare NETS data.

2.3.4 Growth rate concepts

In various places below we report statistics based on firm or establishment employment growth rates. We employ the widely used growth rate concept of Davis, Haltiwanger and Schuh (1996) (the so-called "DHS growth rate"). Let $E_{e,t}$ be employment in year t for establishment e. Then the establishment-level DHS growth rate is given by:

assign an *hqduns* to apply to all related establishments.

$$g_{e,t} = \frac{E_{e,t} - E_{e,t-1}}{0.5 \cdot (E_{e,t} + E_{e,t-1})}. \tag{1}$$

The DHS growth rate differs from standard growth rates by using average two-year employment in the denominator instead of simply employment in year $t - 1$. This growth rate concept has been widely used in the literature because it can accomodate entry (in which case, $E_{e,t-1} = 0$, $E_{e,t} > 0$, and $g_{e,t} = 2$) and exit ($E_{e,t-1} > 0$, $E_{e,t} = 0$, and $g_{e,t} = -2$).

While calculating establishment-level DHS growth rates is straightforward, calculating firm-level growth rates is more complicated due to the possibilities of mergers, acquisitions, and divestitures, which can generate changes in firm-level employment that do not reflect "organic" growth. Following Haltiwanger, Jarmin and Miranda (2013) and related literature, we focus on an "organic" growth concept that abstracts from such reorganizations. The firm-level organic growth rate for firm J is given by:

$$g^f_{J,t} = \frac{\sum_{e \in J}(E_{e,t} - E_{e,t-1})}{\sum_{e \in J} 0.5 \cdot (E_{e,t} + E_{e,t-1})}. \tag{2}$$

The summation subscript $e \in J$ limits the set of establishments being included to those that belong to firm J in year t, regardless of whether they belonged to firm J in year $t - 1$. That is, the firm growth rate is constructed as if all of the firm's establishments in year t belonged to the firm in year $t - 1$ (even if they did not in reality belong to firm J in year $t - 1$), and any establishments controlled by firm J in year $t - 1$ that were divested to a different firm between $t - 1$ and t do not affect the growth rate of firm J. Establishments controlled by firm J in year $t - 1$ that exited (i.e., failed or closed) between $t - 1$ and t do contribute to measured growth, with $E_{e,t-1} > 0$ and $E_{e,t} = 0$ as mentioned above.[14]

3 Imputation

3.1 Employment imputation

NETS data include an imputation flag (*empc*) that takes on four possible values: (0) actual figure, (1) bottom of range, (2) D&B estimate, and (3) Walls & Associates estimate. The first two categories ($empc \in \{0, 1\}$) indicate values reported to D&B by survey respondents (or found in other source materials), with the "bottom of range" category indicating that the respondent or other source data reported a range rather than a specific count. D&B uses proprietary cross-sectional imputation methods in cases of non-reporters ($empc = 2$). Walls & Associates estimates employment for all non-reporters in each year using a longitudinal

[14]It is straightforward to show that $g^f_{J,t}$ is equivalent to the employment-weighted average of the growth rates of all of the firm's year-t establishments (and exiters), where the employment weight is defined as average two-year employment as in the DHS denominator above.

Size class	Imputation rates (%) 2000	2007	2014
1 to 4	42	55	72
5 to 9	22	20	38
10 to 19	19	17	17
20 to 49	18	14	9
50 to 99	17	14	7
100 to 249	15	13	7
250 to 499	19	15	10
500 to 999	18	18	12
1000+	24	22	16

Source: NETS
Note: Establishments with imputed NETS employment as a percent of total NETS establishments by establishment size. NETS sample is restricted to CBP scope but does not merge lines of business.

Table 1: NETS imputation rates by establishment size

imputation method; in cases where this longitudinally imputed estimate differs from the D&B cross-sectionally imputed method, Walls & Associates inserts their own estimate and sets $empc = 3$. The Walls & Associates method uses regressions based on the time series of sales and employment for the establishment and its industry.[15] We consider all values of *empc* except $empc = 0$ to be imputed, where the imputation can be done by the respondent ($empc = 1$), D&B ($empc = 2$), or Walls & Associates ($empc = 3$).

Table 1, taken from Barnatchez, Crane and Decker (2017), reports the share of establishments with imputed employment data ($empc = 0$) by establishment size for selected years. Imputation is prevalent; we observe the highest imputation rates among the smallest establishments, followed by the largest establishments. Imputation of small establishments' employment data rises markedly over time; in 2014—the last year in our data—more than two-thirds of the smallest establishments have imputed employment data. In this respect, NETS data—particularly among small establishments—is better thought of as a sample of businesses than as a measure of the business universe. That said, imputation problems can be managed through the omission of the smallest establishments, which is also where NETS differs most markedly from official data (as we show below).[16]

[15] NETS imputation details are provided with NETS marketing materials, *Understanding Data in the NETS Database* (2009).

[16] In ongoing work, we are assessing the value of dropping imputed observations then applying sampling weights to remaining data.

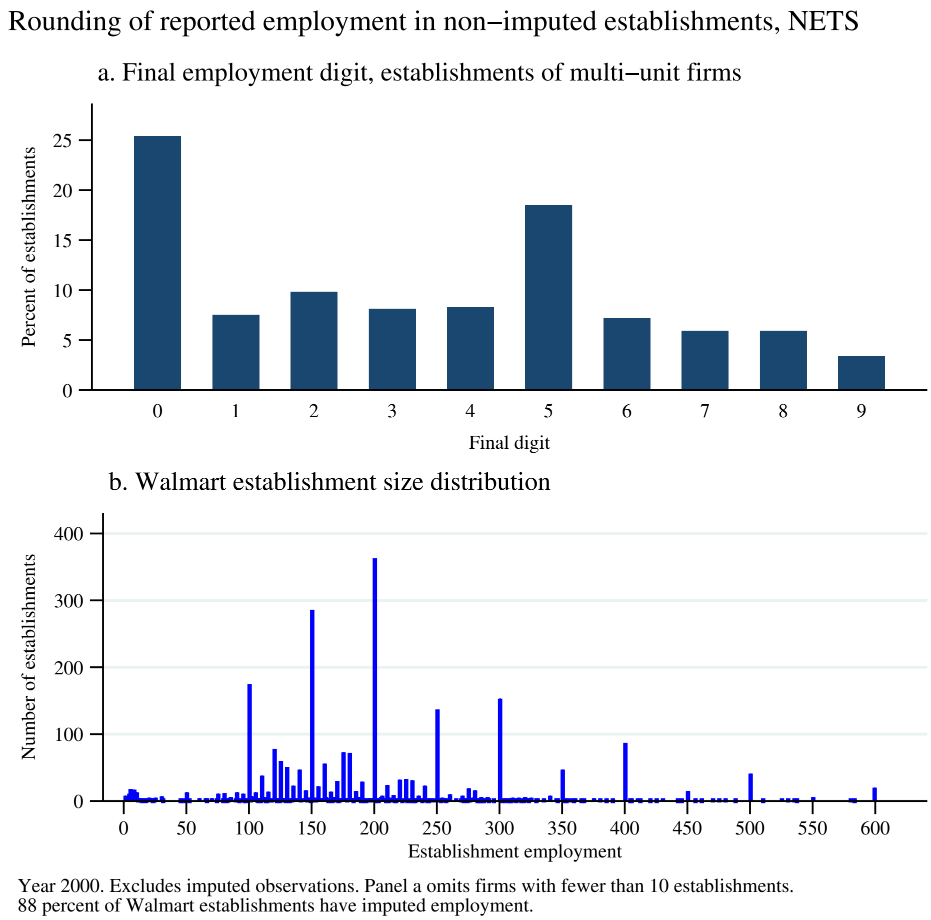


Figure 2

We also find evidence that respondents implicitly impute some data by rounding their reported employment figures, leading to a potential understatement of true imputation rates in NETS. Panel a of Figure 2 reports the distribution of last-digits of reported employment numbers among establishments of multi-unit firms (restricted to firms with at least 10 establishments), omitting establishments with positive imputation flags. Among these firms, we observe high counts of establishments with employment apparently rounded to the nearest 0 or 5. Panel b of Figure 2 shows the distribution of total establishment employment among Walmart establishments in the year 2000. In that year, 88 percent of Walmart establishments' employment data are reported as not being imputed; that is, they are coded with $empc = 0$. Yet among ostensibly non-imputed establishments we still observe overwhelming prevalence of employment figures that appear to be rounded to the nearest 50. This kind of rounding among large establishments such as Walmart's may seem benign, but in some contexts it may matter a great deal; for example, in small geography-by-industry cells, the rounding error on a Walmart establishment may be large relative to the size of competing retail establishments.[17]

This kind of rounding by respondents is a well-known issue in the survey methodology literature. We see more reasonable last-digit distributions among establishments generally (i.e., single-unit and smaller multi-unit firms), yet within large firms there appears to be significant rounding. This kind of rounding may have little cost in static or cross sectional settings, but below we make the case that the cost is higher in dynamic research.

A second source of non-imputation measurement error is seasonality. Official data sources include careful attention to the seasonal timing of employment measurement; for example, Census Bureau employment data typically reflect establishment employment as of March 12 of a given year. D&B data collection presumably proceeds on a rolling basis throughout the year. This potentially introduces an element of non-comparability between employment snapshots of different years within an establishment. In industries with significant seasonal employment patterns, employment growth could be materially mismeasured.

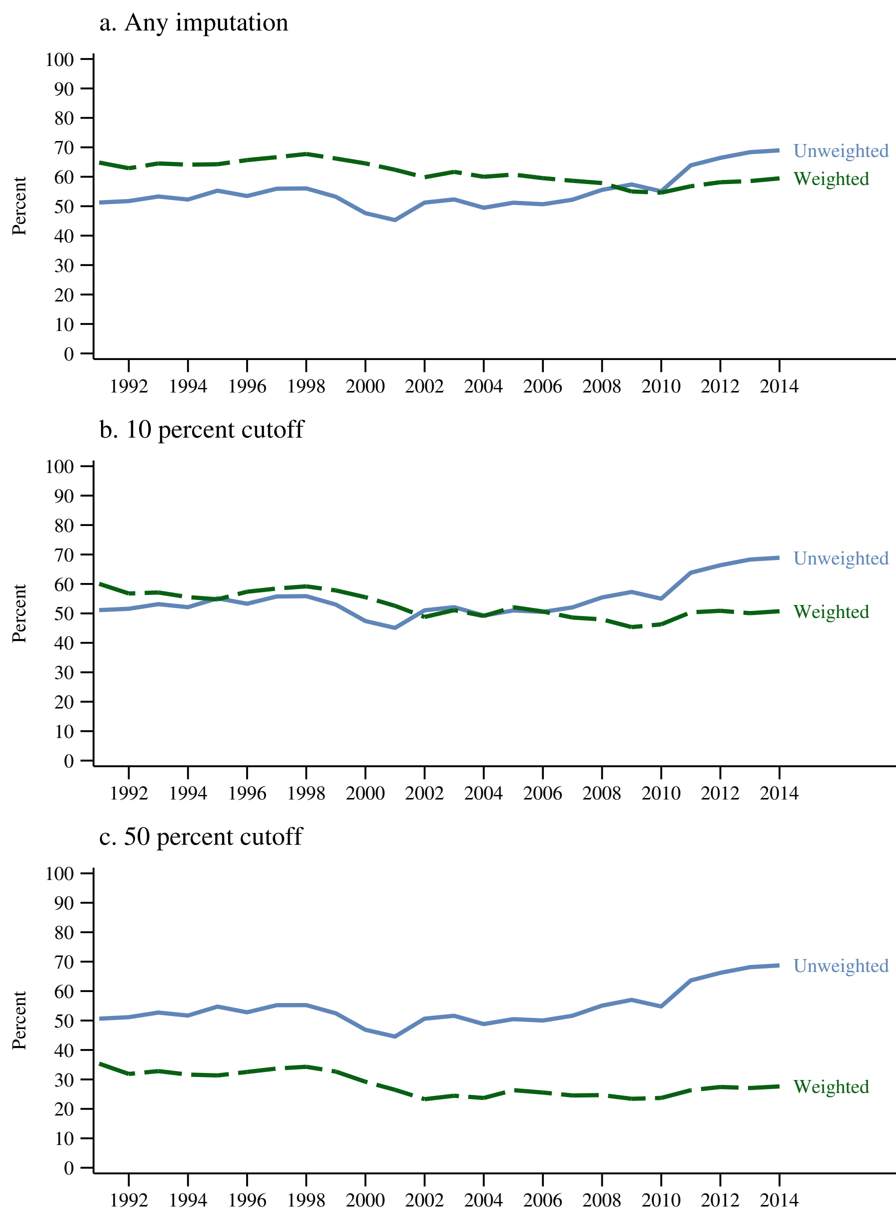
Given our focus on firm dynamics, we also explore *firm* imputation rates. Figure 3 reports the share of firms with imputed employment data; in panel a, the presence of *any* imputed establishments within a firm results in the firm counting as imputed (and establishments count as imputed if $empc \neq 0$). The solid blue line reports the share of firms that count as imputed, while the dashed green line reports the employment-weighted imputation rate (that is, the total employment—imputed or not—of imputed firms divided by total NETS employment). In early years, about half of NETS firms are imputed, but this share rises above two thirds by the end of the sample. Weighted imputation rates—the share of employment that is at imputed firms—are more steady, suggesting that the recent rise in unweighted imputation is primarily driven by smaller firms.

Panel b of Figure 3 uses a more restrictive (and NETS-friendly) definition of "imputed": we count firms as imputed if and only if at least 10 percent of their employment is at imputed establishments. This has no visible effect on the unweighted imputation rate, but the weighted rate moves down. The relaxation of the imputation definition has little effect on the unweighted imputation rate because it primarily affects a relatively small number of large firms. Panel c further relaxes the imputation standard, defining as imputed only those firms with at least half of their employment at imputed establishments, and substantial imputation is still evident. Even under an extreme definition of imputation (not shown) in which firms count as imputed only if at least 90 percent of their employment is at imputed establishments, about one fifth of employment is at imputed firms.

Firm imputation is therefore nontrivial. Imputation may cause only limited problems

[17] For example, this could affect studies like Rossi-Hansberg, Sarte and Trachter (2018), who investigate patterns of local industry concentration.

Firm employment imputation rates, NETS



Cutoffs indicate minimum share of firm reported employment that is imputed to count firm as imputed.
Weighted rate is share of reported employment at imputed firm; unweighted rate is share of firms imputed.

Figure 3

for cross-sectional studies, but there are several reasons imputation is much more costly in research on dynamics. First, the longitudinal imputation method of Walls & Associates necessarily uses data on the establishment time series, implicitly assuming that past and future behavior is indicative of present behavior and thereby dampening dynamic volatility. Moreover, Walls & Associates rely on industry and other data that may serve to minimize the dispersion of measured outcomes. Second, measures of dynamics depend on multiple consecutive data observations such that imputation is magnified. Concretely, employment growth from year $t-1$ to year t depends on employment levels in years $t-1$ and t; if either year's employment value is imputed, the overall employment growth value is necessarily imputed. Third, in the case of *firm* (rather than *establishment*) dynamics, imputation of any establishments within a multi-unit firm implies that the overall firm employment value is necessarily imputed. We find that this problem is particularly salient among firms with many establishments.

A striking way to see the longitudinal costs of imputation is to consider imputation "spells." We define the imputation spell as the number of consecutive years that a firm counts as imputed. For example, suppose a firm first counts as imputed in 1995. Then in 1995, the firm's imputation spell is equal to 1. If the firm is again imputed in 1996, then in that year its imputation spell is equal to 2. If the firm is not imputed in 1997, then its imputation spell in that year resets to 0. Figure 4 characterizes the distribution of imputation spells, where we count a firm as imputed if any of its establishments are imputed. Panel a reports the spell distribution among all firms. The solid green line (the highest line) reports the 90th percentile imputation spell. For example, in 1998, the 90th percentile firm had an imputation spell of 8, meaning that 10 percent of firms had been imputed for 8 or more consecutive years. The median firm had an imputation spell of zero for most years in the sample, but by the end of the sample the median had risen to 2 years.

Panel b reports the same exercise but restricts the sample to imputed firms in any given year; that is, the figure reports the distribution of imputation spells *conditional on* firms being imputed, rather than including non-imputed firms. Among imputed firms, even the 25th percentile reflects multiple consecutive years of imputation in many years, the median firm bounces between 2-year and 4-year imputation spells, and the 75th percentile shows imputation spells between 4 and 6 years.

The problem of consecutive imputation is particularly pronounced among large firms. Figure 5 reports the employment-weighted distribution of imputation spells; again, panel a reports the distribution among all firms. The 90th percentile of the weighted distribution has the maximum possible imputation spell throughout most of the sample (i.e., a spell of imputation beginning at the origin of the sample), as does the 75th percentile. This means that 25 percent of overall employment is at firms that have been imputed for the maximum possible number of consecutive years.[18] Panel b reports the distribution of imputation spells

[18] Of course, relaxing the definition of imputation improves these figures somewhat. Still, in the weighted

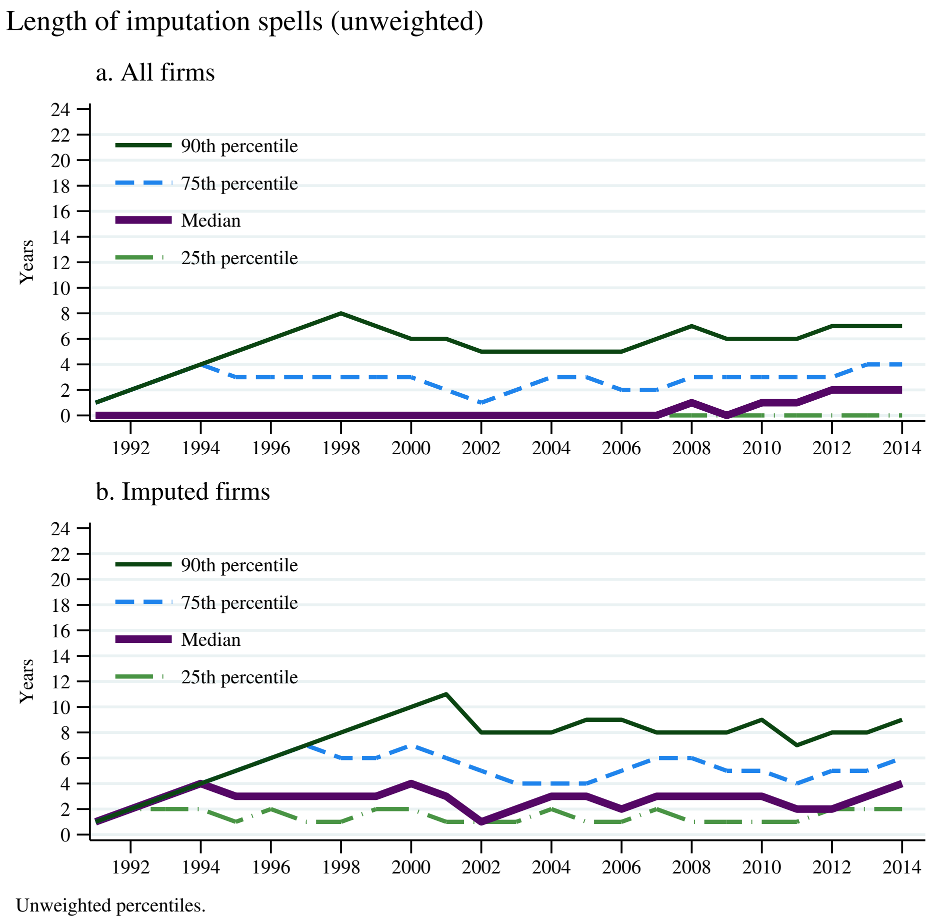


Figure 4

among imputed firms, with the striking pattern that the median of the weighted distribution exceeds 14 in recent years; that is, among imputed firms, half of employment is attached to firms with more than 14 consecutive years of imputed data.

Needless to say, the longitudinal integrity of data in which substantial shares of activity reflect firms whose data have been imputed for multiple consecutive years is limited.

We develop one other imputation measure to track longitudinal imputation on a year-to-year basis. For the rest of the paper, we define a firm as being longitudinally imputed in year t if it counts as imputed in *either* year t or year $t-1$. This definition is highly relevant when studying year-to-year firm-level growth or dynamics; as noted above, in a dynamic setting

distribution among all firms, if imputation is defined with the 25 precent cutoff we still observe 10 percent of employment concentrated at firms with at least 7 years of consecutive imputation.

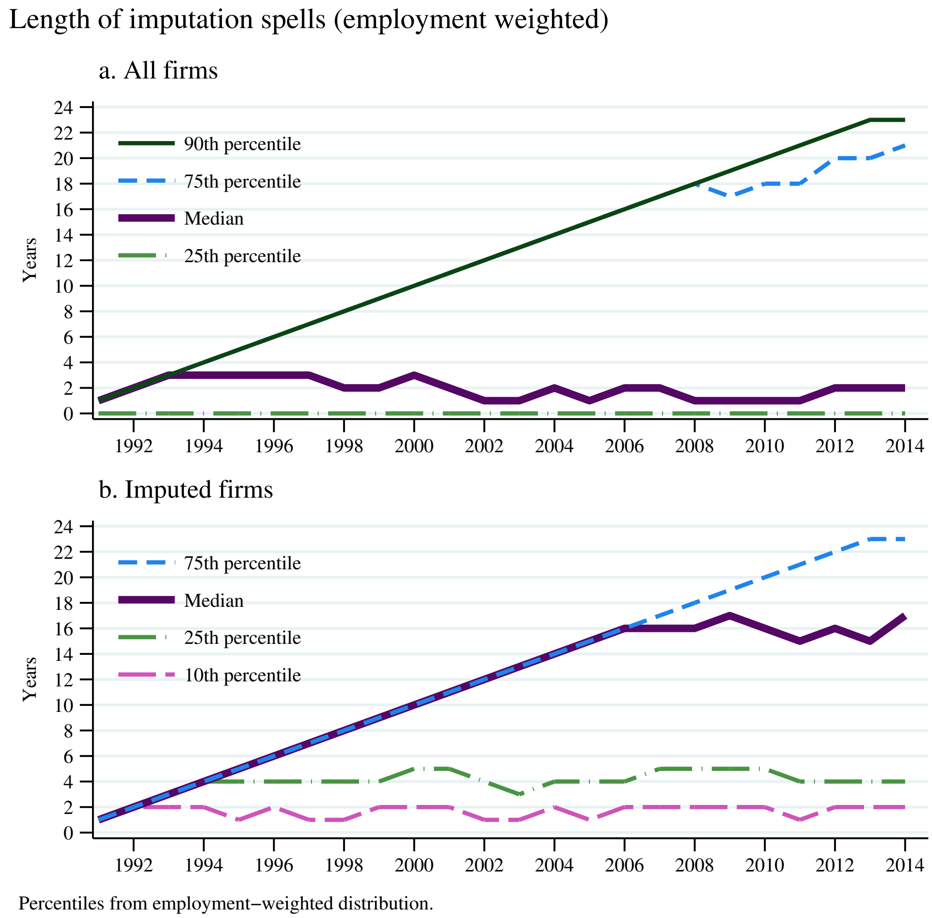


Figure 5

imputation binds in two consecutive years even if only one of the years has imputed data. We find that longitudinal imputation varies by firm age in critical ways. For example, Figure 6 reports longitudinal imputation rates by firm age for two snapshot years, 2003 and 2014. As noted elsewhere and in existing literature, the most recent year of data sees particularly acute imputation problems. But even in revised data, imputation is extremely prevalent among young firms, with rates above two thirds prior to age 3. These high imputation rates among young firms will prove to be problematic in the exercises below.

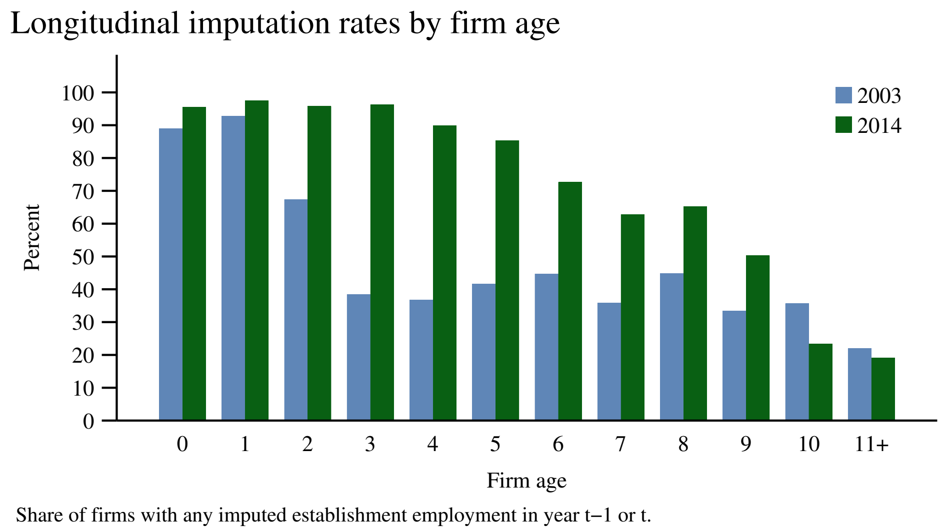


Figure 6

3.2 Sales imputation

Recent literature in firm dynamics relates firm employment growth with firm productivity (Decker, Haltiwanger, Jarmin and Miranda (2018), Alon, Berger, Dent and Pugsley (2018)) by calculating real sales per worker at the firm level.[19] Researchers may wish to study similar topics with NETS/D&B, but we find severe limitations of NETS sales data.

The sales variable in NETS is somewhat more complicated than the employment variable.[20] While a respondent may be able to report current point-in-time employment to D&B surveyors at any time, a respondent is not likely to know current-year sales at the time of contact. NETS documentation suggests that respondents are likely to report some combination of the prior year's sales and an estimate of the current year's sales. Moreover, establishment-level sales is a more complicated object than firm-level sales (indeed, Census Bureau researchers who bring sales data from the Business Register to the LBD study sales at the firm level). Therefore, reported establishment sales figures are estimates at best. For non-reported sales figures, D&B and Walls & Associates rely on imputation methods that are similar to those used for employment (described above), including reliance on firm or industry sales/employment ratios, with some additions. In particular, for multi-establishment firms, when firm-level sales are known (e.g., in the case of publicly traded firms), sales

[19]This literature follows the construction of the Revenue-Enhanced LBD (RE-LBD) by Haltiwanger, Jarmin, Kulick and Miranda (2017), which linked firm revenue data from the Census Bureau's Business Register to the LBD.

[20]This paragraph draws heavily on Walls (2008), part of NETS marketing material.

Firm size (employees)	Year 2000	2014
1 to 4	80	80
5 to 9	78	85
10 to 19	77	82
20 to 49	79	84
50 to 99	85	88
100 to 249	89	91
250 to 499	93	94
500 to 999	94	94
1,000 to 2,499	93	93
2,500 to 4,999	95	92
5,000 to 9,999	95	94
10,000+	96	94

Source: NETS
Notes: Percent of firms with imputed establishment sales data.

Table 2: Establishment sales imputation rates

are allocated among establishments using employment shares. Note that sales figures are attributed even to establishments that do not sell products or services but instead produce inputs used by other establishments within the same firm; in such cases, the establishment sales data provide no marginal information beyond the establishment employment data.

NETS does include an imputation flag for the sales variable, *salesc*, with the same coding as the employment imputation variable (*empc* described above). That is, *salesc* can take on the following values: 0 (actual figure or estimate provided by respondent), 1 (bottom of range reported by respondent), 2 (D&B estimate), or 3 (Walls & Associates estimate).[21] Imputation is common. In both the years 2000 and 2014, just under 20 percent of establishments report $salesc = 0$, indicating that the sales figure is a true reported value or respondent estimate. This likely overstates the accuracy of the figures, however, for the reasons above—even reported sales figures may be rough estimates. In any case, these establishments account for only about 10 percent of total (imputed and non-imputed) employment and total sales (imputed and non-imputed) in both years, indicating that imputation is *more* common among larger establishments. Remaining establishments are imputed, almost entirely reflecting imputation by Walls & Associates ($salesc = 3$).

Sales imputation varies widely by *firm* size. Table 2 reports *establishment* imputation rates by *firm* size bins for the years 2000 and 2014. Small firms have establishment im-

[21] We also observe an extremely small number of establishments with missing sales data and sales imputation flags.

putation rates around 80 percent, while around 95 percent of establishments of large firms have imputed data. The high imputation rates among large firms appear to be driven by firms with multiple establishments; we find that close to 95 percent of establishments of multi-establishment firms have imputed sales data, compared with about 80 percent among single-establishment firms. The interpretation of these imputation rates is not entirely clear. For example, there may be cases (particularly among publicly traded firms) where D&B receive accurate firm-level sales data, but establishment-level sales data must be imputed. Since our NETS data do not provide firm-level sales information, if we require firm sales figures we must construct them by summing across establishments within firms. So it is possible that the imputation rates we report for establishments of large firms overstate the rate of imputation of firm-level sales among large firms; that is, there may be cases where a firm's establishments have sales data imputed from total firm sales such that summing across establishments results in true firm sales figures. However, the number of firms for which D&B receive true sales data is probably small (for example, there are fewer than 5,000 publicly traded firms in the U.S.), so if there is some overstatement, it is likely to be minimal.[22] Moreover, the research for which establishment-level microdata like NETS would be most useful are likely to require geographic breakdowns of activity, in which case establishment imputation is the most relevant. In any case, establishment imputation rates are high across the firm size distribution, even among small firms that are likely to have only one establishment.

Sales data would be particularly useful for the study of productivity; however, we find large discrepancies between NETS and official data on this topic. For example, using the LBD, Decker, Haltiwanger, Jarmin and Miranda (2018) find that the within-industry dispersion (standard deviation) of sales per worker has risen in recent decades; in NETS, we find the opposite pattern. Moreover, the *level* of labor productivity dispersion is much lower in NETS than in the LBD, likely owing to the industry average rules of thumb used for NETS sales imputation. For example, Decker, Haltiwanger, Jarmin and Miranda (2018) find that among young (age less than five) high-tech firms, a firm that is one standard deviation more productive than its corresponding industry-by-year mean is about 2.5 times as productive as the mean in 1996 (the first year in which LBD sales data are available) and 3.0 times as productive in 2012. In NETS, this ratio is about 1.8 in 1996 and 1.7 in 2012.

The limitations of the sales data can be well illustrated by the case of Walmart, a large multi-establishment firm. We find remarkable correlations of sales per worker across establishments. In particular, in every year of our data, more than 90 percent of Walmart establishments have sales per worker equal to the median Walmart establishment, indicating a straight imputation of sales based on employment within the firm. This rules out the use of within-firm establishment-level sales data as a measure of activity separate from employment.

[22] Additionally, Dinlersoz, Kalemli-Ozcan, Hyatt and Penciakova (2019) find evidence that sales data in Compsutat are overstated relative to official data, perhaps due to the inclusion of international sales.

The prevalence of sales imputation—which is more common than employment imputation—and the reliance of the imputation methods on employment data imply that the marginal value of the sales data is very low. Moreover, popular business dynamics topics such as productivity dispersion, decompositions of aggregate productivity growth, or the relationship between business-level productivity and growth cannot be studied with NETS.

4 Cross-sectional static moments in NETS and official data

Barnatchez, Crane and Decker (2017) describe exhaustive cross-sectional comparisons between NETS and both CBP and QCEW. We report key highlights here with a focus on CBP comparisons. In all cases we restrict NETS data to match CBP industry scope and focus on the "employer" restriction described above.[23]

4.1 Establishment size

Table 3 (taken from Barnatchez, Crane and Decker (2017)) reports employment and establishment counts in NETS data relative to CBP data for selected years. The comparisons are reported as a percent of CBP values; for example, the table shows that in the year 2000, total NETS employment in the 1 to 4 employee size category was 82 percent higher than total CBP employment in the same category.

NETS and CBP activity are similar in many size categories. However, NETS has substantially more employment and establishments in the smallest two size cateogries, and the excess increases over time. In other words, small establishments account for the rise of NETS employer establishment counts relative to CBP counts shown on Figure 1. NETS also shows excess activity in the largest size category (establishments with at least 1000 employees), though this discrepancy declines after 2000.

The bottom panel of Table 3 reports overall discrepancies. NETS employment counts exceed CBP by 16 percent in 2000, rising to 19 percent in 2014. Establishment count discrepancies are larger, rising from 33 percent of CBP counts in 2000 to more than 100 percent in 2014. However, omitting establishments with fewer than 10 employees narrows these discrepancies significantly.

4.2 Sector

Table 4 reports NETS versus CBP employment discrepancies by NAICS sector for selected years. The largest discrepancies are seen in the agriculture sector (which may partly reflect NETS coverage of farming-related activities that are classified differently in NETS than in

[23] All of the results described in this section are similar if QCEW is used instead of CBP, as shown by Barnatchez, Crane and Decker (2017).

Size class (Employees)	Percent difference 2000 Emp.	2000 Estab.	2007 Emp.	2007 Estab.	2014 Emp.	2014 Estab.
1 to 4	82	54	113	130	196	242
5 to 9	16	15	19	18	44	41
10 to 19	5	3	9	4	1	0
20 to 49	8	6	9	6	2	−1
50 to 99	10	9	8	8	12	10
100 to 249	−1	1	−2	0	3	4
250 to 499	0	0	−1	0	−4	−3
500 to 999	9	9	−1	0	−6	−5
1000+	49	39	13	10	18	4
Aggregate	16	33	13	76	19	140
Ex. small ests	12	5	5	5	5	1
Ex. small & large ests	5	5	4	5	2	1

Source: NETS and CBP
Note: Difference between NETS and CBP employment as percent of CBP employment. NETS sample restricted to CBP scope. "Small" establishments have fewer than 10 employees; "large" establishments have at least 1000 employees.

Table 3: NETS versus CBP by establishment size

official data and therefore are not dropped in our adjustment of NETS to CBP scope) and education (which may reflect state-owned universities that are out of scope for CBP).[24] But large discrepancies also appear in many other sectors that are not subject to CBP scope; for example, in the year 2000 discrepancies of greater than 25 percent appear in mining; manufacturing; real estate, rental, and leasing; professional, scientific, and technical services; administrative and waste management; arts, entertainment, and recreation; and other services. Omitting small establishments reduces discrepancies in some, but not all, sectors.

Notably, discrepancies in manufacturing rise over time as NETS does not fully capture the decline in U.S. manufacturing employment of the 2000s;[25] and the large year-2000 discrepancy in mining disappears by 2007 as NETS does not capture the strong rise in mining employment associated with the U.S. shale oil and gas boom.

[24] Barnatchez, Crane and Decker (2017) investigate the discrepancy arising from educational services in some detail. NETS excluding small establishments and educational services mimics aggregate CBP and QCEW patterns reasonably well.

[25] From 1998 to 2014, NETS data show an 18 percent decline in U.S. manufacturing employment among employer firms; the CBP and QCEW show much larger declines of 33 percent and 32 percent, respectively.

	Percent difference					
	2000		2007		2014	
Industry	Ex Sm	All	Ex Sm	All	Ex Sm	All
11 Ag., For., Fish., Hunt	74	68	67	66	71	73
21 Mining	70	69	−2	3	−8	−5
22 Utilities	−40	−37	−39	−35	−46	−40
23 Construction	−8	3	−5	3	7	23
31-33 Manufacturing	32	34	39	43	50	54
42 Wholesale Trade	7	15	7	17	−4	8
44-45 Retail Trade	−11	−2	−11	−1	−3	4
48-49 Trans., Warehous.	15	19	−6	0	−19	−7
51 Information	15	20	15	21	10	18
52 Finance, Insurance	25	23	5	10	7	12
53 Real Est., Rent., Leas.	70	70	66	68	75	86
54 Prof., Sci., Tech. Svcs	35	39	12	17	5	17
56 Admin., Waste Mgmt	−34	−26	−39	−18	−47	−8
61 Education Svcs	287	278	252	244	261	260
62 Health, Social Asst.	5	7	−8	−2	−9	1
71 Arts, Entertain., Rec.	33	44	15	27	2	22
72 Accom., Food Svcs	−1	2	−8	−1	−14	−2
81 Other Svcs	39	41	10	25	6	29

Source: NETS and CBP
Note: Difference between NETS and CBP employment as percent of CBP employment by NAICS sector. NETS sample restricted to CBP scope. "Ex Sm" excludes establishments with fewer than 10 employees.

Table 4: NETS versus CBP by sector

4.3 Correlations

Table 5 combines insights from establishment size and industry comparisons with geographic comparisons using simple correlations. We construct employment and establishment counts in geography-size-sector cells in NETS and CBP then calculate the correlation in these counts between the two data sources. The first two rows of Table 5 report correlations of state-size-sector cells, where the second row omits establishments with fewer than 10 employees. These correlations are generally reasonable. Correlations of simple county cells are higher, perhaps reflecting offsetting effects of industry categorizations or simply the fact that economic activity is highly correlated with population, and population drives numbers in all data sources.

Cell	Exclusions	Correlation					
		2000		2007		2014	
		Emp.	Estab.	Emp.	Estab.	Emp.	Estab.
State-Size-Sector	None	0.88	0.96	0.83	0.86	0.77	0.71
State-Size-Sector	Small	0.88	0.97	0.84	0.97	0.82	0.96
County	None	0.99	0.98	1.00	0.99	1.00	0.99
Cty.-Size-Sector	None		0.94		0.87		0.73
Cty.-Size-Sector	Small		0.96		0.97		0.96

Source: NETS and CBP.
Notes: Simple correlations of cell-level employment and establishment counts, NETS and CBP. "Small" exclusions refer to the exclusion of establishments with fewer than 10 employees. NETS sample restricted to CBP scope.

Table 5: Cell-based correlations

Barnatchez, Crane and Decker (2017) also conduct these exercises at the zip code-size-sector level, which shows somewhat lower correlations (between 80 and 85 percent when small establishments are omitted); and they conduct these exercises for the QCEW (in the state-size-sector specification), where correlations are substantially lower among all establishments but only modestly lower when small establishments are omitted.

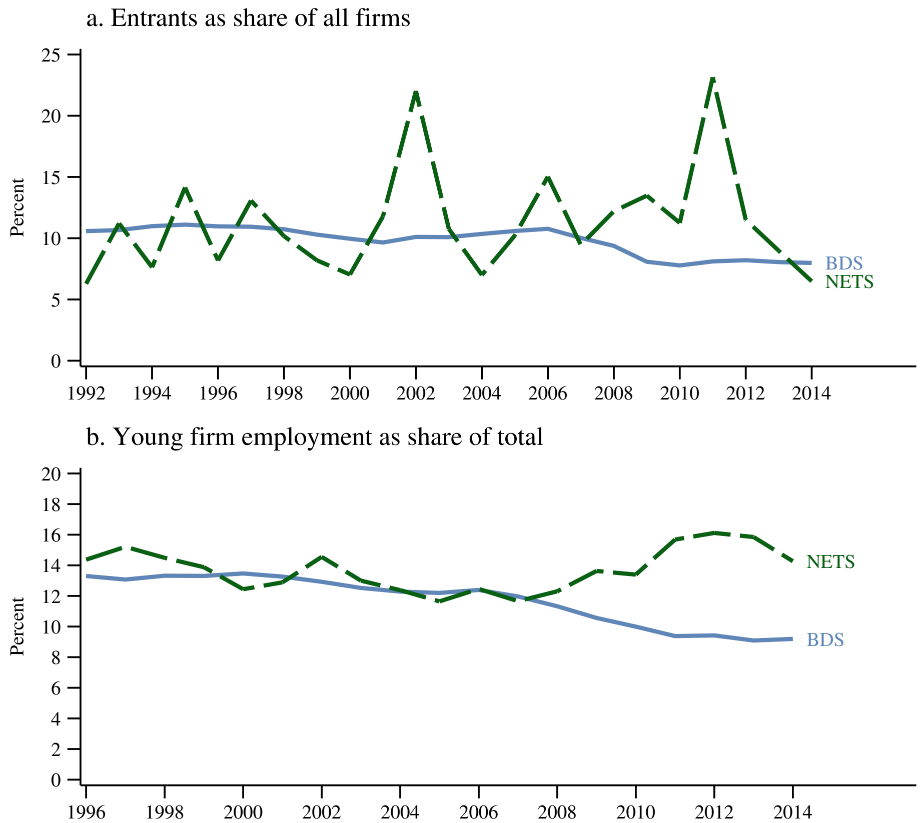
5 Business dynamics in NETS and official data

5.1 Aggregate patterns

We now turn to business dynamics. We first characterize the NETS data in terms of aggregate measures that are well known in the business dynamics literature. Panel a of Figure 7 reports the share of firms that have age zero (often referred to as the startup rate or entry rate). The dashed green line reports the entry rate from NETS, while the blue solid line reports the entry rate from the BDS. The NETS entry rate is more volatile than the official data, though in many years the NETS rate bounces around the BDS rate. NETS sees an excess surge in entry in 2002 then again in 2011, consistent with the the result from Figure 1 showing the NETS employer establishment count surged above the levels of official data after 2000, which likely reflects an expansion of D&B scope or coverage rather than true entry.

Panel b of Figure 7 broadens our study of young firm behavior to include all firms of age less than five, a cutoff that is common in the literature. Here we report the young firm *employment* share. The surge in new businesses appearing in NETS but not in the official data is readily apparent here, with a divergence starting in 2008 and the cumulative effects of differing coverage becoming notable by the end of the sample. Importantly, the well-documented decline in young firm formation and activity described in a large and growing

Entrants and young firms in BDS and NETS



Entrants have age zero; young firms have age less than five.

Figure 7

literature (e.g., Decker, Haltiwanger, Jarmin and Miranda (2014)) is reversed in NETS data due to this late-2000s divergence. While official data show young firm shares moving below 10 percent by 2010, young firm shares in NETS exceed 16 percent in 2012 and 2013, a level not seen in official data since the 1980s.

In short, while a large and growing literature explores the puzzling decline of young firm activity in official data, NETS data tell the opposite story due to (likely spurious) measured entry brought on by an apparent scope expansion.[26] The shape of the gap between total

[26]NETS marketing materials point out the rise in entry and argue that this reflects growth of self employment or gig economy work brought on by changes in the nature of entrepreneurship and the weak labor markets of the Great Recession and aftermath. As noted above, we drop firms with only one reported employee, which should roughly eliminate true nonemployers from the data. Thus, the discrepancy we observe reflects

NETS employment and total CBP employment shown on Figure 1 closely mimics the shape of the gap between NETS young firm shares and BDS young firm shares shown on Figure 7.

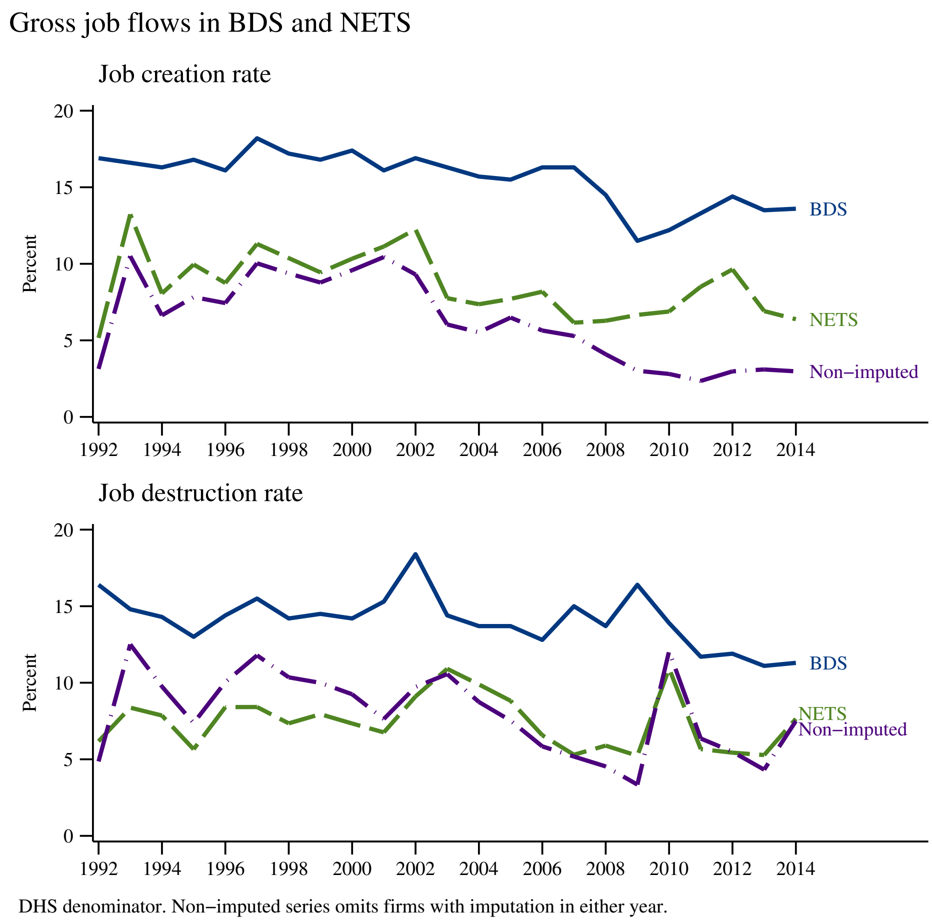


Figure 8

We next study patterns of gross job flows; first, consistent with related literature we define "job creation" as the number of jobs created by entering or expanding establishments, and we define "job destruction" as the number of jobs destroyed by exiting and downsizing establishments (these definitions are consistent with the literature). We express each of these as a rate by dividing by total employment, averaged over years t and $t-1$ in usual DHS fashion. The top panel of Figure 8 reports the job creation rate from the BDS (solid blue

apparent differences in measured employment at employer businesses. See Abraham, Haltiwanger, Sandusky and Spletzer (2018) for discussion of recent patterns of self employment and nonemployer activity in both BLS and Census Bureau data.

line), NETS (dashed green line), and NETS omitting firms with longitudinal imputation (dot-dashed purple line). The bottom panel reports corresponding job destruction rates.

In general, NETS exhibits much lower rates of gross job flows than the official data, as one might expect given the foregoing discussion of imputation and rounding. But it is somewhat surprising that the non-imputed NETS series are sometimes even lower than the overall NETS series, suggesting that imputation alone does not explain the low volatility of NETS firms. One likely reason is that, as shown on Figure 6, imputation is most prevalent among young firms. Dropping imputed firms means shifting the firm distribution heavily toward more mature firms that tend to have lower job creation rates. More precisely, new entrants necessarily contribute positively to gross job creation, no matter their initial employment (imputed or not). Other problems arise from the simple fact that dropping imputed firms significantly reduces the sample, and likely in a non-random way, so any statistics calculated on the residual sample are biased.

In any case, the patterns of gross job flows in NETS are substantially different from the BDS, both in terms of levels and in terms of time series behavior, and imputation alone does not account for these discrepancies.

5.2 Cell-based comparisons

We can drill down further by comparing detailed "cells" in the BDS and NETS. We focus on two disaggregations available in the publicly available BDS files: firm size by firm age by state, and firm size by firm age by industry. Comparing individual cells along these dimensions allows for a more complete picture of the two data sources. We focus on three critical measures of business dynamics: job creation rates, job destruction rates, and net employment growth rates. We also study simple employment levels measured as the DHS denominator (i.e., two-year employment averages).

Firm size bins, in terms of employees (based on DHS denominator), are defined as follows: 1-4, 5-9, 10-19, 20-49, 50-99, 100-249, 250-499, 500-999, 1,000-2,499, 2,500-4,999, 5,000-9,999, and 10,000 or above; these are the narrowest size bins available in the BDS. Firm age bins are defined as follows: 0, 1, 2, 3, 4, 5, 6-10, and 11 or above. The BDS provides more age detail in the 11 and above category (11-15, 16-20, 21-25, and beyond), but given the shorter time series available in NETS, we combine the 11+ categories for better coverage. All states plus the District of Columbia are used, as are all SIC sectors available in the BDS: agricultural services, forestry, and fishing (SIC 7); mining (SIC 10); construction (SIC 15); manufacturing (SIC 20); transportation and public utilities (SIC 40); wholesale trade (SIC 50); retail trade (SIC 52); finance, insurance, and real estate (SIC 60); and services (SIC 70). Therefore, the size by age by state disaggregation has potential for up to 4,896 cells, and the size by age by industry disaggregation has potential for up to 864 cells. When a cell in one data source is missing but that cell is not missing in the other data source, we populate each of job creation, job destruction, and DHS employment as zero in the missing source.[27]

Both location and industry are *establishment* characteristics, and multi-establishment firms can operate in multiple states and industries. When creating cell aggregates we implement BDS methodology, which follows.[28] A single firm's activity can appear in any industry or state cell in which that firm has establishments, but only the establishments that belong to a given cell contribute data to that cell aggregate. However, *firm* characteristics are firm-wide and apply to all of a firm's establishments. That is, firm size and firm age information are the same for all establishments of a given firm. For example, consider a firm with two establishments, one in New York (first opened in 2000) and the other in New Jersey (first opened in 2002). Suppose we observe this firm in the year 2003 and find that the New York establishment has five employees and the New Jersey establishment has ten employees. Then in 2003, the firm has firm age of three and firm size of fifteen. The employment and job flows of the New York establishment will appear in the cell defined as firm size of 10-19 employees, firm age of 3, and New York state. The employment and job flows of the New Jersey establishment will appear in the cell defined as firm size of 10-19 employees, firm age of 3, and New Jersey state. That is, the New York and New Jersey establishments appear in the same firm size and age bins since they belong to the same firm, but they appear in different states. Industry is treated in the same manner.

		Correlations			
Cells	Year	Job Creation	Job Destruction	Net	Denominator
Size-Age-State	2003	0.891	0.937	0.651	0.984
Size-Age-State	2014	0.756	0.567	0.554	0.968
Size-Age-Sector	2003	0.893	0.910	0.685	0.971
Size-Age-Sector	2014	0.735	0.671	0.598	0.966

Source: NETS, BDS

Notes: Cross-cell, unweighted Pearson correlations of BDS and NETS levels. “Denominator” is the average of employment in years $t-1$ and t

Table 6: Cell Correlations: Levels

Table 6 reports simple cell-based correlations between the BDS and NETS in terms of job creation, job destruction, net employment growth, and the DHS employment level; these correlations refer to actual levels (i.e., number of jobs created). For brevity we focus on two snapshot years, 2003 and 2014. We choose 2003 because this is the first year in which NETS is available given our firm age scheme, and we choose 2014 because it is the latest year in our data. The first two rows of Table 6 refer to the size-age-state cells. As can be seen from the first row, in 2003 the levels of job creation and job destruction were highly correlated

[27]If a cell is missing in both sources, we do not generate an empty cell to populate in both sources.

[28]We confirmed the BDS methodology through correspondence with Census Bureau staff.

between BDS and NETS, though net growth is less correlated. These correlations generally deteriorate by 2014. The correlation of employment levels, in the last column, remains extremely high throughout. The size-age-industry cell scheme shows similar results.

However, these correlations hide an underlying divergence. Figure 9 plots job creation (panel a) and job destruction (panel b) in BDS cells against NETS cells in 2003. The job creation pattern illustrates how correlations can overstate the correspondence between the two data sources; a tight linear relationship is apparent, resulting in a high correlation, but the slope of the relationship is clearly steeper than the 45-degree line (dashed red line) that would indicate perfect correspondence. That is, NETS tends to understate job creation in 2003, relative to the BDS. The job destruction pattern has a less clear story. The high correlations shown on Table 6, therefore, partly reflect the fact that NETS and BDS show roughly similar magnitudes in an ordinal sense without always matching well in actual levels. These divergences in levels also help explain the lower correlations for net job creation seen on Table 6, since net job creation is the difference between creation and destruction.

These results on levels of job flows may be of limited importance, however, since much research focuses on *rates* of job flows. We calculate cell-level job creation rates, job destruction rates, and net employment growth rates by dividing each level by overall DHS employment for the cell. We drop all firm age zero cells since, in both sources, these have job creation and employment growth rates of 200 percent and job destruction rates of 0 percent by construction. Table 7 reports these cell correlations, again for the two different disaggregation schemes and for the years 2003 and 2014. These correlations are generally quite low, again suggesting that the level correlations mostly reflect common employment scale effects, and that once things are normalized by employment the rates lack a close relationship across the data sources.

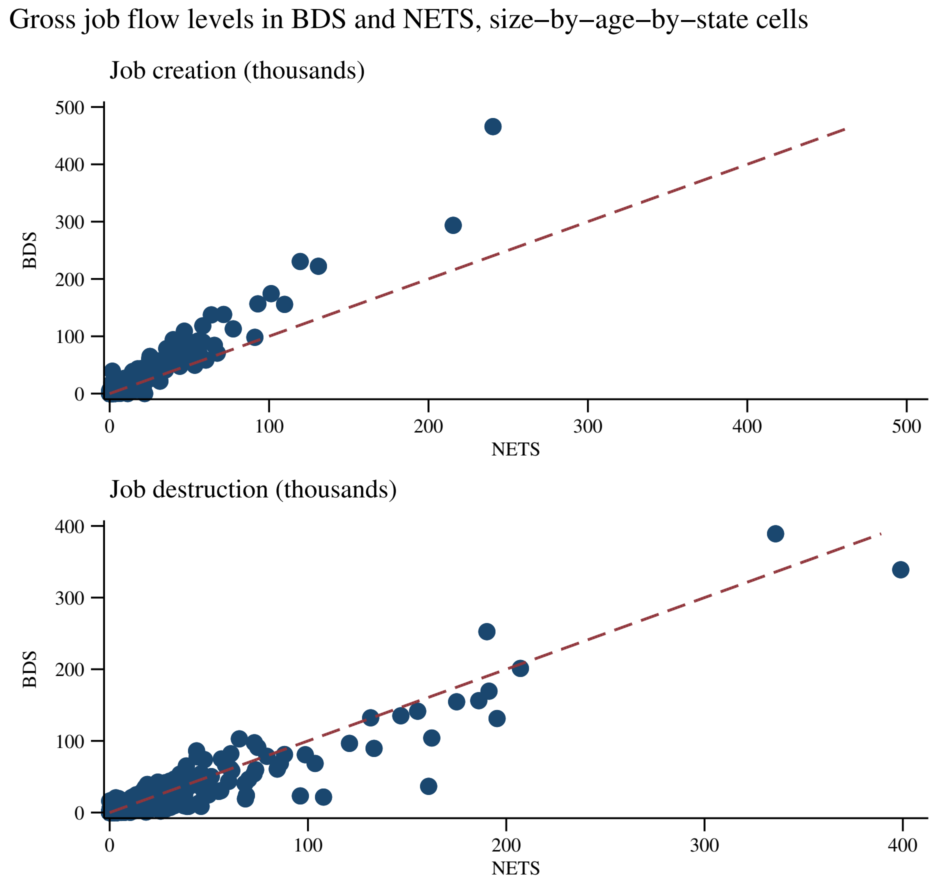
The cell-based comparisons generally support the concerns suggested by the aggregate analysis. NETS appears to have dampened rates of business dynamics compared with the BDS, and cell-level job flow rates are not strongly correlated between the two sources.

		Correlations		
Cells	Year	Job Creation	Job Destruction	Net
Size-Age-State	2003	0.000	0.233	0.139
Size-Age-State	2014	0.078	0.158	0.095
Size-Age-Industry	2003	-0.081	0.312	0.181
Size-Age-Industry	2014	0.134	0.070	0.045

Source: NETS, BDS

Notes: Cross-cell, unweighted Pearson correlations of BDS and NETS rates.

Table 7: Cell Correlations: Rates



Job flow levels in cells defined by firm size, firm age, and state.

Figure 9

5.3 Lifecycle dynamics

Many questions in firm dynamics focus on the firm lifecycle. Indeed, firm age and the behavior of young firms are at the center of many key firm dynamics questions because young firms play a disproportionate role in aggregate job growth (Haltiwanger, Jarmin and Miranda (2013)) and aggregate productivity growth (Alon, Berger, Dent and Pugsley (2018); Decker, Haltiwanger, Jarmin and Miranda (2014) and references therein). As such, accurate measurement of entry and young firm behavior is critical for any dataset used to study firm dynamics. In this section, we proceed by using NETS to investigate critical firm lifecycle patterns that have been documented by LBD-based literature.

5.3.1 Average growth

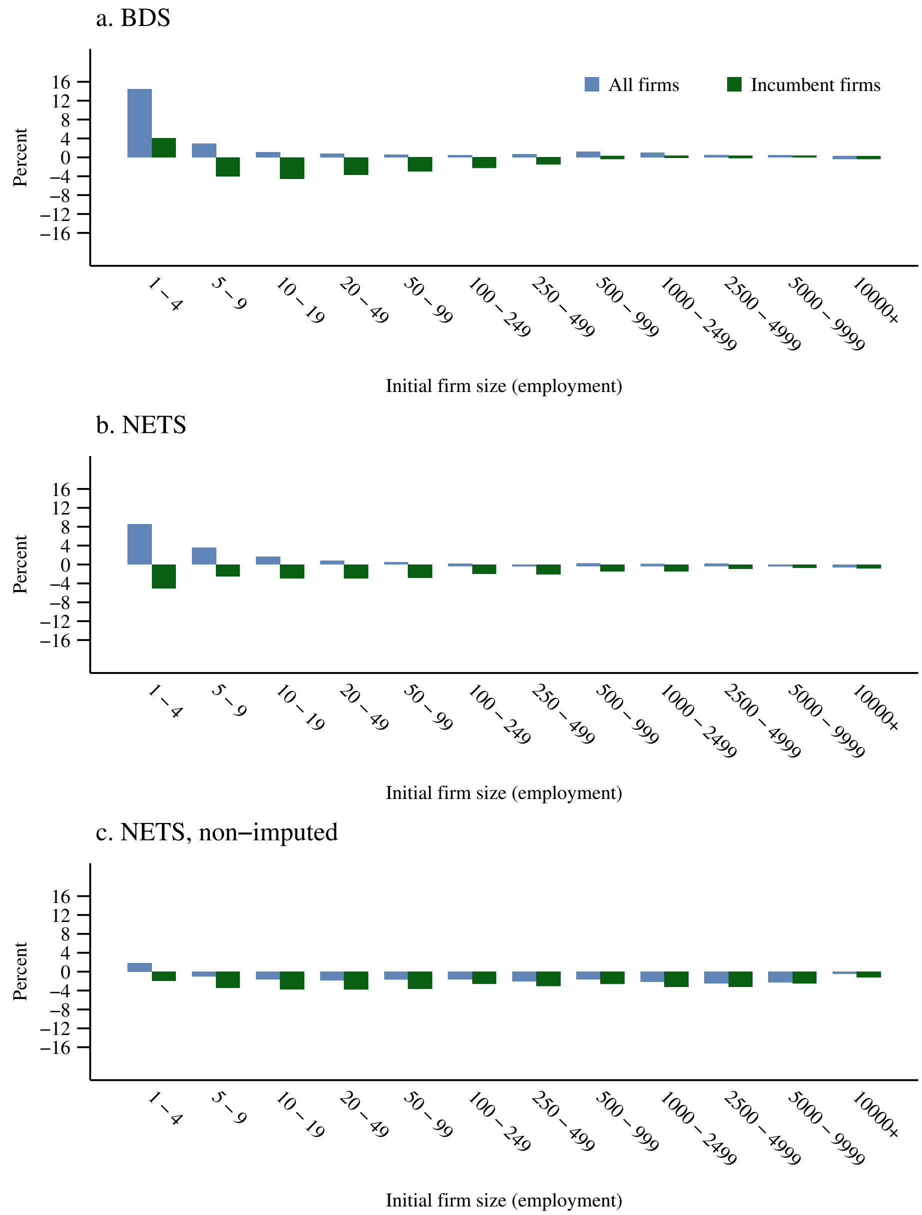
A highly cited study in empirical firm dynamics is Haltiwanger, Jarmin and Miranda (2013). Using the LBD, the authors show that the widely held view that small businesses are the primary job creators—a view reinforced by NETS-based evidence (Neumark, Wall and Zhang (2011))—was clouded by data limitations. Rather, Haltiwanger, Jarmin and Miranda (2013) show that *young* firms are the key job creators; while small businesses do create jobs disproportionately, once the econometrician controls for firm age, the small firm advantage diminishes. The empirical regularity of small firms disproportionately creating jobs arises because young firms tend to be small. The evidence that young firms are critical for job creation has motivated a wide literature seeking to better understand young firms. Here we do not replicate the specific exercises of Haltiwanger, Jarmin and Miranda (2013) but instead illustrate the concept with a simpler exercise.

Panel a of Figure 10, which relies on BDS data for 1992-2014, reports net firm employment growth rates by firm size bin, where size bins are set using *initial* firm employment and growth rates are averaged over the years in the sample.[29] Exiting firms are included (which have growth of -200 percent). The light blue bars use all firms in the BDS and illustrate the view that was common prior to Haltiwanger, Jarmin and Miranda (2013): firm growth rates decline with firm size (at least among the smaller size bins) then hover near zero for larger sizes. The dark green bars feature incumbent firms only (that is, new entrants are omitted). A starkly different picture emerges. The smallest size bin still has some growth advantage, though it is much diminished compared to the all-firm sample. Aside from the smallest class, all size classes below 500 employees actually see negative net growth on average. The figure illustrates the notion that the small-firm growth advantage is driven almost entirely by new entrants.

Panel a of Figure 10 illustrates a critical stylized fact about the firm size and age distribution, so it is important that NETS data exhibit similar properties. Panel b reports the same exercise with NETS data. Rather reassuringly, NETS results are qualitatively (though not quantitatively) similar to those seen in the BDS. Panel c repeats the same excercise omitting firms in which at least 10 percent of reported employment is at establishments with longitudinally imputed employment figures. The result is starkly different and suggests that, oddly, the ability of NETS qualitatively to replicate panel a is heavily dependent on imputed observations. In particular, it appears that much of entrants' contribution to the employment growth of the small firm bins reflects imputed employment data assigned to new firms. Indeed, as shown on Figure 6, close to 90 percent of new entrants (age 0) have imputed employment data. In 2014, of the new firms with imputed employment data, less than 1 percent reflect respondent imputation (i.e., "bottom of range"), while D&B and Walls & Associates estimates each comprise about half of imputations.

[29]Initial firm size means size in $t-1$, where growth is calculated from $t-1$ to t.

Net employment growth by firm size



DHS denominator. 1992−2014 average.
Firms with at least 10 percent of employment at imputed establishments omitted from panel c.

Figure 10

5.3.2 Skewness and churn

Haltiwanger, Jarmin and Miranda (2013) showed that young firms account for the high average growth rates of small firms. Decker, Haltiwanger, Jarmin and Miranda (2014) explore higher moments of the growth rate distribution over the lifecycle, documenting two key characteristics of young firm growth: skewness and churn. The growth outcomes of young firms are highly skewed, with a small number of extreme growth events. And young firms undergo considerable "churn": the growth outcomes of young firms are highly dispersed, with a large amount of both very positive and very negative growth events, and young firms exhibit strong "up-or-out" dynamics as high incidence of failure among some young firms coexists with rapid growth of many survivors. These characteristics of young firms are not captured by average growth statistics but instead require study of the full distribution of growth outcomes, including outcomes of survivors and the prevalence of firm exit.

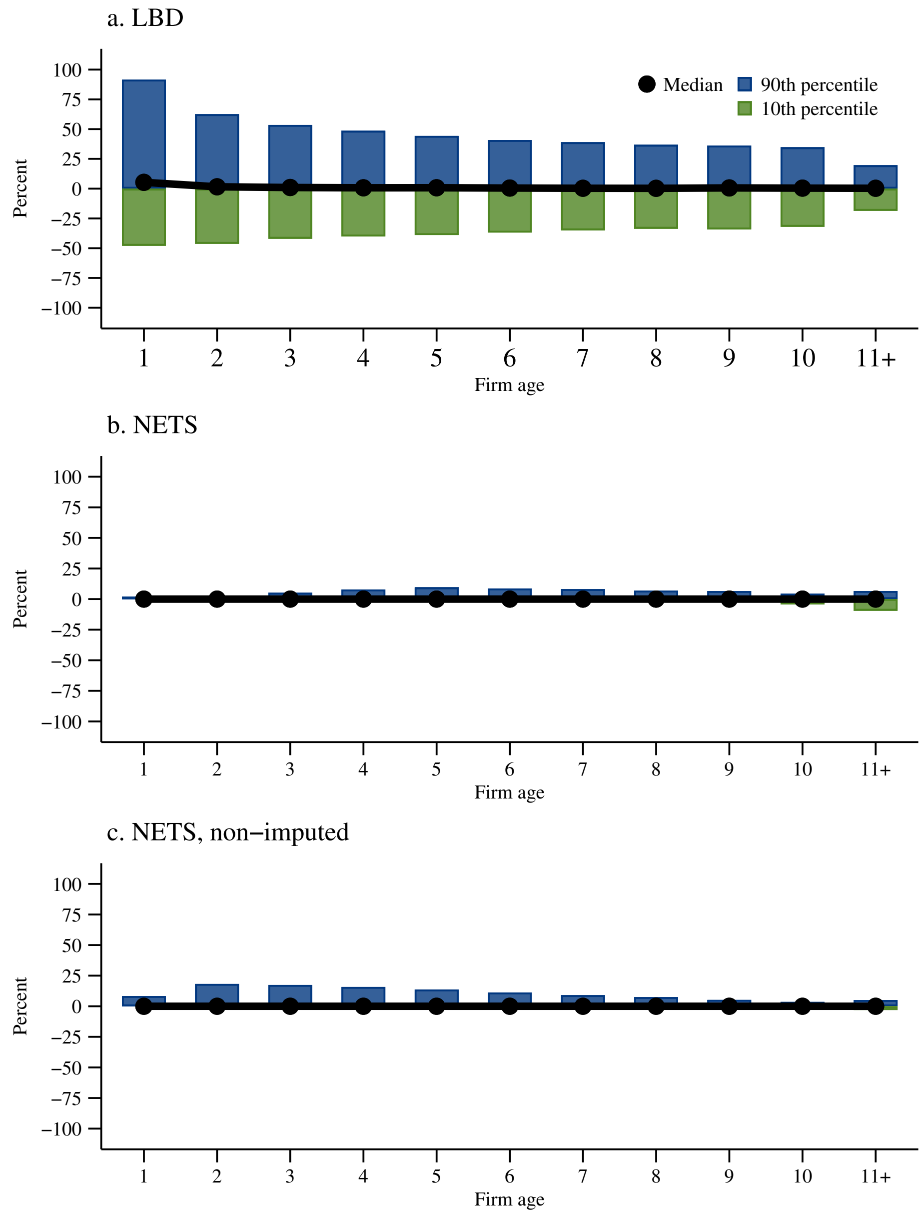
Figure 11, which is taken from Decker, Haltiwanger, Jarmin and Miranda (2014) exercises on LBD data, reports the growth rate distribution of surviving firms (i.e., those that do not exit) by age, averaged over the years 1992-2011.[30] The solid line with dots is the median of the employment-weighted growth rate distribution for the corresponding age bin; that is, for each age bin, half of all employment is at firms with growth rates at or below the black line. The top of the dark blue bars indicates the 90th percentile of the employment-weighted growth rate distribution, while the bottom of the light green bars indicates the 10th percentile of the employment-weighted growth rate distribution. Each statistic is calculated for every year in the sample, then averaged across years.[31]

A few key patterns are evident from Figure 11 (see Decker, Haltiwanger, Jarmin and Miranda (2014) for more discussion). First, median employment growth is only positive among young firms; the typical mature firm has zero employment growth, consistent with the age profiles described above. Second, growth outcomes are highly dispersed among young firms, with dispersion declining as firm cohorts age. This fact illustrates the high pace of churn among young firms, with many outcomes of both extreme growth and extreme decline. Third, the growth rate distribution of young firms is characterized by skewness, shown as the distance from the 90th percentile to the median compared with the distance from the 10th percentile to the median; this skewness illustrates that the substantial job growth contribution of young firms includes not widespread growth but in fact a few firms with extremely high

[30]Decker, Haltiwanger, Jarmin and Miranda (2014) report 16 age bins, with the top bin including all firms age 16 and above. Given the shorter time series of NETS, to improve the comparison we report only 11 age bins. Since our project lacks access to the LBD microdata, in our reproduction of the Decker, Haltiwanger, Jarmin and Miranda (2014) figure we collapse age bins 11 and higher using simple averages of the reported percentiles.

[31]The population of firms included in Figure 11 differs from the population included in Figure 10 in that 10 potentially includes all firms or all incumbents (including firms that exit, with a growth rate of -200 percent), but 11 includes only surviving firms. That is, in Figure 11, the bars corresponding with firm age 1 include firms that survived to reach age 1, omitting those that exited between ages 0 and 1.

Distribution of net employment growth rates for surviving firms



DHS denominator. Employment−weighted (DHS) distribution. 1992−2011 average. Panel a from Decker et al. (2014).
Firms with at least 10 percent of employment at imputed establishments omitted from panel c.

Figure 11

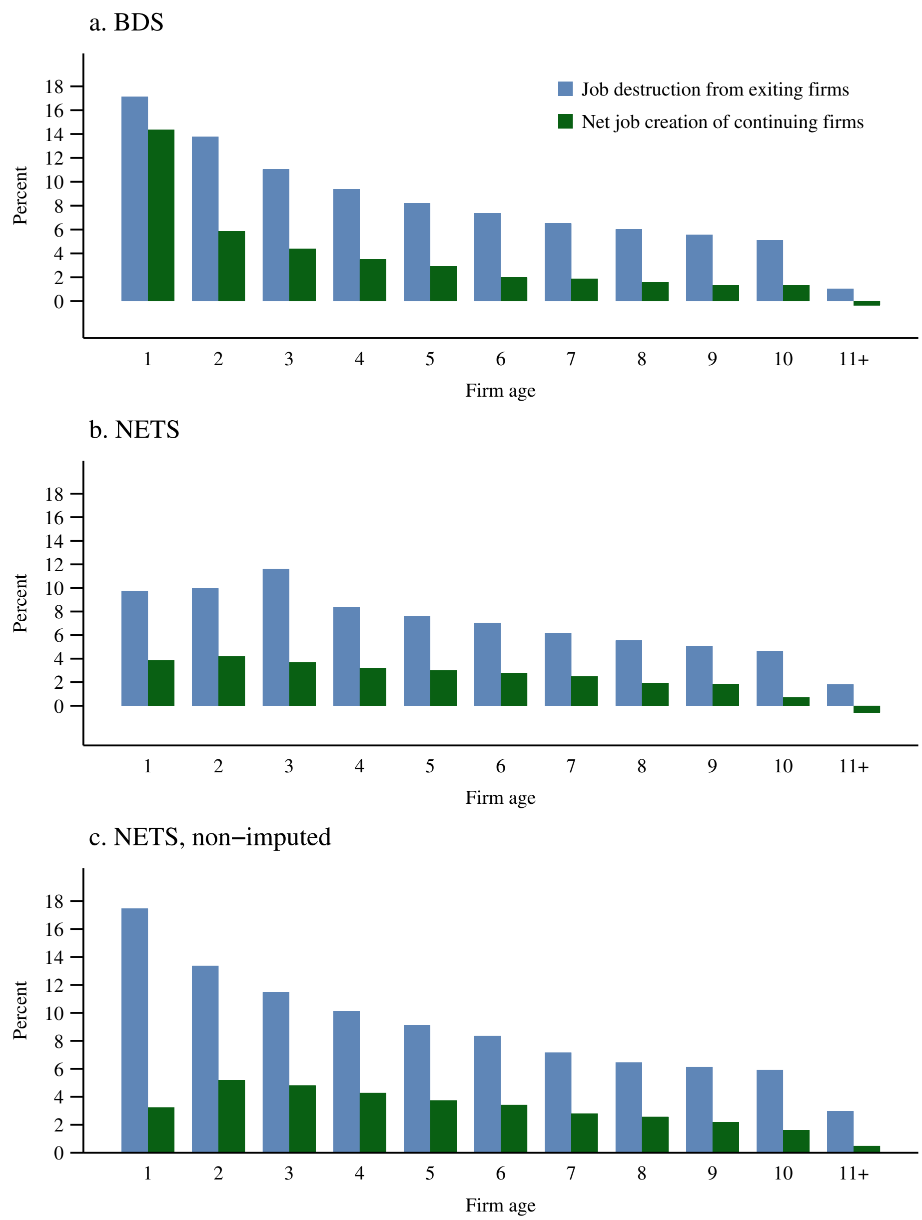
growth. Skewness disappears entirely by age five, a reason that much of the literature studies young firms with an age cutoff around five. High growth is a characteristic of (some) young firms.

As with the data on firm growth by size and age, the patterns of dispersion and skewness over the (surviving) firm lifecycle evident in panel a of Figure 11 are critical stylized facts about the behavior of young firms and the sources of aggregate employment growth. We evaluate the ability of NETS to exhibit these patterns on panel b, which mimics panel a. The difference between the figures is very concerning: While LBD data exhibit significant growth rate dispersion among firms of all ages (and particularly young firms), very little dispersion is evident in the NETS data shown on panel b. Since these are employment-weighted distributions, the latter figure indicates that 90 percent of surviving-firm employment is at firms with a growth rate around zero percent or higher for almost all age groups, while in the LBD we observe very young firms with growth approaching negative 50 percent and even many mature firms with growth around negative 25 percent. And while negative growth is nearly absent from NETS data, positive growth is almost as rare. For example, in the LBD we observe *young* firms that account for around 10 percent of employment growing at a rate of 50 percent or more, but no firm age group in NETS reports a 90th percentile growth rate beyond 20 percent. The median growth rate in NETS, shown by the black line with dots, is close to zero for firms of all ages, in contrast to the positive growth rates seen by young firms in the LBD. NETS data therefore miss virtually the entire distribution of firm growth outcomes, whatever their performance tracking average growth patterns. This is a significant limitation of NETS generally and is particularly problematic for the study of young firms, which (as shown on panel a of Figure 11) are especially characterized by wide dispersion and high skewness of firm growth rates. Panel c shows that omitting imputed observations from NETS does not materially alter Figure 11.

Decker, Haltiwanger, Jarmin and Miranda (2014) also document the "up-or-out" nature of the young firm experience by contrasting exit and survival. Panel a of Figure 12, which uses BDS data to replicate Decker, Haltiwanger, Jarmin and Miranda (2014), reports the experiences of firm cohorts as follows. The light blue bars report jobs destroyed (over one year) by firms that exit just before reaching a given age; that is, the blue bar for age 1 reflects exits of firms between age 0 and age 1, the blue bar for age 2 reflects exits of firms between age 1 and age 2, and so on. The dark green bars report net job creation (over one year) among firms that survive to a given age; that is, the green bar for age 1 reports jobs created by firms continuing from age 0 to 1, the green bar for age 2 reports jobs created by firms continuing from age 1 to 2, and so on. All figures are scaled by the DHS employment denominator for the entire cohort, and rates are calculated by year then averaged over all years 1992-2011.[32]

[32]As with the previous set of figures, we collapse all age categories above 10 into a single "11+" category, which is simple in this exercise since we rely on BDS data. We do this for comparability with NETS data but make a note of it because it differs slightly from the setup in Decker, Haltiwanger, Jarmin and Miranda (2014).

Distribution of net employment growth rates for surviving firms



DHS denominator.
Firms with at least 10 percent of employment at imputed establishments omitted from panel c.

Figure 12

Figure 12, panel a illustrates three key points about the firm lifecycle. First, both job destruction from exits and job creation from entrants are high among young firms and decline monotonically with firm age, consistent with evidence above that young firm outcomes are volatile and highly dispersed. Second, an "up-or-out" pattern is evident in the sense that, while many jobs are destroyed by exiting firms, surviving firms have high average growth rates. Third, job creation from survivors is more than offset by job destruction from exiting firms for all age groups. Note that, by construction, age zero firms (not shown on Figure 12) only create (i.e., do not destroy) jobs, so creation far offsets destruction upon entry. A reasonable characterization of young firm dynamics, then, is that each new cohort creates a large number of jobs upon entry, but firms immediately begin failing, destroying many jobs as firms age but with continued growth among surviving firms that partially offsets the job destruction.[33]

Panel b of Figure 12 mimics panel a using NETS data to assess the presence of "up-or-out" dynamics in NETS. The performance of NETS in this exercise is not as weak as in the previous skewness and dispersion exercise: the lifecycle pattern of exit-driven destruction and creation of continuers is not quite monotonic but is qualitatively similar to BDS data in that destruction from exit outpaces creation among continuers for all age groups. Moreover, among age groups above 5 the magnitudes of job destruction and creation appear reasonably accurate. However, the magnitudes illustrated by the figure indicate particularly poor measurement of *young* firm dynamics. The differences between young and mature firms, in terms of both job destruction and creation, are much smaller in NETS than in the BDS, and the monotonicity-by-age is wrong for the youngest age groups.

Panel c documents the same exercise in NETS omitting firms that have longitudinally imputed data comprising at least 10 percent of their employment; interestingly, NETS' inability to track the pattern of exit-driven job destruction among young firms shown in Figure 12 appears to be due to imputed observations; job destruction rates across the lifecycle look reasonably accurate among non-imputed firms. This suggests that imputation may prevent measured job destruction by creating persistent employment values for downsizing or even exiting firms. Moreover, job creation rates among young firms appear little better among the non-imputed observations than among NETS firms generally. Again we observe that NETS is particularly limited in its measurement of young firm dynamics, and young firm dynamics are a critical component of the overall firm dynamics literature.

[33] Decker, Haltiwanger, Jarmin and Miranda (2014) note that the post-entry job destruction of exiting firms is still not enough to completely offset the jobs created upon entry: five years after entry, the employment of the typical cohort is still equal to 80 percent of the cohort's entry employment, such that new cohorts of firms make permanent contributions to aggregate employment despite high failure rates in early years.

6 Discussion

6.1 Suggestions for researchers

Our analysis above leads to the following suggestions for researchers considering the use of NETS or D&B data:

1. Do not rely on NETS/D&B to define the business universe. Instead, link the data to outside data sources with well-defined and appropriate universe definitions (e.g., Compustat, which is the universe of publicly traded companies).

2. Focus on employment *bins* or ranges instead of precise figures; imputation, rounding, and seasonality imply that precise employment figures are likely to be inaccurate. Moreover, carefuly analyze the effects of imputation on empirical results, and be transparent about imputation prevalence.

3. Focus more on cross-sectional moments than on time series moments, and study averages rather than higher moments of the business distribution. Cell-based correlations (in terms of business size, industry, and geography) between NETS and official data are reasonably high, particularly when small establishments are omitted.

4. Do not rely on high-frequency (annual) dynamics (e.g., annual employment growth or the specific timing of business entry). Focus instead on low-frequency (e.g., 3-year) dynamics if at all.

5. Do not rely heavily on the measurement of young firm activity or the first five years of firm lifecycles when businesses tend to change significantly from year to year and timely, accurate measurement is difficult.

6. Do not use sales data, which provide negligible or no marginal information about business size or success beyond measured employment.

We consider these suggestions to be *minimum* requirements for conducting convincing research with NETS/D&B data. Measurement challenges in these data are formidable, and researchers should carefully consider whether their application is sensitive to these challenges. More broadly, it is critical to remember that these data are collected for business intelligence purposes that, while featuring certain incentives to ensure accuracy, may have different standards of accuracy than are common in academic research; and while the data have the attractive property of being collected routinely for use in business activities, they are not collected or processed in a manner that necessarily ensures precision.[34]

[34] For example, Cajner, Crane, Decker, Hamins-Puertolas, Kurz and Radler (2018) and various subsequent papers study ADP payroll microdata, which are generated automatically through paycheck issuance pro-

6.2 On the reliability of official data

One potential response to the NETS/LBD discrepancies we document here is that government data sources are also imperfect and may not have a claim on being the benchmark against which private data sources should be judged. We readily acknowledge that official sources have many limitations. For example, users of the LBD encounter problems with firm identifier longitudinal linkages, staleness of industry codes and firm organization details between census years, and lack of easily intergrated coverage of the nonemployer universe. Indeed, methods of defining firm age and organic employment growth, now widely used in empirical literature but pioneered by Haltiwanger, Jarmin and Miranda (2013) and Davis, Haltiwanger, Jarmin and Miranda (2007), are designed to minimize the errors introduced by these limitations. Barnatchez, Crane and Decker (2017) discuss limitations of official data more broadly, including the Census Bureau's County Business Patterns (which uses the same source data as the LBD), and show discrepancies between Census and Bureau of Labor Statistics (BLS) sources even when restricted to common industry scope (though these discrepancies are smaller than those between NETS and either official source).

However, there are at least two reasons to treat the official sources as authoritative. First, official data collection efforts are characterized by systematic focus on consistency and measurement best practices. For example, as noted previously, the employment data on which the LBD is based always measure employment as of March 12 of a given year; in contrast, the Dun & Bradstreet employment figures could be recorded at any point during the year, rendering them vulnerable to seasonal fluctuations. Other LBD variables are continually updated with information from administrative and survey sources, such as the annual Report of Organization survey, and Census surveys are conducted scientifically.[35] More broadly, the U.S. statistical agencies employ large staffs of experts tasked with ensuring data quality as well as active researchers exploring and performing research and development on data products.[36] These efforts are supplemented by robust exchanges between statistical agency staff and outside experts, such as those facilitated by the Federal Economic Statistics Advisory Committee (FESAC). For the purposes of D&B, scientific best practice is likely to be both excessively costly and unnecessary; for example, an estimated or imputed employment observation is often good enough for the needs of D&B clients while being much less useful for researchers of business dynamics.

Second, official sources are based in part on administrative data that are accurate by

cesses with strong mechanisms for assuring accuracy. This differs markedly from the collection process for NETS/D&B, which consists of a combination of non-scientific surveys and imputation.

[35]See https://www.census.gov/programs-surveys/cos/about.html for details about the Report of Organization survey, also known as the Company Organization Survey.

[36]For example, the Census Bureau's Center for Economic Studies employs many social scientists who actively evaluate the research uses and limitations of the LBD and other Census data products. Other Census offices have similar features, and additional quality control is performed by authorized outside researchers using the Federal Statistical Research Data Centers.

construction. The LBD source data are ultimately tax records, so the LBD represents the universe of in-scope employer businesses that are known to U.S. tax authorities—a clear and reasonable definition of business activity that contrasts with D&B's looser goal of covering a less-defined employer and nonemployer business universe with large potential for undercoverage relative to its goal (as appears to have been the case prior to the likely scope improvement of the 2000s). To the extent that NETS differs from the combined Census employer and nonemployer universe (i.e., CBP and NES) in terms of establishment coverage, it must be that NETS is either including businesses defined in some other way than taxable entities or omitting taxable businesses. Likewise, annual employment snapshots in the LBD represent data that are routinely used for administrative purposes by the IRS and the Social Security Administration, limiting the scope for inaccuracy and imputation. While some LBD establishments only receive industry code and company organization updates after the semidecennial Economic Censuses, employment data come from administrative sources. NETS, by contrast, exhibits high rates of imputation of employment data, which is particularly problematic for the study of business dynamics.

Weaknesses and limitations of the official sources notwithstanding, then, data from the statistical agencies are, in our view, best treated as more authoritative than NETS. The discrepancies between the sources are therefore cause for concern about the usefulness of NETS for business dynamics research.

7 Conclusion

NETS/D&B data can be made reasonably consistent with static, cross-sectional distributions of employment and businesses as reported in official data. We find high correlations of total employment and establishment counts in NETS and official sources across geography-by-industry cells, particularly when small establishments are omitted, so NETS data may have value for cautious cross-sectional static analysis.

But our comparisons also reveal serious discrepancies between NETS and official administrative data. The NETS/D&B universe is uncertain, with too much apparent activity to correspond with the U.S. employer universe and too little to correspond with the total U.S. business universe. Coverage of the universe appears to rise during the 2000s, raising concerns about the measurement of business entry. Coverage discrepancies between NETS and official sources (both CBP and QCEW) are largest among small establishments, which also have extremely high rates of imputation.

NETS displays patterns of young firm activity, in terms of both aggregate activity shares and the micro behavior of young firms, that differ markedly from official sources. NETS businesses generally exhibit patterns of business dynamics that are far less volatile than those seen in official sources. A key driver of these discrepancies is the high rate of imputation in NETS, particularly among young firms, most of which lack fresh data observations in typical years. But restricting the sample to omit imputed data is no panacea, as imputation is extremely prevalent among young businesses and restricting the sample to non-imputed observations creates composition effects that raise more concerns than they resolve.[37] These limitations of NETS are serious and oblige researchers use caution. Topics including reallocation, entrepreneurship, firm growth and exit, and inaction are highly vulnerable to the limitations of NETS, but other topics may be studied with careful attention to our suggestions above and to the broader principles revealed by our analysis.

We therefore urge caution in using this particular data source. By paying careful attention to the nature of the measurement error present in the data, researchers may study a range of topics using NETS/D&B. But some topics, such as business dynamics or the lifecycle behavior of young firms, are difficult to study with these data. More broadly, the measurement challenges associated with these data are formidable, and researchers using the data should take these into account in the process of research design.

While we view our results as compelling, there are many aspects of NETS that we do not investigate. NETS includes a some information on variables other than employment, industry, location, and sales, such as credit information and legal form of organization. We leave investigation of these and other variables for future research.

References

Abraham, Katharine G., John C. Haltiwanger, Kristin Sandusky, and James R. Spletzer, "Measuring the Gig Economy: Current Knowledge and Open Issues," Technical Report 24950, National Bureau of Economic Research 2018.

Alon, Titan, David Berger, Robert Dent, and Benjamin Pugsley, "Older and Slower: The Startup Deficit's Lasting Effects on Aggregate Productivity Growth," *Journal of Monetary Economics*, 2018, *93*, 68–85.

Barnatchez, Keith, Leland D. Crane, and Ryan A. Decker, "An Assessment of the National Establishment Time Series (NETS) Database," Finance and Economics Discussion Series 2017-110, Federal Reserve Board 2017.

Cajner, Tomaz, Leland Crane, Ryan Decker, Adrian Hamins-Puertolas, Christopher Kurz, and Tyler Radler, "Using Payroll Processor Microdata to Measure Aggregate Labor Market Activity," Technical Report 2018-005, Board of Governors of the Federal Reserve System 2018.

[37]In ongoing work, we are investigate the application of sampling weights to non-imputed observations.

Chava, Sudheer, Alex Oettl, Manpreet Singh, and Linghang Zeng, "The Dark Side of Technological Progress? Impact of E-Commerce on Employees at Brick-and-Mortar Retailers," 01 2018.

Cho, Clare, Patrick Mclaughlin, Eliana Zeballos, Jessica Kent, and Chris Decken, "Capturing the Complete Food Environment With Commercial Data: A Comparison of TDLinx, ReCount, and NETS Databases," Technical Bulletin 1953, USDA Economic Research Service March 2019.

Davis, Steven J., John C. Haltiwanger, and Scott Schuh, *Job Creation and Destruction*, The MIT Press, 1996.

_ , John Haltiwanger, Ron S. Jarmin, and Javier Miranda, in Daron Acemoglu, Kenneth Rogoff, and Michael W, eds., *NBER Macroeconomics Annual 2006, Volume 21*, The MIT, 2007, chapter Volatility and Dispersion in Business Growth Rates: Publicly Traded versus Privately Held Firms, pp. 107–180.

Decker, Ryan A., John Haltiwanger, Ron S. Jarmin, and Javier Miranda, "Changing Business Dynamism and Productivity: Shocks vs. Responsiveness," *NBER working paper no. 24236*, 2018.

Decker, Ryan, John Haltiwanger, Ron S. Jarmin, and Javier Miranda, "The Role of Entrepreneurship in US Job Creation and Economic Dynamism," *Journal of Economic Perspectives*, 2014, *28* (3), 3–24.

Denes, Matthew, Sabrina Howell, Filippo Mezzanotti, Xinxin Wang, and Ting Xu, "Investor Tax Credits and Entrepreneurship: Evidence from U.S. States," Technical Report 2020.

DeSalvo, Bethany, Frank Limehouse, and Shawn D. Klimek, "Documenting the Business Register and Related Economic Business Data," Technical Report, Center for Economic Studies working paper CES-WP-16-17 2016.

Dinlersoz, Emin, Sebnem Kalemli-Ozcan, Henry Hyatt, and Veronika Penciakova, "Leverage over the firm life cycle, firm growth, and aggregate fluctuations," Technical Report, Federal Reserve Bank of Atlanta 2019.

Echeverri-Carroll, Elsie and Maryann Feldman, "Chasing Entrepreneurial Firms," Technical Report 2017.

Faccio, Mara and Hung-Chia Hsu, "Politically Connected Private Equity and Employment," *The Journal of Finance*, 2017, *72* (2), 539–574.

Guzman, Jorge and Scott Stern, "The State of American Entrepreneurship: New Estimates of the Quantity and Quality of Entrepreneurship for 15 US States, 1988-2014," Working Paper 22095, National Bureau of Economic Research March 2016.

Haltiwanger, John, Ron S. Jarmin, and Javier Miranda, "Who Creates Jobs? Small versus Large versus Young," *The Review of Economics and Statistics*, 2013, *95* (2), 347–361.

_ , _ , Robert Kulick, and Javier Miranda, "High growth young firms: Contribution to job growth, revenue growth, and productivity," in "Measuring entrepreneurial businesses: Current knowledge and challenges," University of Chicago Press, 2017.

Heider, Florian and Alexander Ljungqvist, "As certain as debt and taxes: Estimating the tax sensitivity of leverage from state tax changes," *Journal of Financial Economics*, 2015, *118* (3), 684 – 712. NBER Symposium on New perspectives on corporate capital structures.

Jarmin, Ron S. and Javier Miranda, "The Longitudinal Business Database," Technical Report, Center for Economic Studies 2002.

Ma, Song, Justin Murfin, and Ryan Pratt, "Young Firms, Old Capital," Technical Report 2020.

Neumark, David, Brandon Wall, and Junfu Zhang, "Do Small Businesses Create More Jobs? New Evidence for the United States from the National Establishment Time Series," *The Review of Economics and Statistics*, 2011, *93* (1), 16–29.

_ , Junfu Zhang, and Brandon Wall, "Employment Dynamics and Business Relocation: New Evidence from the National Establishment Time Series," Working Paper 11647, National Bureau of Economic Research October 2005.

Rossi-Hansberg, Esteban, Pierre-Daniel Sarte, and Nicholas Trachter, "Diverging Trends in National and Local Concentration," Working Paper 25066, National Bureau of Economic Research September 2018.
Understanding Data in the NETS Database

***Understanding Data in the NETS Database**, Technical Report, Walls & Associates NETS marketing materials 2009.*

Walls, Donald, *"Understanding Data in the NETS Database," NETS Documentation, 2008.*