Sample Bias in Decompositions of Economic Dynamics*

Jevan Cherniwchan
McMaster University

Nouri Najjar
Western University

March 2026

Abstract

Decompositions are a common method for quantifying within- and across-agent contributions to aggregate economic dynamics. We show that the standard practice of applying decompositions to sample data yields biased estimates of these contributions, and for common sample designs, these biases can be addressed by reformulating the decomposition as an estimation problem and applying standard statistical techniques. An application to India suggests sample bias meaningfully changes our understanding of how firm dynamics contribute to productivity growth. We also demonstrate our method enables the study of settings traditionally impeded by data limitations, such as productivity and firm dynamics in Sub-Saharan Africa.

JEL Codes: C18, D24, E24, O47
Keywords: Decomposition, Sample Bias, Economic Dynamics, Firm Dynamics, Productivity

*We would like to thank Adam Lavecchia, Zach Mahone, Chris Muris, Pau Pujolas, Brandon Schaufele, Angela Zheng, and audience members at Carleton University, McMaster University, Queen's University, Western University, and the Canadian Economic Association Meetings for useful discussions and feedback, and Fatemeh Tehranikia and Meghdad Rahimian for valuable research assistance. Funding from the Social Science and Humanities Research Council of Canada (Najjar), and the Canada Research Chairs program and the Spencer Family Professorship at McMaster University (Cherniwchan) is gratefully acknowledged. The usual disclaimer applies.

1 Introduction

Decompositions are a common empirical technique for quantifying how changes in economic activity both within and across individual economic agents contribute to aggregate fluctuations in economic outcomes.[1] These methods are appealing, in part, due to their clear link between micro and macro; a typical decomposition starts with an accounting identity that specifies how aggregate changes in an outcome across two periods can be linked to both within-agent changes in outcomes and changes due to re-allocations of economic activity across agents for the universe of agents in an economy. Yet, data limitations, such as those created by statistical survey designs, mean that the full population of interest is often not observed in practice.

The purpose of this paper is twofold. First, we show that the common practice of applying a decomposition based on an accounting identity to sample data yields biased estimates of the contributions of within and across agent changes in economic activity to aggregate changes in economic outcomes, regardless of the sample design. Second, we show that, for the survey designs underlying many of the data sets used in previous research, these biases can be addressed by reformulating the decomposition as an estimation problem and applying a novel estimator. With this new estimator, we first revisit the study of firm dynamics and manufacturing productivity growth in India. Doing so challenges previous findings: our estimates suggest that the reallocation of economic activity across firms has played a much larger role in the Indian economy than previously thought. Second, we show our estimator can be used to overcome the data limitations that have previously impeded the use of decompositions. We do so by providing novel evidence of firm dynamics and productivity growth in Sub-Saharan Africa.

The starting point for our analysis is the pioneering decomposition proposed by Baily et al. (1992) (hereafter, BHC).[2] This "accounting" decomposition starts with an identity

[1] This approach is, perhaps, most commonly associated with the direct study of firm dynamics and aggregate productivity (e.g. Baily et al. (1992); Griliches and Regev (1995); Foster et al. (2001); Melitz and Polanec (2015)) or with its use as a tool in the study of related questions in macroeconomics (e.g. Lentz and Mortensen (2008); Acemoglu et al. (2018)), international trade (e.g. Pavcnik (2002); Bernard et al. (2003)), and industrial organization (e.g. Van Biesebroeck (2003); Backus (2020); Barwick et al. (2025)). These methods have also been used throughout economics, including in the study of labor markets (e.g. Autor et al. (2020)), development (e.g. Chari et al. (2021)), the environment (Cherniwchan et al. (2017); Najjar and Cherniwchan (2021)), health (e.g. Chandra et al. (2016, 2024); Bloom et al. (2025)) and inequality (e.g. Gomez (2023)).

[2] We focus on the BHC decomposition for expositional convenience as it is the basis of much of the subsequent literature. However, the biases we identify also arise with variants of this decomposition, such as those proposed by Griliches and Regev (1995), Haltiwanger (1997) and Foster et al. (2001), and with alternative decompositions such as Olley and Pakes (1996) and Melitz and Polanec (2015). We discuss this

that states that an aggregate outcome of interest at time t, Y_t, can be written as a share-weighted sum of outcomes from the universe of agents, $\mathbf{U}_t$:

$$Y_t = \sum_{i \in \mathbf{U}_t} s_{it} x_{it} \tag{1}$$

where s_{it} and x_{it} are agent i's share and outcome, respectively, and i's share is calculated as $s_{it} = q_{it} / \sum_{i \in \mathbf{U}_t} q_{it}$ where q_{it} is a measure of economic importance (such as output, employment, etc.). Following BHC, the change in the aggregate outcome of interest across two periods t_0 and t_1 can be decomposed as:

$$\Delta Y_{t_1} = \sum_{i \in \mathbf{C}} s_{it_0} \Delta x_{it_1} + \sum_{i \in \mathbf{C}} \Delta s_{it_1} x_{it_1} + \sum_{i \in \mathbf{E}} s_{it_1} x_{it_1} - \sum_{i \in \mathbf{L}} s_{it_0} x_{it_0} \tag{2}$$

where $\Delta Z_{t_1} = Z_{t_1} - Z_{t_0}$, $\mathbf{C} = \{i | i \in \mathbf{U}_{t_1} \text{ and } i \in \mathbf{U}_{t_0}\}$ is the set of continuing agents, $\mathbf{E} = \{i | i \in \mathbf{U}_{t_1} \text{ and } i \notin \mathbf{U}_{t_0}\}$ is the set of agents that enter at t_1, and $\mathbf{L} = \{i | i \notin \mathbf{U}_{t_1} \text{ and } i \in \mathbf{U}_{t_0}\}$ is the set of agents that exit after t_0. The first term on the right hand side of Equation (2) is the *within-effect* that captures the change in Y between t_0 and t_1 created by changes in the outcomes of continuing agents, holding their relative shares fixed. The second term is the *between-effect* that captures the change in Y from changes in continuing agent shares, holding agent outcomes fixed. The last two terms are the *entry* and *exit effects*, respectively, that are often jointly referred to as the *selection* or *net-entry effect*; these terms capture changes in Y from agent entry and exit.

When the complete set of agents in both populations, $\mathbf{U}_{t_1}$ and $\mathbf{U}_{t_0}$, are observed, computing each component of Equation (2) is straightforward; the researcher calculates the within and between effects by summing over the set of continuing agents that are observed in both periods, and calculates the exit and entry effects by summing over the set of agents that exit after period t_0 and the set of agents that enter in period t_1.

Applying an accounting decomposition such as Equation (2) to sample data creates bias for two reasons that we term *misclassification bias* and *misweighting bias*, respectively. Misclassification bias arises because an agent's true membership in **C**, **E**, or **L** is not observed by the researcher. With a sample, if an agent is observed at t_0, but not at t_1, then the observation could reflect a true economic exit, or the agent could continue to exist, but simply not be sampled in t_1. In the latter case, the continuing agent's outcome in t_0 will be misclassified and included in the calculation of the exit effect, while the change in the agent's outcome and economic share across periods will be excluded from the calculation of the within and between effects. If, instead, a continuing agent is not

issue further in Section 2.

observed at t_0, but is observed at t_1, its outcome at t_1 will be misclassified and included in the entry effect. The change in the agent's outcome and economic share will again not be included in the calculation of the within and between effects.

Misweighting bias arises because some agents are never observed in either **C**, **E**, or **L**.[3] When an agent is not observed, their outcome and economic share will not be included when calculating the within and between effects, the exit effect or the entry effect. We term this misweighting because each component of the accounting decomposition is a weighted sum of some variable for a subset of the population; if an agent is not observed, it is "misweighted" and effectively assigned a weight of zero in this calculation.

We begin our analysis by formally showing that misclassification and misweighting lead to biased estimates of the terms on the right-hand side of Equation (2) when it is applied to sample data. We also show that different samples will produce different estimates of each component of the decomposition, reflecting sample uncertainty. Furthermore, our estimates from Monte Carlo exercises using simulated data suggest that the bias from applying an accounting decomposition to sample data may be substantial.

The biases created by misclassification and misweighting have not been given a formal treatment in previous work. Moreover, our theoretical results suggest that the approaches used in the few studies that have tried to address the biases created by misclassification and misweighting fail to do so.[4] The response to misclassification, for example, has typically been to either forgo the study of entry and exit and focus on within and between effects (e.g. Van Biesebroeck (2005b); Bollard et al. (2013)). Our theoretical results suggest that even if entry and exit are ignored, the use of sample data will still result in biased estimates of the within and between effects. Some have responded to misweighting by employing sample weights observed in each period in the calculation of the decomposition (e.g. Griliches and Regev (1995); Bollard et al. (2013); Harrison et al. (2013)). This approach still results in biased estimates, as the appropriate weight must jointly account for the sampling routine utilized in both periods.

To make progress in addressing these biases, we reformulate the decomposition as an estimation problem and propose a novel estimator based on the work of Horvitz and Thompson (1952). In doing so, we focus our attention on a class of survey designs in

[3] A third source of bias is the fact that the units that are observed are given the wrong economic share. We relegate discussion of this source of bias to the Online Appendix because, as we show below, Equation (2) is biased even if the true economic shares are observed. Our estimator also accounts for this source of bias.

[4] It is worth noting that some researchers have also highlighted some potential challenges arising from the use of accounting decompositions with sample data without producing formal analytical results. For example, Li and Rama (2015) discuss the potential biases that could arise from ignoring micro-enterprises in the Melitz and Polanec (2015) decomposition.

which all agents in the population of interest are sampled with a non-zero probability.[5] The canonical Horvitz and Thompson (HT) estimator produces an unbiased estimate of a population total from sample data by re-weighting the elements of the dataset using their first-order inclusion probability, that is, the probability with which each observation is sampled. We extend this logic to our setting and obtain unbiased estimates of the within and between effects by re-weighting observations by their second-order inclusion probabilities, that is, the joint probability that a continuing agent is observed in both t_0 and t_1 given the sampling design. We then use these within- and between-effect estimates and HT-derived estimates of the change in the aggregate outcome to obtain an unbiased estimate of the selection effect. Using this approach, we also obtain estimates of the variance of each effect, which allows for the possibility of hypothesis testing. Further Monte Carlo exercises validate our estimators.

As the final step in our analysis we use our estimation approach to study how firm dynamics have contributed to productivity growth in several developing countries. Whether aggregate productivity gains are due to productivity improvements within firms or the reallocation of economic activity from low to high productivity firms is a long-standing question (Bartelsman and Doms, 2000; Foster et al., 2001; Syverson, 2011), particularly for developing countries (Tybout, 2000; Li and Rama, 2015; McMillan and Zeufack, 2022). The measurement of these channels has been complicated by the fact that the full population of interest is not observed in many countries due to how the relevant statistical agencies collect data. Here, we illustrate the utility of our estimation approach via two applications. We first use it to revisit the study of firm dynamics and aggregate productivity growth in India's manufacturing sector, a setting in which the statistical agency collects firm-level data using a survey.[6] Second, we apply it to a setting where the use of decompositions has proved difficult due to issues related to the availability and reliability of firm-level data: Sub-Saharan Africa.[7]

For our first application, we use data from India's Annual Survey of Industries, which is a survey of formal firms in India's manufacturing sector, over the period 1998-2015. Here, we compare estimates of the within, between and selection effects obtained from

[5] We center our attention on this class of designs as they feature prominently in many of the data sets that have been used in previous work (e.g. Griliches and Regev (1995), Disney et al. (2003), Bartelsman et al. (2009), Brandt et al. (2012), Bloom et al. (2025)).

[6] This is also an attractive setting because the productivity of this sector has been studied widely (e.g. Hsieh and Klenow (2009), Allcott et al. (2016), Martin et al. (2017), Boehm and Oberfield (2020), and Bau and Matray (2023)), including via decompositions (e.g. Bollard et al. (2013), and Harrison et al. (2013)).

[7] These limitations mean previous research has examined continuing firms (e.g. Van Biesebroeck (2005b)) and firm entry and exit (e.g. Frazer (2005); Söderbom et al. (2006)) in isolation. Other related work has decomposed aggregate productivity growth using industry level data instead of data from firms (e.g. McMillan et al. (2014)).

our proposed HT estimator for both total factor productivity (TFP) and labor productivity with those obtained from two approaches akin to those used elsewhere in the literature that fail to address sample bias: the application of Equation (2) with sample weights, and the direct application of Equation (2) without weighting.[8] The results from this exercise illustrate that failing to account for sample bias lead to drastically different estimates of how firm dynamics affect aggregate productivity. For example, our estimates for TFP suggest that, on average, the application of Equation (2) with sample weights overstates the annual within effect by 28%, understates the annual between effect by 345%, and overstates the annual selection effect by 568%. Results from our HT estimator suggest that the reallocation of economic activity across firms accounts for 90% of India's productivity growth during our period of study.

For our second application, we use our proposed HT estimator and firm-level data on TFP and other firm characteristics from the World Bank Enterprise Survey (WBES) (World Bank, 2025b) to obtain novel estimates of the within, between, and selection effects for 17 countries in Sub-Saharan Africa over a variety of sample periods.[9] For most of the countries in our sample, there are no previous estimates of how firm dynamics have contributed to aggregate productivity growth. For this application we also produce standard errors for each estimate to allow for statistical inference. These results suggest that most countries we study either experienced no statistically significant growth in aggregate productivity or aggregate productivity declined. Strikingly, many of our estimates of within, between and selection effects are both negative and statistically different from zero, suggesting that in many African countries –in contrast to what has been documented in most developing countries– existing firms have been getting less productive, and economic activity has been reallocated from relatively productive to relatively unproductive firms. These results highlight how our proposed HT estimator can facilitate the use of decompositions in settings where data limitations have otherwise impeded the use of these methods.

Altogether, this paper makes two main contributions to the literature. Our primary contribution is methodological and adds to the large body of work that uses decompositions to study how agent-level changes in economic activity impact aggregate economic

[8] As we describe further in Section 4, we estimate TFP following the approach of Ackerberg et al. (2015), and measure labor productivity as value added per worker.

[9] The WBES are nationally representative firm level surveys, and they, and their precursors, have featured prominently in the study of questions related to firm dynamics and productivity in developing countries (e.g. Frazer (2005); Van Biesebroeck (2005a,b); Söderbom et al. (2006); Asker et al. (2014)). Our sample contains data from the WBES for Angola, Botswana, Cameroon, Cote d'Ivoire, the Democratic Republic of the Congo, Ethiopia, Ghana, Kenya, Mali, Nigeria, Rwanda, Senegal, Sierra Leone, South Africa, Tanzania, Uganda, and Zambia.

outcomes. Of this literature, our work is most closely related to the methodological work focused on the development of alternative accounting decompositions to refine the measurement of the relative importance of reallocations of economic activity across agents and within-agent changes in economic activity (Baily et al., 1992; Griliches and Regev, 1995; Olley and Pakes, 1996; Haltiwanger, 1997; Foster et al., 2008; Petrin and Levinsohn, 2012; Melitz and Polanec, 2015; Gomez, 2023). Our work also seeks to improve measurement, but its novelty stems from instead formally highlighting the potential biases from applying accounting decompositions to sample data, and from developing an estimation approach for correcting these biases. By focusing on the application of existing decompositions, our work also relates to that of Van Biesebroeck (2008a,b) and Bartelsman et al. (2009) who each highlight other measurement considerations in the use of decompositions.

Our secondary contribution is to the large body of work that seeks to understand the linkages between firm dynamics and aggregate productivity growth around the world. For many economies, the application of accounting decompositions is straightforward because the entire population of interest is observed.[10] However, for many countries (such as the United Kingdom (Disney et al., 2003; Bloom et al., 2025), Canada (Baldwin and Gu, 2006), China (Brandt et al., 2012), India (Bollard et al., 2013; Harrison et al., 2013) and the Netherlands and Brazil (Bartelsman et al., 2009), for example) the survey designs used by statistical agencies mean that only a subset of the population of interest is sampled. As our empirical applications demonstrate, our approach improves the study of firm dynamics in such settings, facilitating a richer understanding of aggregate productivity growth worldwide. ‘

The remainder of this paper is organized as follows. Section 2 illustrates the biases that arise when an accounting decomposition, such as Equation (2), is applied to changes in an aggregate outcome using sample data, and provides Monte-Carlo evidence of the potential magnitudes of these biases. Section 3 outlines our estimator, and provides Monte-Carlo evidence as to its performance. Section 4 presents the results from our applications to India and Sub-Saharan Africa. Finally, Section 5 concludes.

2 The Bias From Using Sample Data

As our discussion above highlights, the application of an accounting decomposition such as Equation (2) to understand the change in an aggregate outcome Y between two

[10]For example, the population is observed in studies of the manufacturing sector in the United States (Foster et al., 2008), Taiwan (Aw et al., 2001), and Slovenia (Melitz and Polanec, 2015).

periods t_0 and t_1 is straightforward when the full set of agents in both populations, $\mathbf{U}_{t_1}$ and $\mathbf{U}_{t_0}$ is observed. We now turn to formally consider the consequences of applying an accounting decomposition when this is not the case.

To this end, suppose the researcher applies the decomposition given by Equation (2) to data comprised of samples in periods t_0 and t_1, respectively. That is, suppose the researcher has access to data defined by the sets $\mathbf{D}_{t_0} \subset \mathbf{U}_{t_0}$ and $\mathbf{D}_{t_1} \subset \mathbf{U}_{t_1}$. In this case, the decomposition captures the following:

$$\Delta Y_{t_1}^D = \sum_{i \in \tilde{\mathbf{C}}} s_{it_0} \Delta x_{it_1} + \sum_{i \in \tilde{\mathbf{C}}} \Delta s_{it_1} x_{it_1} + \sum_{i \in \tilde{\mathbf{E}}} s_{it_1} x_{it_1} - \sum_{i \in \tilde{\mathbf{L}}} s_{it_0} x_{it_0} \tag{3}$$

where $Y_t^{D_t} = \sum_{i \in \mathbf{D}_t} s_{it} x_{it}$, $\tilde{\mathbf{C}} = \{i | i \in \mathbf{D}_{t_0} \text{ and } i \in \mathbf{D}_{t_1}\}$ is the set of continuing agents in the sample, $\tilde{\mathbf{E}} = \{i | i \notin \mathbf{D}_{t_0} \text{ and } i \in \mathbf{D}_{t_1}\}$ is the set of agents that enter the sample at t_1, and $\tilde{\mathbf{L}} = \{i | i \in \mathbf{D}_{t_0} \text{ and } i \notin \mathbf{D}_{t_1}\}$ is the set of agents that exit the sample after t_0. The four terms on the right hand side of Equation (3) are the sample within (WE), between (BE), entry (EE), and exit (LE) effects. The biases arising from the application of an accounting decomposition to sample data can be determined by comparing each of these terms to the analogous term in Equation (2). However the nature of the bias will depend on the underlying properties of the sample.

To see this, suppose the sample drawn by the researcher is such that $\tilde{\mathbf{C}} \subset \mathbf{C}$, $\tilde{\mathbf{E}} \subset \mathbf{E}$, and $\tilde{\mathbf{L}} \subset \mathbf{L}$ and let a_{it} be an indicator for any observation sampled at t such that $a_{it} = 1$ if and only if $i \in \mathbf{D}_t$. It is useful to then rewrite Equation (3) as:

$$\Delta Y_{t_1}^D = \sum_{i \in \mathbf{C}} a_{it_0} a_{it_1} s_{it_0} \Delta x_{it_1} + \sum_{i \in \mathbf{C}} a_{it_0} a_{it_1} \Delta s_{it_1} x_{it_1} + \sum_{i \in \mathbf{E}} a_{it_1} s_{it_1} x_{it_1} - \sum_{i \in \mathbf{L}} a_{it_0} s_{it_0} x_{it_0} \tag{4}$$

Given that s_{it} and x_{it} are fixed characteristics of observation i and a_{it} is determined by sampling, a_{it_0} and a_{it_1} are the only random variables in this expression. As such, taking expectations yields:

$$\begin{aligned}
\mathbb{E}\left[\Delta Y_{t_1}^D\right] &= \sum_{i \in \mathbf{C}} \mathbb{E}\left[a_{it_0} a_{it_1}\right] s_{it_0} \Delta x_{it_1} + \sum_{i \in \mathbf{C}} \mathbb{E}\left[a_{it_0} a_{it_1}\right] \Delta s_{it_1} x_{it_1} \\
&\quad + \sum_{i \in \mathbf{E}} \mathbb{E}\left[a_{it_1}\right] s_{it_1} x_{it_1} - \sum_{i \in \mathbf{L}} \mathbb{E}\left[a_{it_0}\right] s_{it_0} x_{it_0} \\
&= \sum_{i \in \mathbf{C}} \Pr(i \in \mathbf{D}_{t_0} \cap \mathbf{D}_{t_1}) s_{it_0} \Delta x_{it_1} + \sum_{i \in \mathbf{C}} \Pr(i \in \mathbf{D}_{t_0} \cap \mathbf{D}_{t_1}) \Delta s_{it_1} x_{it_1} \\
&\quad + \sum_{i \in \mathbf{E}} \Pr(i \in \mathbf{D}_{t_1}) s_{it_1} x_{it_1} - \sum_{i \in \mathbf{L}} \Pr(i \in \mathbf{D}_{t_0}) s_{it_0} x_{it_0}
\end{aligned} \tag{5}$$

Equation (5) indicates that the expected value of the accounting decomposition in this case depends on the probability with which agents in the true sets of continuing, entering and exiting agents are observed in the sample the decomposition is applied to. It is worth noting that for any i, $\Pr(i \in \mathbf{D}_t)$ will depend on the underlying sample design and may be individual and time specific. Subtracting Equation (5) from Equation (2) immediately yields the following:

Proposition 1. *Suppose the researcher has access to sample data given by the set $\boldsymbol{D}_{t_0} \cup \boldsymbol{D}_{t_1}$ where $\boldsymbol{D}_{t_0} \subset \boldsymbol{U}_{t_0}$ and $\boldsymbol{D}_{t_1} \subset \boldsymbol{U}_{t_1}$, and suppose the sample design is such that $\tilde{\boldsymbol{C}} \subset \boldsymbol{C}$, $\tilde{\boldsymbol{E}} \subset \boldsymbol{E}$, and $\tilde{\boldsymbol{L}} \subset \boldsymbol{L}$. Then the within, between, entry and exit effects computed by applying Equation* (2) *to the sample are biased. The bias of each effect is given by:*

1. *Within:* $\sum_{i \in \boldsymbol{C}} \left[\Pr(i \in \boldsymbol{D}_{t_0} \cap \boldsymbol{D}_{t_1}) - 1\right] s_{it_0} \Delta x_{it_1}$,
2. *Between:* $\sum_{i \in \boldsymbol{C}} \left[\Pr(i \in \boldsymbol{D}_{t_0} \cap \boldsymbol{D}_{t_1}) - 1\right] \Delta s_{it_1} x_{it_1}$,
3. *Entry:* $\sum_{i \in \boldsymbol{E}} \left[\Pr(i \in \boldsymbol{D}_{t_1}) - 1\right] s_{it_1} x_{it_1}$,
4. *Exit:* $-\sum_{i \in \boldsymbol{L}} \left[\Pr(i \in \boldsymbol{D}_{t_0}) - 1\right] s_{it_0} x_{it_0}$.

Proof. Follows from subtracting Equation (2) from Equation (5). □

As Proposition 1 shows, the application of an accounting decomposition to sample data results in biased estimates of the within and between effects when only a subset of continuing agents are observed, and biased estimates of the entry and exit effects when subsets of the sets of entering or exiting agents are observed, respectively. In each case bias arises due to what we term "misweighting." Misweighting occurs because each effect is computed as a weighted sum over the relevant population; with a sample, agents that are not observed are "misweighted" and effectively assigned a weight of zero in this calculation.

Of course, the sample design underlying Proposition 1 is unrealistic and is unlikely to be encountered in practice. In the settings where decompositions have been applied to sample data, the underlying sample design is such that a sampled agent's continuing status can be determined by simply observing the agent at both t_0 and t_1, but the researcher is unlikely to know whether an agent that enters the data and is first observed at t_1 is a true economic entry, or leaves the data after t_0 is a true economic exit. This creates an issue we refer to as "misclassification" whereby some continuing agents are misclassified as entering or exiting because they are not observed at both t_0 and t_1.

We now consider this case and suppose the sample possessed by the researcher is subject to misclassification, meaning the true status of entering and exiting agents is not

observed. That is, we again suppose that $\tilde{\mathbf{C}} \subset \mathbf{C}$, but now assume $\tilde{\mathbf{E}} \subset \mathbf{E} \cup \{\mathbf{C} \backslash \tilde{\mathbf{C}}\}$, and $\tilde{\mathbf{L}} \subset \mathbf{L} \cup \{\mathbf{C} \backslash \tilde{\mathbf{C}}\}$. Equation (3) can then be rewritten as:

$$\begin{aligned}\Delta Y_{t_1}^D = &\sum_{i \in \mathbf{C}} a_{it_0} a_{it_1} s_{it_0} \Delta x_{it_1} + \sum_{i \in \mathbf{C}} a_{it_0} a_{it_1} \Delta s_{it_1} x_{it_1} \\ &+ \sum_{i \in \mathbf{E}} a_{it_1} s_{it_1} x_{it_1} + \sum_{i \in \mathbf{C}} a_{it_1} [1 - a_{it_0}] s_{it_1} x_{it_1} \\ &- \sum_{i \in \mathbf{L}} a_{it_0} s_{it_0} x_{it_0} - \sum_{i \in \mathbf{C}} [1 - a_{it_1}] a_{it_0} s_{it_0} x_{it_0} \quad (6)\end{aligned}$$

where, as before, the first two terms on the right hand side of the equation are the sample within and between effects, but the sample entry effect is now divided into the third and fourth terms of the equation, while the sample exit effect is divided into the fifth and sixth terms. Again taking expectations yields:

$$\begin{aligned}\mathbb{E}\left[\Delta Y_{t_1}^D\right] = &\sum_{i \in \mathbf{C}} \Pr\left(i \in \mathbf{D}_{t_0} \cap \mathbf{D}_{t_1}\right) s_{it_0} \Delta x_{it_1} + \sum_{i \in \mathbf{C}} \Pr\left(i \in \mathbf{D}_{t_0} \cap \mathbf{D}_{t_1}\right) \Delta s_{it_1} x_{it_1} \\ &+ \sum_{i \in \mathbf{E}} \Pr\left(i \in \mathbf{D}_{t_1}\right) s_{it_1} x_{it_1} + \sum_{i \in \mathbf{C}} \left[\Pr\left(i \in \mathbf{D}_{t_1}\right) - \Pr\left(i \in \mathbf{D}_{t_0} \cap \mathbf{D}_{t_1}\right)\right] s_{it_1} x_{it_1} \\ &- \sum_{i \in \mathbf{L}} \Pr\left(i \in \mathbf{D}_{t_0}\right) s_{it_0} x_{it_0} - \sum_{i \in \mathbf{C}} \left[\Pr\left(i \in \mathbf{D}_{t_0}\right) - \Pr\left(i \in \mathbf{D}_{t_0} \cap \mathbf{D}_{t_1}\right)\right] s_{it_0} x_{it_0} \quad (7)\end{aligned}$$

The issue created by misclassification can be seen immediately from the last two lines of Equation (7). For example, consider the sample entry effect given on the second line of the equation. As the first term shows, as in the case when only a subset of the population of entering agents is observed, misweighting will occur as some entering agents are effectively assigned a weight of zero in the calculation of the sample entry effect. However, as the second term of line two of Equation (7) shows, the calculation of the sample entry effect will also include observations from the set of continuing agents that are only sampled in the second period. The third line of Equation (7) highlights that a similar issue of misclassification arises for the calculation of the sample exit effect. Thus, we have the following:

Proposition 2. *Suppose the researcher has access to sample data given by the set $\boldsymbol{D}_{t_0} \cup \boldsymbol{D}_{t_1}$ where $\boldsymbol{D}_{t_0} \subset \boldsymbol{U}_{t_0}$ and $\boldsymbol{D}_{t_1} \subset \boldsymbol{U}_{t_1}$, and suppose the sample design is such that $\tilde{\boldsymbol{C}} \subset \boldsymbol{C}$, $\tilde{\boldsymbol{E}} \subset \boldsymbol{E} \cup \{\boldsymbol{C} \backslash \tilde{\boldsymbol{C}}\}$, and $\tilde{\boldsymbol{L}} \subset \boldsymbol{L} \cup \{\boldsymbol{C} \backslash \tilde{\boldsymbol{C}}\}$. Then the within, between, entry and exit effects computed by applying Equation* (2) *to the sample are biased. The bias of each effect is:*

1. *Within:* $\sum_{i \in \boldsymbol{C}} \left[\Pr(i \in \boldsymbol{D}_{t_0} \cap \boldsymbol{D}_{t_1}) - 1\right] s_{it_0} \Delta x_{it_1}$,

2. *Between:* $\sum_{i \in \boldsymbol{C}} \left[Pr(i \in \boldsymbol{D}_{t_0} \cap \boldsymbol{D}_{t_1}) - 1 \right] \Delta s_{it_1} x_{it_1}$,

3. *Entry:* $\sum_{i \in \boldsymbol{E}} \left[\Pr(i \in \boldsymbol{D}_{t_1}) - 1 \right] s_{it_1} x_{it_1}$
$- \sum_{i \in \boldsymbol{C}} \left[\Pr\left(i \in \boldsymbol{D}_{t_0} \cap \boldsymbol{D}_{t_1} \right) - \Pr\left(i \in \boldsymbol{D}_{t_1} \right) \right] s_{it_1} x_{it_1}$,

4. *Exit:* $- \sum_{i \in \boldsymbol{L}} \left[\Pr(i \in \boldsymbol{D}_{t_0}) - 1 \right] s_{it_0} x_{it_0}$
$+ \sum_{i \in \boldsymbol{C}} \left[\Pr\left(i \in \boldsymbol{D}_{t_0} \cap \boldsymbol{D}_{t_1} \right) - \Pr\left(i \in \boldsymbol{D}_{t_0} \right) \right] s_{it_0} x_{it_0}$.

Proof. Follows from subtracting Equation (2) from Equation (7). □

Together, Propositions 1 and 2 suggest that the common practice of applying an accounting decomposition to sample data leads to biased estimates of how within-agent changes and reallocations of economic activity across agents contribute to changes in aggregate outcomes. This implies that any conclusions drawn as to the relative importance of within and across agent changes in driving changes in aggregate outcomes will potentially be incorrect. This shortcoming is worth emphasizing as decompositions are often explicitly motivated as a means to understand these relative contributions.[11]

Propositions 1 and 2 also suggest that further sample restrictions will not address the biases created from applying an accounting decomposition to sample data. This is worth noting in light of the practice of limiting analyses to focus on within and between effects for continuing agents due to concerns over the reliability of data on agent entry and exit.[12] Propositions 1 and 2 imply that the sample within and between effects produced by this approach will typically still be biased unless the underlying sample design ensures the entire set of continuing agents is observed. In Supplemental Appendix A.2, we show that if the data possessed by the researcher is a random sample, the magnitude and direction of bias of the within and between effects can be directly measured, but the magnitude and direction of bias for the entry and exit effects can be determined if the sample is only subject to misweighting. As a result, even under random sampling one is unlikely to correctly measure the relative contributions of within and across agent changes to economic dynamics.

Propositions 1 and 2 assume that the researcher can calculate s_{it} for every observation in the sample. This requires the researcher to have information on the aggregate value of the variable used to calculate these shares. That is, the researcher must observe $Q_t = \sum_{i \in \mathbf{U}_t} q_{it}$. In many cases this information (such as aggregate output, employment, etc) can be obtained by the researcher from other data sources. In Supplemental Appendix A.3,

[11] See Bloom et al. (2025) for a recent example.

[12] See the work of Van Biesebroeck (2005b) or Bollard et al. (2013) for examples of this approach in the context of the study of firm dynamics and aggregate productivity.

we show that if Q_t is not observed, calculating the shares using data from the sample introduces an additional source of bias into the decomposition.

It is also worth emphasizing that the Propositions 1 and 2 are not specific to the BHC decomposition; it is relatively straightforward to extend these results to similar decompositions, such as those of Griliches and Regev (1995), Haltiwanger (1997) and Foster et al. (2001). Similar results can also be obtained for panel decompositions based on Olley and Pakes (1996); we show this formally for the Melitz and Polanec (2015) decomposition in Supplemental Appendix A.4.

A final issue worth raising is that of sample variability. Decompositions are often used to motivate economic theory by producing stylized facts as to the relative importance of the underlying components of the decomposition. This approach is natural in a world in which the universe of agents in observed. However, as Propositions 1 and 2 suggest, with sample data, different samples may produce different stylized facts. We formalize this point in Supplemental Appendix A.1.

2.1 Monte Carlo Evidence

We now provide simulation evidence to highlight the magnitudes of the potential biases that arise from applying an accounting decomposition to sample data.

For this exercise, we use simulated data on firm employment and output to study changes in aggregate labour productivity. Specifically, we compare the estimates of the within, between, entry and exit effects that we obtain from applying Equation (2) to subsamples of the data to the "true" effects that we obtain when Equation (2) is applied to the universe of observations.[13] In doing so, we conduct four sets of simulations. The first two sets assume that firm-characteristics are generated via a Pareto distribution, while the second two sets assume that they are generated via a bivariate normal distribution.[14] For each distribution we consider two sampling designs reflective of the types used in common practice: (i) a random sampling design in which all firms are sampled with the same probability, and (ii) a size based sampling design in which larger firms are sampled with higher probability.

We initially generated two data sets consisting of 50,000 observations on firm output, z_{it} and employment, l_{it}, in two periods: t_0 and t_1. For the first dataset, we assume the

[13] In these simulations, we compute bias as the difference between the estimate and the true population term, divided by the true population term. In contrast, the preceding theory does not scale the bias expressions by the population term. We adopt this scaling to simplify the interpretation of the bias in our simulations.

[14] We also considered a third simulation in which we assume firm characteristics are generated via a log-normal distribution. The results of these simulations are in Supplemental Appendix B.2.

Table 1: Simulation Parameters

Distribution	Parameter	Period t_0	Period t_1
Pareto	λ	3	2.5
	α	1	1.5
Bivariate Normal	μ_l	5	6
	μ_z	3	2
	σ_l	1	1
	σ_z	1	1
	ρ	0.4	0.6

Notes: Table reports parameters for distribution used to generate simulated data for the Monte Carlo exercises.

values of l_{it} and z_{it} for each observation in each period are randomly generated from a Pareto distribution with shape parameter λ and scale parameter α, as firm-specific characteristics such as size and productivity are often well approximated by this distribution (Axtell, 2001). In the second dataset, we adopt a starkly different assumption regarding the underlying distribution of firm characteristics and assume that the underlying the values of l_{it} and z_{it} are randomly generated from a bivariate normal distribution in each period, where μ_l and μ_z, and σ_l and σ_z are the means and standard deviations of l and z and ρ is the correlation between l and z. The specific values of the parameters governing each distribution in our simulations are reported in Table 1.

Once each dataset was generated, we randomly assigned observations to be "continuers," "entrants," and "exiters" to generate our population of interest.[15] For our dataset drawn from a Pareto distribution, this yielded a dataset with 38,675 observations at t_0 and 40,588 observations at t_1. Of these observations, 29,263 were continuers and observed at t_0 and t_1, 11,325 were entrants that were only observed at t_1, and 9,412 were exiters that were only observed at t_0. For this dataset, the aggregate change in labour productivity was 0.001, and the population within-, between-, entry-, and exit-effects calculated using Equation (2) were 0.059, -0.093, 0.277, and -0.242, respectively. For our dataset drawn from a bivariate normal distribution, our assignment process yielded a dataset with 38,631 observations at t_0 and 39,622 observations at t_1. Of these observations, 28,555 were assigned to be continuers, 11,067 were assigned to be entrants, and 10,076 were assigned to be exiters. For this dataset, the aggregate change in labour productivity was -0.266, and the population within-, between-, entry-, and exit-effects calculated using Equation (2) were -0.197, -0.006, 0.094, and -0.156, respectively.

We then used these datasets to conduct our simulations. For each dataset, we first

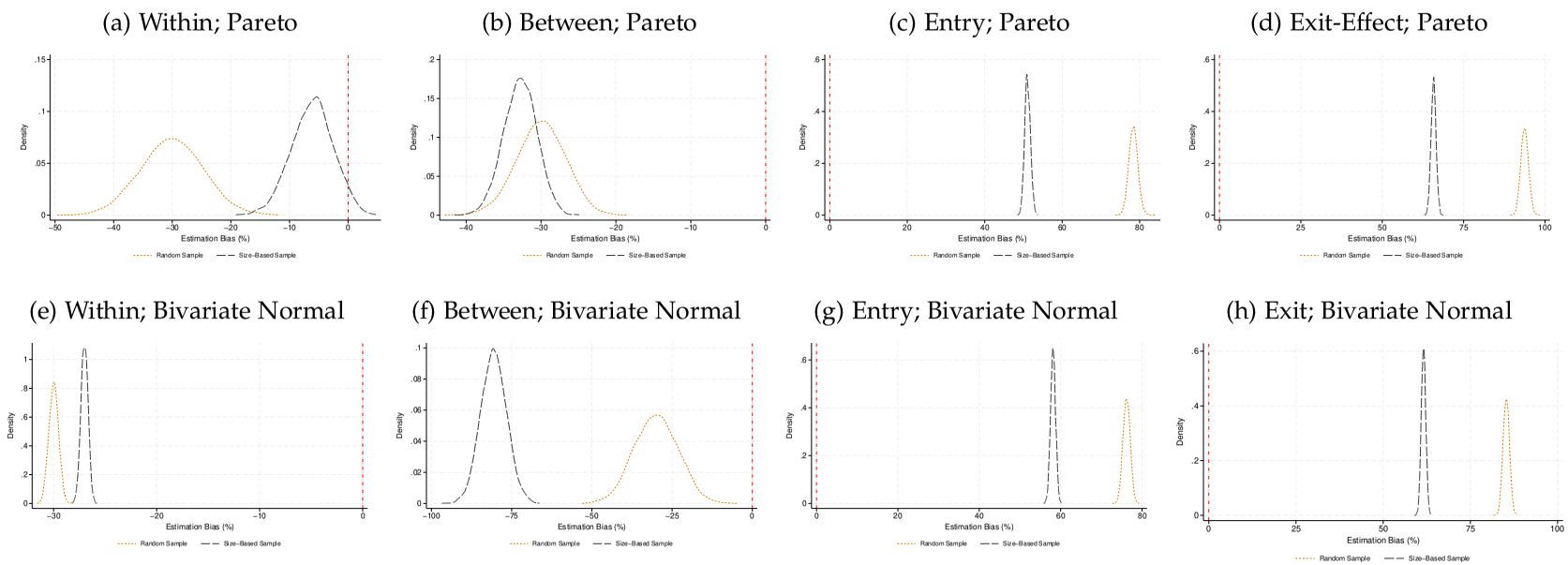
[15]Observations not assigned to these categories were dropped from the data.

drew 5,000 independent samples from the population using a random sample design in which 70% of the population was sampled in each period; for each sample we calculated the within, between, entry, and exit effects using Equation (2) and then calculated the magnitude of bias by comparing these estimates to the true population values. We then drew 5,000 independent samples from the population using a sized based sampling procedure in which 70% of the population was sampled in each period. For this procedure, we split the distribution of l_{it} into quartiles in each period, and then sampled all firms in the top quartile, 80% of the firms in the second quartile, 60% of the firms in the third quartile, and 40% of the firms in the bottom quartile. Firms were randomly (and independently) sampled from within their quartile in each period. We again calculated the within, between, entry, and exit effects for each sample using Equation (2) and then calculated the magnitude of bias by comparing these estimates to the true population values.

The results from these exercises are displayed in the eight panels of Figure 1, which plot kernel density estimates of the distribution of estimation bias (expressed in percentage terms) resulting from our simulations under both the random sampling design and the size-based sampling design. The top row of the figure (Panels (a) through (d)) plot the distribution of estimation bias for the within-, between-, entry, and exit-effects under the assumption the underlying distributions of firm characteristics follow a Pareto distribution. Similarly, the bottom row of the figure (Panels (e) through (h)) plot the distribution of estimation bias for the within-, between-, entry, and exit-effects under the assumption the underlying distributions of firm characteristics follow a bivariate normal distribution. In each plot, the distribution of estimation bias from the random sampling design is depicted in yellow with short dashed lines, while the distribution for the sized based sampling design is depicted in black with long dashed lines. In all cases, a red vertical dashed line is depicted at 0, to highlight when there is no bias and the estimate from using sample data corresponds to the true population value.

We emphasize two aspects of Figure 1. First, the results suggest that the biases identified in Proposition 1 and Proposition 2 can be large in magnitude. In our simulations, the sample design adopted produces within- and between-effects that are almost always underestimated and entry- and exit-effects that are always overestimated. When the data generating process follows a Pareto distribution and random sampling is used, the mean biases of the within- and between-effects are around -30%, while the mean biases of the entry- and exit-effects are over 70%. With size-based sampling the mean bias for the within and between effects are approximately -6% and -33%, respectively, while the mean biases of the entry- and exit-effects are over 50%. When the data generating pro-

Figure 1: Bias in Accounting Decompositions with Sample Data - Simulation Results



Notes: Figure shows BHC Decomposition estimation bias (in %) from a Monte Carlo Simulation (5,000 repetitions) for Baily et al. (1992) accounting decomposition performed with sample data. Results from two separate simulations are shown. The first simulation (panels (a) through (d)) generates population variables from a Pareto distribution with a fixed shape and scale parameter (N=50,000). The second simulation (panels (e) through (h)) generates population variables from a multivariate normal distribution with a fixed mean and correlation structure (N=50,000). In each panel, results from two separate Monte Carlo simulations are shown. One simulation draws a sample of observations from the population using random sampling. The other simulation follows a size-based sampling design.

cess follows a bivariate normal distribution, the average magnitude of each bias under random sampling are similar to the Pareto case. With size-based sampling, the bias of the within- and between-effects are much larger under a bivariate normal distribution than in the Pareto case, while the entry- and exit-effect biases are of a similar magnitude as observed in the Pareto case.[16]

Second, the results highlight the potential significant variability in the estimated magnitudes of the within, between, entry, and exit effects across samples, as suggested by Proposition A.1. If there was no variation in the estimated magnitudes of each effect across samples, then the biases would be constant and depicted by vertical lines in each panel of the Figure. Clearly, this is not the case. As Figure 1 shows, while there is a relatively small degree of variability in estimation errors in some cases (such as for estimates of the within effect for the bivariate normal distribution in Panel (d)), the degree of variability in estimation error can be quite large (as is the case for the within and between effects generated from a Pareto distribution).

3 The Decomposition as an Estimation Problem

Our analysis thus far suggests that the application of an accounting decomposition, such as that given by Equation (2), to understand the agent level sources of aggregate economic dynamics will lead to biased estimates when data on the population of interest is unavailable. To make progress in addressing these biases, we reformulate the decomposition as an estimation problem, and propose estimators of the decomposition terms based on the work of Horvitz and Thompson (1952).[17] In doing so, we focus our attention on the class of survey designs in which all agents in the population of interest are sampled with non-zero probability in each period.

Formally, again suppose the researcher has access to sample data defined by the sets $\mathbf{D}_{t_0} \subset \mathbf{U}_{t_0}$ and $\mathbf{D}_{t_1} \subset \mathbf{U}_{t_1}$, where $\mathbf{D}_{t_0} \cap \mathbf{D}_{t_1}$ is non-empty. For convenience, let the probability with which unit i is sampled in time t be denoted $\rho_{it} = \Pr(a_{it} = 1) = \Pr(i \in \mathbf{D}_t)$. We focus our attention on sample data that are generated by the class of survey designs that satisfy the following:

Assumption 1. *a) The probability with which an active unit is sampled in period t is non-zero (i.e., $\rho_{it} > 0 \; \forall \; i \in \boldsymbol{U}_t$) and known ex-post.*

[16]The bias in the accounting approach implemented with the random sample is typically larger than with size-based sampling because the size-based regime typically samples more large firms. Thus, size-based sampling usually captures more "economically meaningful" observations.

[17]HT estimation is the basis for inverse probability weighting, matching, and related estimators (Cerulli, 2015). For further discussion of HT estimation see Section 2.8 of Särndal et al. (1992).

b) Each continuer is sampled in both periods t_0 and t_1 with non-zero probability.

Assumption 1 is a non-negligible refinement to the set of survey designs we consider. Part a) requires no subset of the population of interest be left out of the sampling process systematically.[18] Part b) requires that the survey design does not systematically omit certain continuers across periods.[19] We make these assumptions as they underpin the sampling designs of many of the data sets that have been used in previous research, particularly on the study of firm dynamics. For example, size-based sampling regimes in which firms are placed into strata based on size with different inclusion probabilities across strata –such as those that underlie Canada's Annual Survey of Manufacturing and Logging Industries, or the Annual Survey of Manufacturers in the United States– satisfy Assumption 1 provided all strata are sampled and continuers from previous years are not intentionally omitted in subsequent years. Similarly, random sampling regimes that draw from the entire population of interest – such as the Decision Maker Panel in the United Kingdom (Bloom et al., 2025) – also satisfy Assumption 1.

It is also useful to make the following assumption about the sampling process:

Assumption 2. *The population shares, s_{it}, can be calculated for all $i \in \boldsymbol{D}_t$ and t.*

Assumption 2 requires that some additional information, namely the relative economic size of every agent, be available for the entire population of interest. While this is a convenient assumption that holds in many settings –such as when s_{it} is calculated using employment and industry and firm employment are observed– information of this kind is not always collected and available to the researcher. Below we show how this assumption can be relaxed in these cases.

3.1 Estimating the Within- and Between-Effects

We start by proposing estimators for the within and between effects. Let $\rho^c_{it_1} = \Pr(i \in \mathbf{D}_{t_0} \cap \mathbf{D}_{t_1} | i \in \mathbf{C})$ denote the probability with which agent i is sampled at both t_0 and t_1 conditional on surviving across both periods.[20] Then the HT estimator of the within-

[18] Notably, this assumption will fail in settings where there are reporting thresholds based on size, output or other agent characteristics, as in the case for many establishment-level datasets used in environmental economics, such as the Toxic Release Inventory maintained by the Environmental Protection Agency in the United States.

[19] This rules out non-overlapping and certain rotating panel schemes.

[20] Note that $\rho^c_{it_1}$ must account for both survey design and non-response.

effect ($\widehat{WE}_{t_1}^{HT}$) is defined as:

$$\widehat{WE}_{t_1}^{HT} = \sum_{i \in \tilde{\mathbf{C}}} \frac{s_{it_0} \Delta x_{it_1}}{\rho_{it_1}^{c}}, \tag{8}$$

and the HT estimator of the between-effect ($\widehat{BE}_{t_1}^{HT}$) is defined as:

$$\widehat{BE}_{t_1}^{HT} = \sum_{i \in \tilde{\mathbf{C}}} \frac{\Delta s_{it_1} x_{it_1}}{\rho_{it_1}^{c}}. \tag{9}$$

Equations (8) and (9) are straightforward extensions of the original estimator proposed by Horvitz and Thompson (1952). The HT estimator produces an unbiased estimate of a population total in settings where Assumption 1 holds by re-weighting the elements of the dataset using their first-order inclusion probability, that is, the probability with which each observation is sampled.[21] Here we extend this logic by re-weighting elements of the dataset using their second-order inclusion probability. Doing so delivers unbiased estimates of the within- (WE_{t_1}) and between-effects (BE_{t_1}):

Proposition 3. *Under Assumptions 1 and 2:*

1. $\widehat{WE}_{t_1}^{HT}$ *is an unbiased estimator of* WE_{t_1}.
2. $\widehat{BE}_{t_1}^{HT}$ *is an unbiased estimator of* BE_{t_1}.

Proof. To show $\widehat{WE}_{t_1}^{HT}$ is an unbiased estimator of WE_{t_1}, first let $a_{it_1}^{c}$ be an indicator such that $a_{it_1}^{c} = 1$ if $i \in \mathbf{D}_{t_0} \cap \mathbf{D}_{t_1}$ and $a_{it_1}^{c} = 0$ otherwise, so $\mathbb{E}\left[a_{it_1}^{c}\right] = \Pr(i \in \mathbf{D}_{t_0} \cap \mathbf{D}_{t_1} | i \in \mathbf{C}) = \rho_{it_1}^{c}$. By construction, $\tilde{\mathbf{C}} \subseteq \mathbf{C}$, which means $\widehat{WE}_{t_1}^{HT}$ can be expressed as:

$$\widehat{WE}_{t_1}^{HT} = \sum_{i \in \mathbf{C}} \frac{a_{it_1}^{c} s_{it_0} \Delta x_{it_1}}{\rho_{it_1}^{c}}.$$

[21] An alternative is to implement a Hájek-style estimator (Hájek, 1971) that further scales $\widehat{WE}_{t_1}^{HT}$ or $\widehat{BE}_{t_1}^{HT}$ by $\sum_{i \in \tilde{\mathbf{C}}} 1/\rho_{it_1}^{c}$. As a ratio estimator, this alternative estimator would be subject to small bias, but may result in lower variance than the HT estimator.

As $a^c_{it_1}$ is the only random variable, taking expectations of $\widehat{WE}^{HT}_{t_1}$ yields:

$$\begin{aligned}\mathbb{E}\left[\widehat{WE}^{HT}_{t_1}\right] &= \sum_{i\in\mathbf{C}} \frac{\mathbb{E}\left[a^c_{it_1}\right] s_{it_0}\Delta x_{it_1}}{\rho^c_{it_1}} \\ &= \sum_{i\in\mathbf{C}} \frac{\Pr(i\in\mathbf{D}_{t_0}\cap\mathbf{D}_{t_1}|i\in\mathbf{C}) s_{it_0}\Delta x_{it_1}}{\rho^c_{it_1}} \\ &= WE_{t_1}.\end{aligned}$$

Thus, $\widehat{WE}^{HT}_{t_1}$ is unbiased. The proof for $\widehat{BE}^{HT}_{t_1}$ follows analogously. □

It is worth noting that Proposition 3 implies that as long as $\rho^c_{it_1}$ is known for all observations in the sample, then unbiased estimates of WE_{t_1} and BE_{t_1} can be recovered for any sampling design that satisfies Assumptions 1 and 2. Importantly, this includes many classes of non-random sampling designs.

It is also worth noting that $\rho^c_{it_1}$ can be expressed as: $\rho^c_{it_1} = \Pr(i\in\mathbf{D}_{t_0}|i\in\mathbf{C})\Pr(i\in\mathbf{D}_{t_1}|\mathbf{D}_{t_0}, i\in\mathbf{C})$. With independent sampling, $\Pr(i\in\mathbf{D}_{t_0}|i\in\mathbf{C}) = \Pr(i\in\mathbf{D}_{t_0})$, which gives $\rho^c_{it_1} = \Pr(i\in\mathbf{D}_{t_0})\Pr(i\in\mathbf{D}_{t_1})$. That is, with independent sampling across periods, the second-order inclusion probability for an agent is the product of its two first-order inclusion probabilities from periods t_0 and t_1. This is useful to note because: (i) the data from many surveys used in previous work feature independent sampling across periods, and (ii) inverse first-order inclusion probabilities are often reported as sample weights.

While Proposition 3 provides the intuition behind our estimation approach, some settings may exist in which Assumption 2 is violated and the true values of s_{it} are not observed. To extend Proposition 3 to accommodate this case, suppose Assumption 2 fails. We now formally show that the HT estimators of the within and between effects will be biased, but that adjusted HT estimators can be derived that are approximately unbiased.

To see why the estimators in Proposition 3 will be biased in this case, note that if Assumption 2 fails, the researcher replaces true sample shares s_{it} with observed sample shares given by $s^s_{it} = q_{it}/[\sum_{i\in\mathbf{D}_t} q_{it}]$. Let the total share of q_{it} observed in the sample be given by $S^s_t = \frac{\sum_{i\in\mathbf{D}_t} q_{it}}{\sum_{i\in\mathbf{U}_t} q_{it}}$. By construction, $s^S_{it} = s_{it}/S^s_t$. The HT estimator of the within effect derived using sample shares ($\widehat{WE}^{HT,S}_{t_1}$) is:

$$\widehat{WE}^{HT,S}_{t_1} = \sum_{i\in\tilde{\mathbf{C}}} \frac{s^s_{it_0}\Delta x_{it_1}}{\rho^c_{it}} = \frac{1}{S^S_{t_0}} \sum_{i\in\tilde{\mathbf{C}}} \frac{s_{it_0}\Delta x_{it_1}}{\rho^c_{it_1}} = \frac{1}{S^S_{t_0}} \widehat{WE}^{HT}_{t_1}$$

As $S_t^s \in (0,1)$, employing the HT estimator with sample shares will overstate the absolute magnitude of the within effect. A similar issue arises for the HT estimator of the between effect. This can be addressed with an adjusted estimator that explicitly accounts for the fact that the population shares must be estimated.

Proposition 4. *If Assumption 2 fails but Assumption 1 holds, then adjusted HT estimators of WE_{t_1} and BE_{t_1} can be constructed for which the first order Taylor-series approximations are unbiased.*

The adjusted HT estimators of the within- and between-effects are given by:

$$\widehat{WE}_{t_1}^{AHT} = \widehat{S}_{t_0}^s \widehat{WE}_{t_1}^{HT,S} \tag{10}$$

$$\widehat{BE}_{t_1}^{AHT} = \widehat{S}_{t_1}^s \sum_{i \in \tilde{\boldsymbol{C}}} \frac{s_{it_1}^s x_{it_1}}{\rho_{it_1}^c} - \widehat{S}_{t_0}^s \sum_{i \in \tilde{C}} \frac{s_{it_0}^s x_{it_1}}{\rho_{it_1}^c}, \tag{11}$$

where $\widehat{S}_t^s = \sum_{i \in \boldsymbol{D}_t} q_{it} / \sum_{i \in \boldsymbol{D}_t} [q_{it}/\rho_{it}]$ is a HT estimator of the total q_{it} sampled.

Proof. We start by proving the proposition for the within effect. First, define $\widehat{Y} = \sum_{i \in \tilde{C}} \frac{q_{it_0} \Delta x_{it_1}}{\rho_{it_1}^c}$ and $\widehat{Q} = \sum_{i \in D_{t_0}} \frac{q_{it_0}}{\rho_{it_0}}$ and their analogous population terms as $Y = \sum_{i \in C} q_{it_0} \Delta x_{it_1}$ and $Q = \sum_{i \in U_{t_0}} q_{it_0}$. $\widehat{Y}$ and $\widehat{Q}$ are HT estimators of Y and Q, respectively. Under Assumption 1 they are unbiased.

Note that the adjusted HT estimator of the within effect can be rewritten as a ratio of these two estimators $\widehat{WE}_{t_1}^{AHT} = \frac{\widehat{Y}}{\widehat{Q}}$. Taking a first-order Taylor series expansion around (Y, Q) and using $WE_{t_1} = \frac{Y}{Q}$ gives:

$$\widehat{WE}_{t_1}^{AHT,lin} = \frac{Y}{Q} + \frac{1}{Q}(\widehat{Y} - Y) - \frac{Y}{Q^2}(\widehat{Q} - Q) = WE_{t_1} + \frac{1}{Q}(\widehat{Y} - Y) - \frac{WE_{t_1}}{Q}(\widehat{Q} - Q).$$

As $\mathbb{E}\left[\widehat{Y}\right] = Y$ and $\mathbb{E}\left[\widehat{Q}\right] = Q$, clearly the first-order approximation of $\widehat{WE}_{t_1}^{AHT}$ is unbiased. The proof for $\widehat{BE}_{t_1}^{AHT}$ follows analogously. □

Intuitively, the adjusted HT estimators will be biased because they are ratio estimators. However, Proposition 4 suggests this that bias will be small in expectation. This means that even when population shares must be estimated, our proposed estimators may be preferable to the existing practice of applying an accounting decomposition to a sample; as we have shown above the biases from the existing approach can be large.[22]

[22] Though the goal of this paper is to explore the small-sample properties of decompositions, in Supplemental Appendix A.5 we further establish that our HT and adjusted HT estimators are both consistent under standard regulatory conditions.

Below, we examine this further using Monte Carlo simulations.

3.2 The Selection-Effect

When the population is not observed it is generally not possible to recover estimates of the entry- or exit-effects. This is because the "true" status of any unit in the set $\mathbf{O} = \{i : i \in \mathbf{D}_{t_0} \cup \mathbf{D}_{t_1} \text{ and } i \notin \tilde{\mathbf{C}}\}$ is unknown without explicit information on exit for these units, and typical survey designs mean that such information is not collected. However, a combined selection effect can still be estimated.

Define the combined selection-effect as:

$$SE_{t_1} = \sum_{i \in \mathbf{E}} s_{it_1} x_{it_1} - \sum_{i \in \mathbf{L}} s_{it_0} x_{it_0}. \tag{12}$$

When the aggregate outcomes Y_{t_1} and Y_{t_0} are known, then the HT estimator of the selection-effect is:

$$\widehat{SE}_{t_1}^{HT} = \Delta Y_{t_1} - \widehat{WE}_{t_1}^{HT} - \widehat{BE}_{t_1}^{HT}, \tag{13}$$

and the adjusted HT estimator is:

$$\widehat{SE}_{t_1}^{AHT} = \Delta Y_{t_1} - \widehat{WE}_{t_1}^{AHT} - \widehat{BE}_{t_1}^{AHT}, \tag{14}$$

Note that if the aggregate outcomes Y_{t_1} and Y_{t_0} are not known, then they can be estimated using the following HT estimator:

$$\Delta \hat{Y}_{t_1} = \sum_{i \in \mathbf{D}_{t_1}} \frac{s_{it_1} x_{it_1}}{\rho_{it_1}} - \sum_{i \in \mathbf{D}_{t_0}} \frac{s_{it_0} x_{it_0}}{\rho_{it_0}}, \tag{15}$$

and used in place of the ΔY_{t_1} in $\widehat{SE}_t^{HT}$ and $\widehat{SE}_t^{AHT}$.

By the linearity of $\widehat{SE}_t^{HT}$, it is straightforward to show that $\widehat{SE}_t^{HT}$ is unbiased under Assumptions 1 and 2 and that $\widehat{SE}_t^{AHT}$ is approximately unbiased if Assumption 2 is relaxed. These results hold regardless of whether ΔY_{t_1} is estimated.

While this establishes an approach to obtain unbiased (or approximately unbiased) estimates of the combined selection effect from sample data, it is important to note that doing so will lead one to interpret all estimation error as selection. This could be particularly pronounced in settings in which sample designs change from year to year. Though in expectation this estimation error is mean zero (or approximately mean

zero for the adjusted HT estimator), this could create meaningful changes in selection estimates from year to year. Clearly, accounting for sample variability is important for selection, a point to which we now turn.

3.3 Variances

Thus far we have proposed estimators of the within-, between-, and selection-effects found in an aggregate decomposition. The remaining task is to reflect sample variability inherent in these estimators. As each of these is a HT estimator, it is straightforward to derive the following formulas for their variances. Here we derive these variances under both Assumptions 1 and 2. In Supplemental Appendix B.4, we discuss how to handle cases when Assumption 2 is relaxed.

To simplify expressions we move to matrix notation. To that end, we define the following vectors:

$$\mathbf{a}'_{t_1} = \begin{bmatrix} \frac{s_{1t_0}\Delta x_{1t_1}}{\rho^c_{1t_1}} & \frac{s_{2t_0}\Delta x_{2t_1}}{\rho^c_{2t_1}} & \dots & \frac{s_{N_{\tilde{c}}t_0}\Delta x_{N_{\tilde{c}}t_1}}{\rho^c_{N_{\tilde{c}}t_1}} \end{bmatrix},$$
$$\mathbf{b}'_{t_1} = \begin{bmatrix} \frac{\Delta s_{1t_1} x_{1t_1}}{\rho^c_{1t_1}} & \frac{\Delta s_{2t_1} x_{2t_1}}{\rho^c_{2t_1}} & \dots & \frac{\Delta s_{N_{\tilde{c}}t_1} x_{N_{\tilde{c}}t_1}}{\rho^c_{N_{\tilde{c}}t_1}} \end{bmatrix},$$

and the matrices:

$$\mathbf{D}_{t_1} = \begin{bmatrix} \delta_{11t_1} & \frac{1}{2}(\delta_{12t_1} + \delta_{21t_1}) & \dots & \frac{1}{2}(\delta_{1N_{\tilde{c}}t_1} + \delta_{N_{\tilde{c}}1t_1}) \\ \frac{1}{2}(\delta_{21t_1} + \delta_{12t_1}) & \delta_{22t_1} & \dots & \dots \\ \dots & \dots & \dots & \dots \\ \frac{1}{2}(\delta_{N_{\tilde{c}}1t_1} + \delta_{1N_{\tilde{c}}t_1}) & \dots & \dots & \delta_{N_{\tilde{c}}N_{\tilde{c}}t_1} \end{bmatrix},$$
$$\widehat{\mathbf{D}}_{t_1} = \begin{bmatrix} \hat{\delta}_{11t_1} & \frac{1}{2}(\hat{\delta}_{12t_1} + \hat{\delta}_{21t_1}) & \dots & \frac{1}{2}(\hat{\delta}_{1N_{\tilde{c}}t_1} + \hat{\delta}_{N_{\tilde{c}}1t_1}) \\ \frac{1}{2}(\hat{\delta}_{21t_1} + \hat{\delta}_{12t_1}) & \hat{\delta}_{22t_1} & \dots & \dots \\ \dots & \dots & \dots & \dots \\ \frac{1}{2}(\hat{\delta}_{N_{\tilde{c}}1t_1} + \hat{\delta}_{1N_{\tilde{c}}t_1}) & \dots & \dots & \hat{\delta}_{N_{\tilde{c}}N_{\tilde{c}}t_1} \end{bmatrix},$$

where $\delta_{klt_1} = \rho^c_{klt_1} - \rho^c_{kt_1}\rho^c_{lt_1}$ and $\hat{\delta}_{klt_1} = \frac{\rho^c_{klt_1} - \rho^c_{kt_1}\rho^c_{lt}}{\rho^c_{klt_1}}$. Note that $\hat{\delta}_{klt_1} = 1 - \frac{\rho^c_{kt_1}\rho^c_{lt_1}}{\rho^c_{klt_1}}$ if $k \neq l$ and $\hat{\delta}_{klt_1} = 1 - \rho^c_{kt_1}$ if $k = l$. With this notation in place, we have the following:.

Proposition 5. *Under Assumptions 1 and 2, unbiased estimators of the variances of the within-,*

between-, and selection-effects estimated using HT estimators are given by:

$$\widehat{Var}\left(\widehat{WE}_{t_1}^{HT}\right) = \boldsymbol{a}_{t_1}'\hat{\boldsymbol{D}}_{t_1}\boldsymbol{a}_{t_1}, \tag{16}$$

$$\widehat{Var}\left(\widehat{BE}_{t_1}^{HT}\right) = \boldsymbol{b}_{t_1}'\hat{\boldsymbol{D}}_{t_1}\boldsymbol{b}_{t_1}, \tag{17}$$

$$\widehat{Var}\left(\widehat{SE}_{t_1}^{HT}\right) = \left[\boldsymbol{a}_{t_1} + \boldsymbol{b}_{t_1}\right]'\hat{\boldsymbol{D}}_{t_1}\left[\boldsymbol{a}_{t_1} + \boldsymbol{b}_{t_1}\right] \tag{18}$$

Proof. From Result 2.8.1 in Särndal et al. (1992), the variance of $\widehat{WE}_t$ is:

$$Var\left(\widehat{WE}_{t_1}\right) = \sum_{k\in\mathbf{C}_{t_1}}\sum_{l\in\mathbf{C}_{t_1}}\delta_{klt_1}\frac{s_{kt_0}\Delta x_{kt_1}}{\rho^c_{kt_1}}\frac{s_{lt_0}\Delta x_{lt_1}}{\rho^c_{lt_1}}.$$

Expressing this in matrix form, replacing δ_{klt} with its estimator $\hat{\delta}_{klt}$ and $\mathbf{C}_{t_1}$ with $\tilde{\mathbf{C}}_{t_1}$ gives the expression in the proposition. As $\hat{\delta}_{klt}$ is a HT type estimator, $\widehat{Var}(\widehat{WE}_{t_1}^{HT})$ will be an unbiased estimator of $Var(\widehat{WE}_{t_1})$. The variance of the between effect can be derived by analogy.

The variance of the selection-effect estimator is given by:

$$\begin{aligned} Var\left(\widehat{SE}_{t_1}^{HT}\right) &= Var\left(\Delta Y_{t_1} - \widehat{WE}_{t_1}^{HT} - \widehat{BE}_{t_1}^{HT}\right) \\ &= Var\left(\widehat{WE}_{t_1}^{HT}\right) + Var\left(\widehat{BE}_{t_1}^{HT}\right) + 2Cov\left(\widehat{WE}_{t_1}^{HT}, \widehat{BE}_{t_1}^{HT}\right). \end{aligned}$$

From Result 5.4.1 in Särndal et al. (1992), $Cov(\widehat{WE}_{t_1}^{HT}, \widehat{BE}_{t_1}^{HT}) = \mathbf{a}_{t_1}'\mathbf{D}_{t_1}\mathbf{b}_{t_1}$. Substituting into the above expression and replacing $\mathbf{D}_{t_1}$ with its estimator $\hat{\mathbf{D}}_{t_1}$ gives:

$$Var\left(\widehat{SE}_{t_1}^{HT}\right) = \mathbf{a}_{t_1}'\hat{\mathbf{D}}_{t_1}\mathbf{a}_{t_1} + \mathbf{b}_{t_1}'\hat{\mathbf{D}}_{t_1}\mathbf{b}_{t_1} + 2\mathbf{a}_{t_1}'\hat{\mathbf{D}}_{t_1}\mathbf{b}_{t_1}.$$

Factoring gives the formula in the proposition. As $\hat{\mathbf{D}}_{t_1}$ is a HT-type estimator, then $Var(\widehat{SE}_{t_1}^{HT})$ is an unbiased estimator of $Var(SE_{t_1}^{HT})$. Note that Assumption 2 implies that ΔY_{t_1} is known. In Supplemental Appendix B.4, we show how the estimator of the variance of the selection effect when ΔY_{t_1} is estimated. □

The variance expressions in Proposition 5 are straightforward. However, estimating them may be non-trivial for complex survey designs, as they require the availability of pairwise inclusion probabilities. In such a case, an approximation may need to be adopted. We explore one such approximation in Supplemental Appendix B.4, but leave

a full exploration to future work.

3.4 Monte Carlo Evidence

We now turn to evaluate the performance of our proposed estimator using simulations. We compare the performance of our proposed adjusted HT estimator to two alternative approaches that have been used elsewhere in the literature. Here we examine the adjusted HT estimator, rather than the HT estimator, as we use the adjusted HT estimator in our empirical applications. We report simulation results for the HT estimator in Supplemental Appendix B.1. The first estimator, which we term the "unweighted" estimator, estimates the within, between, and selection effects by applying Equation (3) directly to sample data, an approach that is common in the literature (see, e.g., (Van Biesebroeck, 2008b; Chari et al., 2021; Bloom et al., 2025)). The second, which we refer to as the "simple weighted" estimator estimates the within, between, and selection effects by reweighting the terms in Equation (3) using contemporaneous (current period) sample weights. The simple weighted estimator is analogous to our adjusted HT estimator, but ρ_{it_1} is used in place of $\rho^c_{it_1}$. This approach is also common in the literature (see, e.g., Griliches and Regev (1995); Bollard et al. (2013); Harrison et al. (2013)).

For our first exercise, we again rely on the simulated data from Section 2.1. Here, we only utilize the data drawn from the Pareto distribution and the bivariate normal distribution under the sized based sampling regime as the results from Section 2.1 suggest that the estimation bias associated with this sampling regime is smaller than with random sampling. For this exercise, for each distribution, we drew 5,000 independent samples (distinct from those drawn in Section 2.1) from the simulated population and estimated the within, between and selection effects using the simple weighted estimator, the unweighted estimator, and our proposed adjusted HT estimator. For each draw, we calculated the difference between the resulting estimates and the true population values to determine the magnitude of bias from each approach. For each sample, we use the same size-based sampling routine described in Section 2.1, which independently samples 70% of the population in each period.

The results of these simulations are reported in Table 2 and Figure 2. Table 2 reports the mean and median biases (in percents) across all 5,000 simulations for each of the three estimators assessed in each of the Monte Carlo simulations. Panel (a) shows the results for the simulations with the Pareto distribution and panel (b) shows the results for the simulations with the normal distribution. In both panels, rows one through three report the within-, between-, and selection-effects, respectively. Columns (1) and (2)

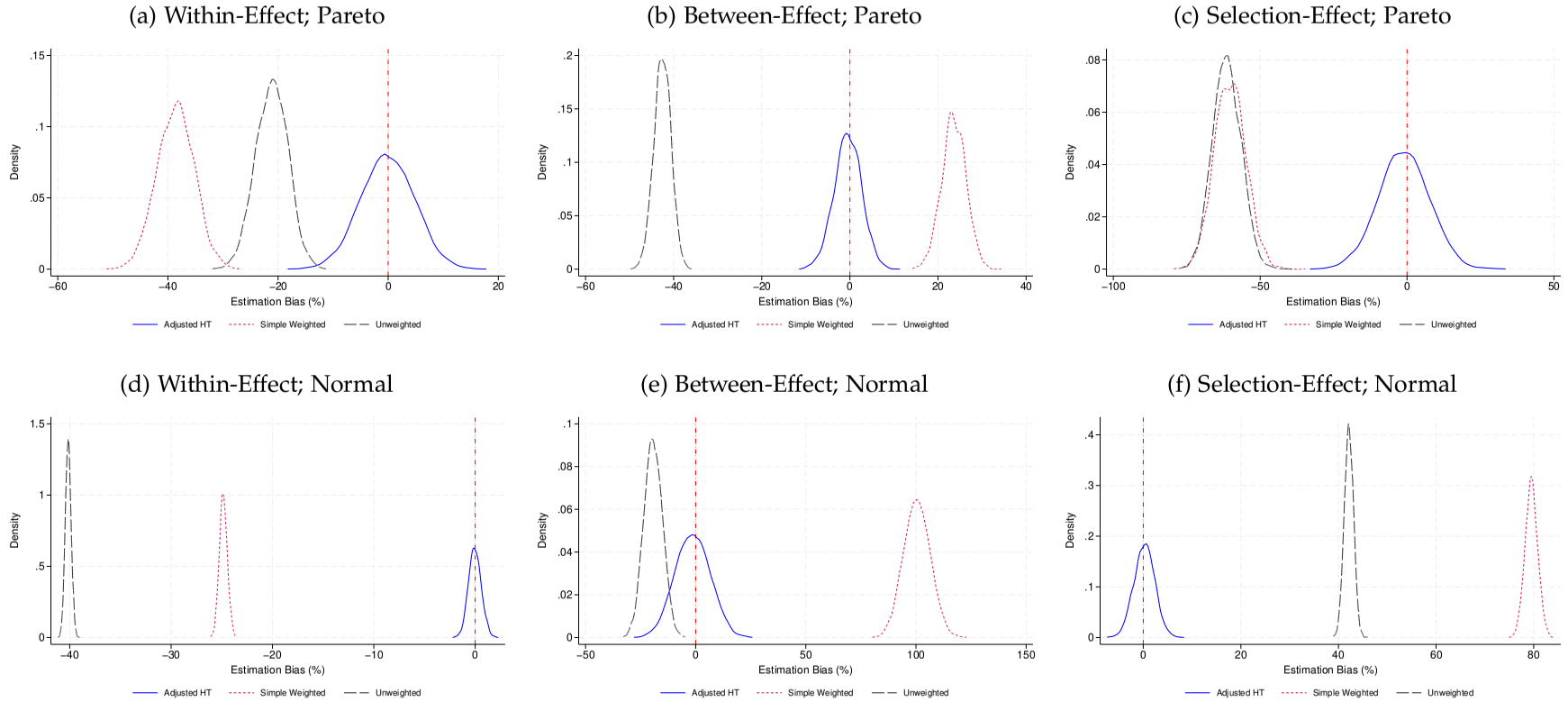
show the results for the adjusted HT estimator. Columns (3) and (4) show the results for the simple weighted estimator. Columns (5) and (6) show the results for the unweighted estimator. Figure 2 reports kernel density estimates of the distribution of simulation results for all three estimators. Panels (a) through (c), respectively, plot the within-, between-, and selection-effect distributions for the Pareto distribution simulation. Panels (d) through (f), respectively, plot the within-, between-, and selection-effect distributions for the normal distribution simulation.

As the results reported in Table 2 and Figure 2 show, the performance of the three different estimation approaches differs substantially. For the adjusted HT estimator, the mean and median bias in all three decomposition effects are very close to zero in both the Pareto and bivariate normal simulations.[23] Examining Figure 2 further shows that the bias of the adjusted HT estimator is indeed centered around zero for all three estimated effects. The same cannot be said of the other two estimators; both appear to be badly biased in our simulations. In both sets of simulations, the unweighted and simple estimators tend to underestimate the within-effect (by between 20% and 40%, on average). Figure 2 shows that the within effect bias is negative in all simulations for these estimators. Interestingly, the simple weighted and unweighted estimators produce quite different estimates of the between-effect. In both simulations, the unweighted estimator tends to underestimate this effect while the simple weighted estimator tends to overestimate the effect. The bias in these effects is also quite large. For example, the between-effect estimated by the simple weighted estimator for the bivariate normal distribution simulation is almost double the true value in the population. Lastly, the selection effects are also badly estimated. Both estimators tend to underestimate this effect for the Pareto distribution simulation (by roughly 60%) but overestimate this effect for the bivariate normal distribution simulation (by 40-80%).

While we focus on the relative performance of the unweighted, simple weighted and adjusted HT estimators here for the sake of brevity, in Supplemental Appendix B.4, we also examine the performance of our proposed variance estimators estimated using the adjusted HT approach. We begin by deriving explicit expressions for the variances of the HT estimators for each decomposition term under a stratified random sample design. As using the adjusted HT estimator also requires estimating each observation's share, s_{it}, these formulas will be subject to small bias if used to estimate the variances of the adjusted HT estimators. An alternative to using these analytical variance expressions is to instead approximate the variances with the Taylor linearization approach of Demnati

[23] In Supplemental Appendix B.3 we examine the robustness of our adjusted HT estimator to small sample coverage. We find similar results in simulations that sample 20% of the population.

Figure 2: Decomposition Simulation Results



Notes: Figure shows BHC Decomposition estimation bias (in %) from a Monte Carlo Simulation (5,000 repetitions) for three estimators: adjusted Horvitz-Thompson, simple weighted, and unweighted. Results from two separate simulations are shown. The first simulation (panels (a) through (c)) generates population variables from a Pareto distribution with a fixed shape and scale parameter (N=50,000). The second simulation (panels (d) through (f)) generates population variables from a multivariate normal distribution with a fixed mean and correlation structure (N=50,000). In both simulations, a sample of observations are drawn from the population following a size-based sampling design with fixed sample probabilities.

Table 2: Decomposition Simulation Results

	Adjusted HT		Simple Weighted		Unweighted	
	Mean	Median	Mean	Median	Mean	Median
	(1)	(2)	(3)	(4)	(5)	(6)
	Panel (a): Pareto Distribution					
Within-Effect	-0.13	-0.15	-38.60	-38.55	-21.15	-21.08
Between-Effect	-0.47	-0.45	23.58	23.55	-42.55	-42.53
Selection-Effect	-1.05	-1.07	-60.22	-60.25	-61.34	-61.43
	Panel (b): Bivariate Normal Distribution					
Within-Effect	-0.03	-0.04	-24.85	-24.86	-40.15	-40.15
Between-Effect	-1.38	-1.41	100.44	100.38	-19.27	-19.35
Selection-Effect	0.22	0.26	79.52	79.51	42.13	42.13

Notes: Table reports results from two sets of Monte Carlo simulations (each with 5,000 repetitions) testing the bias in three estimators based on Equation (3): adjusted Horvitz-Thompson, simple weighted, and unweighted. Mean and median estimation biases (in %) are reported for each estimator. The first simulation (Panel (a)) generates population variables from a Pareto distribution with a fixed shape and scale parameter (N=50,000). The second simulation (Panel (b)) generates population variables from a bivariate normal distribution with a fixed mean and correlation structure (N=50,000). In both simulations, a sample of observations is drawn from the population following a size-based sampling design with fixed sample probabilities.

and Rao (2004). We find the bias is typically smaller than the approximation error in our simulations.[24] As a result, we adopt the analytical variance expressions when we present variance estimates in the empirical applications that follow.

4 Applications

We now turn to apply our adjusted HT estimation approach to study how firm dynamics have contributed to the productivity growth of several developing countries. While decompositions have been used to answer questions in many fields, here we focus on firm dynamics and aggregate productivity because this is, perhaps, the most common application of the decomposition methodology. We focus our applications on developing and emerging countries as the measurement of how firm dynamics have affected productivity growth in this setting is often complicated by the fact that the full population of interest (the universe of firms) is typically not observed, meaning the results from the direct application of an accounting decomposition will be biased.[25] Here we conduct two empirical applications to illustrate the utility of our methodology in such settings.

[24]The mean magnitude of the bias is between 0.2-5.8%, depending on the term and distribution.

[25]An important caveat is that here we are defining the population of interest as formal firms. This omits informal firms, which could be an important element of the complete universe of firms in some developing and emerging economies.

4.1 India

For our first application, we revisit how firm dynamics have contributed to productivity growth in India's manufacturing sector. This is a natural setting for applying our HT estimator. The sources of productivity changes in India's manufacturing sector have been studied widely, including via decompositions (e.g. Bollard et al. (2013) and Harrison et al. (2013)). However, the data sources used in these studies do not provide information for the universe of Indian manufacturing firms, meaning the results from previous decompositions –which rely on "simple weighted" estimators in which the decomposition terms are reweighted using contemporaneous sample weights rather than their second-order inclusion probability– are biased. Here, we employ the same data source as these previous decompositions –the Annual Survey of Industries (ASI), a sample of all formal firms from India's manufacturing sector– and perform a BHC decomposition of the form given by Equation (2). We compare the estimates of the within, between and selection effects obtained from our proposed adjusted HT estimation approach with those from simple weighted estimators and those from "unweighted" estimators in which Equation (2) is applied directly to sample data.[26]

We conduct this exercise for two different measures of firm productivity for fiscal years 1998/99 to 2015/16 (1998-2015 for simplicity).[27] Our first measure is the natural log of labor productivity, measured as value-added per worker. Our second measure is the natural log of total factor productivity (TFP), estimated following the approach of Ackerberg et al. (2015). While previous studies of how firm dynamics have contributed to India's productivity growth have focused on measures of TFP, here we perform decompositions with both measures for the purposes of illustration as labor productivity has been the focus of decompositions in other settings (e.g. Griliches and Regev (1995), Baily et al. (2001)). Our estimates indicate manufacturing TFP increased by 118.9% between 1998-2015, while manufacturing labor productivity increased by 72.9% during this time.

Next, we turn to examine the sources of productivity growth underlying the changes in each measure using the BHC decomposition. We must first determine the appropriate second-order inclusion probability to use as a weight in our adjusted HT estimators. The inverse of the inclusion probability that accounts for non-response for a given unit-year (i.e., $1/\Pr(i \in D_t)$ is provided for each observation in the ASI as a sample weight. As data for the ASI is independently drawn each year, the sample design does not

[26] We use our adjusted HT estimator in this case as we must estimate the shares used in the decomposition. Recall from Section 3 that this estimator has bias that is approximately zero.

[27] See Supplemental Appendix C for details of the data, including productivity measures.

explicitly introduce cross-period dependencies. However, as the sample frame includes both a census and sample segments, there is an implicit cross-period dependency for continuing census units. As such, for our analysis of the ASI we define $\rho^c_{it1} = \rho_{it_0}\gamma$, where $\gamma = \rho_{it_1}$ for all sample units (i.e., independent sampling across periods) and $\gamma = 1$ for all census units (i.e., we adjust the conditional second-period inclusion probability).[28]

Our analysis starts with a series of year-on-year decompositions to obtain annual estimates of the within, between and selection effects for each of the three estimation approaches described above: (i) our proposed adjusted HT estimator, (ii) the simple-weighted estimator, and (iii) the naive, unweighted estimator. The results of this exercise are reported in the six panels of Figure 3. In the first column of the figure (Panels (a), (c) and (e)), productivity is measured using TFP; in the second column (Panels (b), (d), and (f)), productivity is measured using value added per worker. The first row of each column presents annual estimates of the within effect, the second presents annual estimates of the between effect, and the third presents annual estimates of the selection effect.

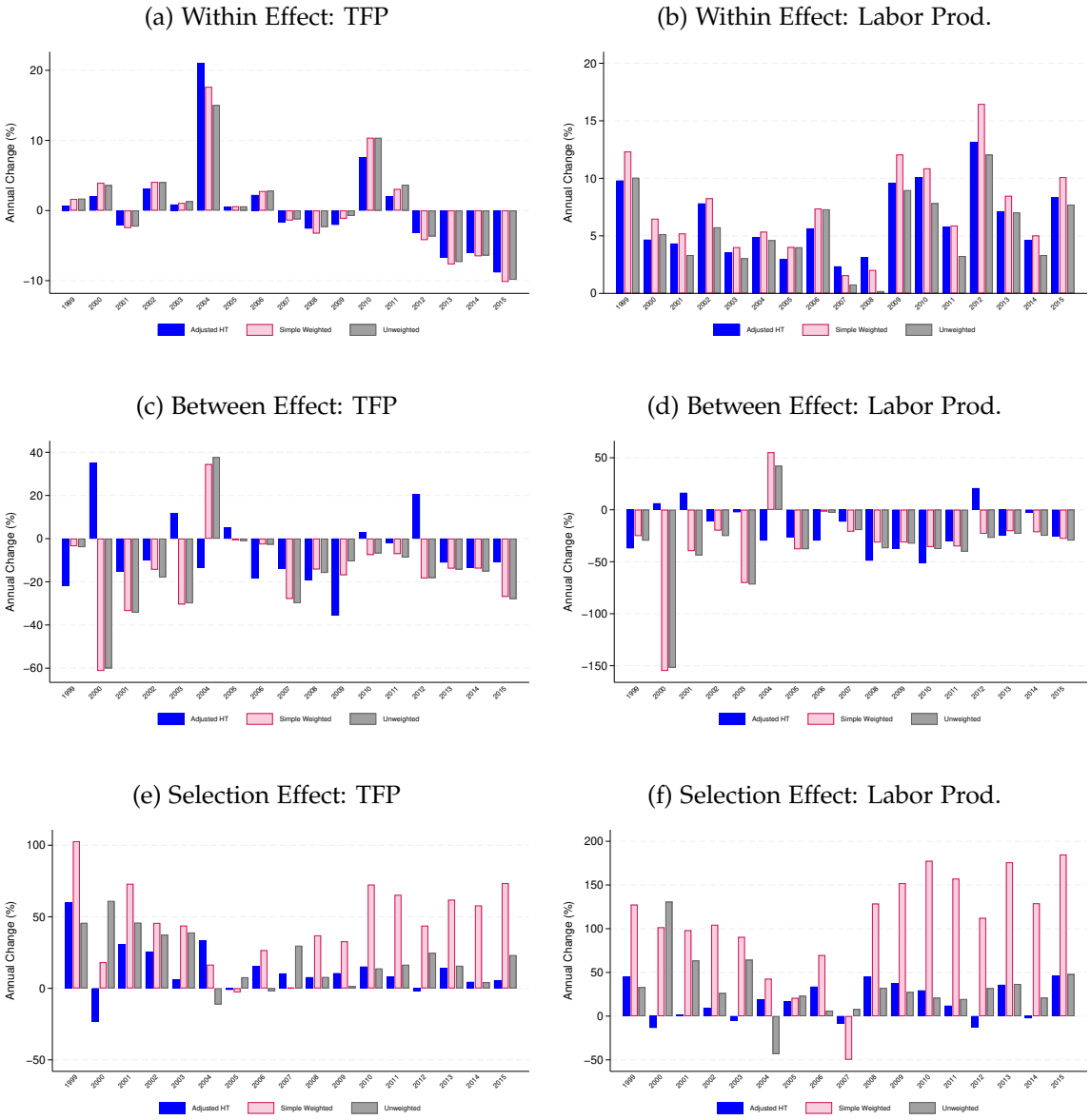
As Figure 3 shows, the estimates for the within, between, and selection effects obtained from the adjusted HT estimator are often substantially different from those produced using the simple weighted and unweighted approaches. For example, consider the estimates for the year 2000. For both TFP and labor productivity, the simple weighted and unweighted estimators both overstate the within effect and return between and selection effect estimates with the wrong sign.

While Figure 3 is illustrative, we also quantify the potential difference across approaches by calculating the average, minimum, and maximum percentage differences between the estimates of the within, between and selection effects from the adjusted HT estimator and those from the simple weighted and unweighted estimators, respectively. These results are presented in the two panels of Table 3. Panel (a) reports the differences between estimators when firm productivity is measured as TFP, while Panel (b) reports the differences when firm productivity is measured as value added per worker.

The results presented in Table 3 further highlight how the simple weighted and unweighted approaches that have been used previously in the literature can produce substantially different estimates than the adjusted HT estimator. For example, the results presented in the first row of Panel (a) indicate that the simple weighted estimator overstates the within effect by 28% on average, with annual deviations that can be as high as a 43% underestimate or a 139% overestimate. The differences for the between and

[28]This change only affects the 7% of census firms that report $\rho_{it_1} \neq 1$. Leaving all ρ_{it_1}s unadjusted does not meaningfully alter our results.

Figure 3: India Decomposition Estimates



Notes: Figure presents annual estimates of the within, between and selection effects from two different BHC decompositions. In the first column of the figure (Panels (a), (c) and (e)), productivity is measured using TFP estimated using the approach of Ackerberg et al. (2015). In the second column of the figure (Panels (b), (d), and (f)), productivity is measured as value added per worker. In all cases, productivity is log transformed. Panels (a) and (b) present estimates of the within effect, panels (c) and (d) present estimates of the between effect, and Panels (e) and (f) present estimates of the selection effect. Each panel shows the year-on-year change in aggregate productivity due to the panel's decomposition effect for three different estimators: the adjusted Horvitz-Thompson (HT), the simple weighted, and the unweighted.

selection effects are much larger, averaging an underestimate of 345% for the between effect and an overestimate of 568% for the selection effect. The results presented in Panel (b) suggest that these difference are not unique to TFP. On average, the simple weighted estimator overstates the magnitude of the within and selection effects and understates

Table 3: Differences in Estimates Across Estimators

	Within			Between			Selection		
	Mean	Min	Max	Mean	Min	Max	Mean	Min	Max
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
					Panel (a): TFP				
Simple	28	-43	139	-345	-5,005	2,666	568	-2,079	6,243
Unweighted	24	-63	147	-359	-4,943	2,602	292	-2,089	4,283
					Panel (b): LP				
Simple	13	-36	39	-308	-3,455	1,749	1,422	-1,759	2,828
Unweighted	-16	-94	35	-358	-3,389	1,480	361	-1,282	3,091

Notes: Table compares estimates of the within, between and selection effects produced by the simple weighted and unweighted estimators with those obtained using the adjusted Horvitz-Thompson estimator. Panel (a) reports these comparisons for total factor productivity (TFP) estimated using the approach of Ackerberg et al. (2015). Panel (b) shows these comparisons for labor productivity (LP) measured as value added per worker. Each panel shows percentage differences between the estimates from the Horvitz-Thompson estimator and those from the simple weighted estimator (row 1) and the unweighted estimator (row 2). The table displays the mean, minimum, and maximum difference across all years for each decomposition term.

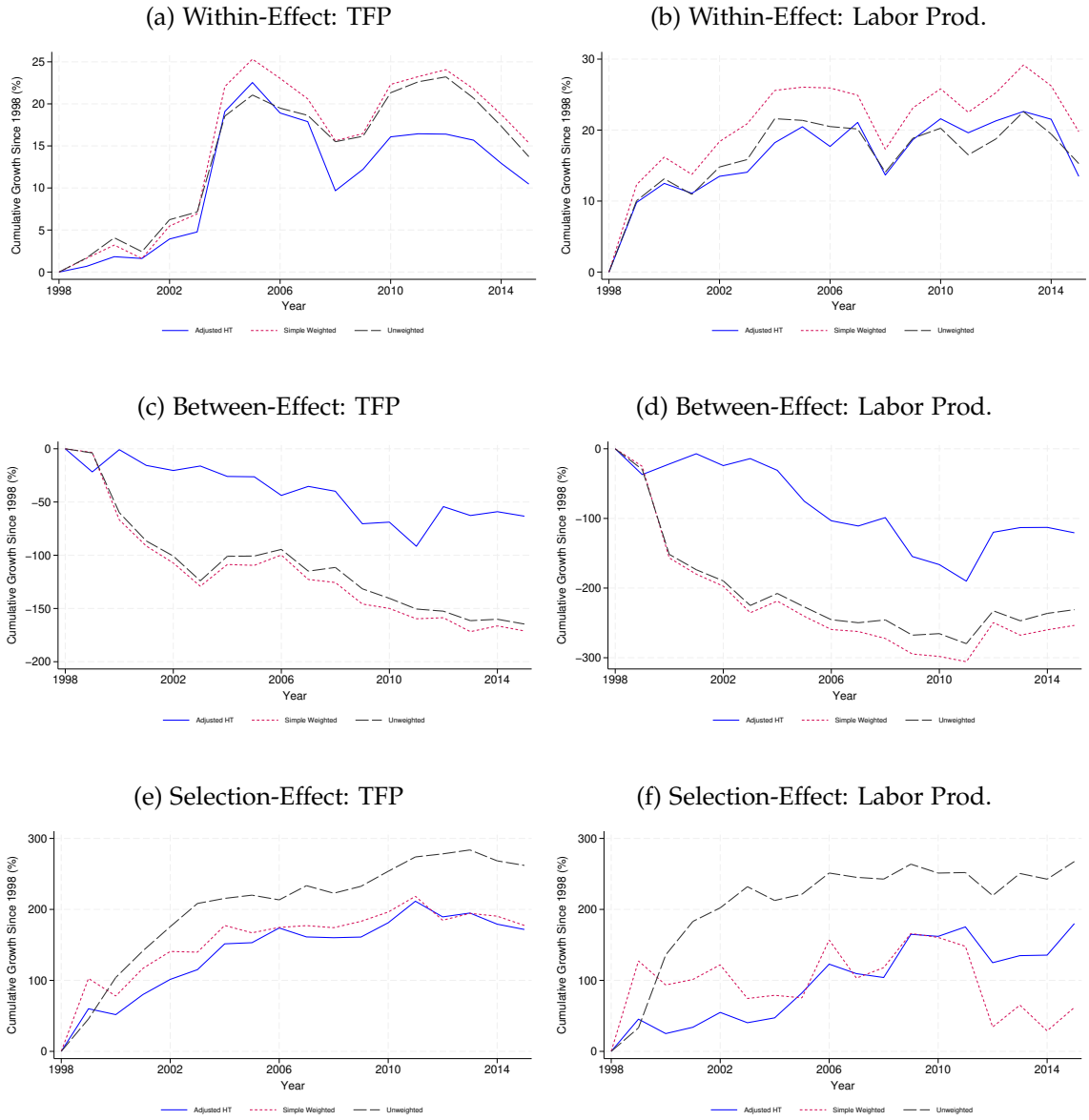
the magnitude of the between effect. On average, the unweighted estimator understates the magnitude of the within and between effects and overstates the selection effect.

Given that the within, between, and selection effects measure the relative importance of within-firm changes in productivity and productivity gains from the reallocation of economic activity across firms, the results presented in Table 3 and Figure 3 show that correcting for the bias from using sample data can drastically alter our understanding of how firm dynamics have contributed to aggregate productivity growth. To further investigate the implications of correcting for sample bias, as a final step, we perform an additional exercise in which we decompose cumulative aggregate productivity growth in each year, rather than on an annual basis (that is t_1 is the year in question and t_0 is always 1998). Again, we perform these decompositions using both labor productivity and TFP. The results of this exercise are reported in the six panels of Figure 4.[29] As in Figure 3, in the first column (Panels (a), (c) and (e)), productivity is measured using TFP, and in the second column (Panels (b), (d), and (f)) productivity is measured using value added per worker. The first row of each column presents estimates of the cumulative within effect, the second presents estimates of the cumulative between effect, and the third presents estimates of the cumulative selection effect.

The estimates presented in Figure 4 further highlight how correcting for sample bias

[29]This exercise defines a continuing firm as one observed in both 1998 and the year in question. An alternative approach would be to compute the running sum of the estimated annual decomposition terms in Figure 3 from 1998 to the year in question.

Figure 4: Cumulative Decomposition Estimates



Notes: Figure shows results of a BHC Decomposition of total factor productivity (TFP) estimated using the ACF approach (panels (a), (c), and (e), and labor productivity (panels (b), (d), and (f)) for Indian manufacturing. Changes since 1998 are decomposed and productivity is log transformed. Panels (a) & (b) report the within-effect. Panel (c) & (d) show the between-effect. Panel (e) & (f) show the selection-effect. Each panel shows the change in aggregate productivity since 1998 due to the panel's decomposition effect for three different decomposition estimators: the adjusted Horvitz-Thompson, the simple weighted, and the unweighted.

can impact our understanding of how firm dynamics shape aggregate productivity. For example, consider the estimates presented in the first column of the Figure. The estimates in Panel (a) indicate that both the simple weighted and unweighted approaches overstate the role of within-firm changes as a source of aggregate TFP growth throughout most of our sample. The estimates in Panel (c) indicate that both the simple weighted

and unweighted approaches overstate the negative effects of reallocations of economic activity across continuing firms. Finally, the estimates in Panel (e) show that the simple weighted approach modestly overstates the role of reallocations of economic activity through firm entry and exit in determining aggregate productivity, while the unweighted approach greatly overstates the importance of selection. It is worth emphasizing that these results are not unique to TFP; the estimates presented in the second column of Figure 4 for labor productivity paint a similar picture. It is also worth emphasizing that these differences are economically important. For example, the estimates for TFP for 2015 from the simple weighted approach suggest reallocations of economic activity across firms –the combined between and selection effects– increased aggregate productivity by 6.3%, whereas the estimates from the adjusted HT estimator suggest that these reallocations increased productivity by 108.4%. Similarly, the simple weighted estimates for labor productivity from 2015 suggest that reallocation served to lower aggregate productivity by 191.7%, whereas estimates from the adjusted HT estimator suggest that these reallocations increased productivity by 59.4%.[30]

These results are notable because previous work relying on the ASI has highlighted within-firm changes as the dominant driver of productivity growth in India (e.g. Bollard et al. (2013); Harrison et al. (2013)). Our approach produces estimates of the within effect that are of a similar magnitude to those produced previously. However, correcting for sample bias substantially alters the relative importance of reallocation to India's productivity growth.

4.2 Sub-Saharan Africa

For our second application we study firm dynamics and aggregate productivity in Sub-Saharan Africa. This is also a useful setting for applying our HT estimator; unlike the case of India, where decompositions have been employed previously to simultaneously estimate how within-firm and across firm changes have impacted aggregate productivity, there are no such estimates for most Sub-Saharan African countries due to issues related to the availability and reliability of firm-level data. Here we use our adjusted HT estimator and firm level data on TFP and output from the World Bank Enterprise Sur-

[30]Figures 3 and 4 provide alternative cumulative differences between selection effect estimates, with the annual decomposition showing larger differences between the estimate from the adjusted HT estimator and those from the simple and unweighted estimators. This likely arises from the fact that the selection effect estimates are more susceptible to sample variability than the within and between estimates. As such, some of the annual selection estimates will reflect changes in sample rather than true selection.

vey (WBES) to overcome these limitations.[31] We use the data from the WBES to perform BHC decompositions of the form given by Equation (2) and provide novel estimates of the within, between, and selection effects for 17 countries in Sub-Saharan Africa over a variety of sample periods.[32] For this application, we also produce standard errors for each estimate to allow for statistical inference.[33][34]

The results from this exercise are presented in Table 4. Each row of the table presents results from a different decomposition. The first column of the table lists the country and period over which the decomposition is performed. The second column reports the adjusted HT estimate of the change in aggregate productivity over the reported period. The final three columns of the table report the corresponding adjusted HT estimates of the within, between, and selection effects. Throughout the table, estimated standard errors are reported in parentheses.

As the estimates reported in the second column of Table 4 show, our adjusted HT estimate of the change in aggregate productivity is not statistically different from zero at conventional levels for eleven countries. This suggests that for these countries aggregate productivity growth did not change meaningfully during the period they were surveyed by the WBES. Moreover, for half of the cases that are statistically significant, aggregate productivity growth declined (Cameroon, Cote d'Ivoire, and Ethiopia).

The estimates reported in the final three columns of Table 4 suggest that firm dynamics have often reduced aggregate productivity growth for many of the countries in our sample. For example, consider the case of Ghana. As the estimate reported in the second column of the table shows, aggregate productivity was essentially unchanged over the period 2006-2012. Yet, as the estimates reported in columns three and four indicate, aggregate productivity grew by 0.81% from within-firm productivity changes and by 7.36% via the reallocation of economic activity to higher productivity continuing firms. These gains were offset by a selection effect of -8.22%, suggesting that net-entry has significantly reduced aggregate productivity in Ghana. Similarly, consider the case

[31] We use the TFP estimates provided by the WBES, which are derived using OLS, rather than the ACF method, due to the small sample sizes available and a lack of available instruments. While these TFP estimates are themselves biased, we use them to demonstrate the potential value of our method. For further details on the WBES data, see Supplemental Appendix C.

[32] Our sample consists of the 17 countries for which TFP data is available for a panel of firms. The sample period for each country is determined by survey coverage.

[33] The WBES uses a stratified random sample design. Unlike the ASI, respondents from earlier waves are systematically contacted in later waves. The WBES provides adjusted panel weights to account for these panel firms (World Bank, 2025a). As a result, we use $\rho^c_{it1} = \rho_{it_0}\gamma$, where $\gamma = \rho_{it_1}$ for all entering firms and $\gamma = \Pr(i \in D_{t_1} | D_{t_0})$ for all continuing firms.

[34] The WBES handles non-response in the interview stage by replacing each non-responder firm with another firm willing to complete the WBES questionnaire that falls into the same sector-size-region cell. As such, we do not adjust the sample inclusion probabilities for non-response.

Table 4: Firm Dynamics and Productivity in Sub-Saharan Africa

Country/Period	ΔProductivity	Within	Between	Selection
Angola; 2005-2009	1.04	2.37	7.19	-8.52
	(1.51)	(0.63)	(10.21)	(10.13)
Botswana; 2005-2009	-0.29	-0.83	0.48	0.06
	(1.09)	(0.58)	(0.28)	(1.22)
Cameroon; 2008-2015	-2.24	6.61	-4.36	-4.50
	(0.94)	(37.56)	(28.85)	(47.43)
Cote d'Ivoire; 2008-2015	-2.37	1.30	5.44	-9.12
	(0.92)	(7.46)	(31.33)	(32.21)
DRC; 2005-2009	1.05	1.15	-0.63	0.53
	(0.22)	(0.45)	(0.31)	(0.50)
Ethiopia; 2010-2014	-1.67	-0.77	-0.55	-0.36
	(0.76)	(0.25)	(0.20)	(1.21)
Ghana; 2006-2012	-0.05	0.81	7.36	-8.22
	(3.12)	(0.47)	(4.32)	(3.11)
Kenya; 2006-2017	0.50	-0.13	0.21	0.41
	(0.13)	(0.06)	(0.13)	(0.32)
Mali; 2006-2015	0.07	-0.06	-0.23	0.36
	(0.09)	(0.08)	(0.09)	(0.09)
Nigeria; 2006-2013	2.08	0.19	12.70	-10.81
	(2.74)	(0.18)	(7.92)	(7.41)
Rwanda; 2005-2018	-1.58	1.67	-2.18	-1.07
	(0.97)	(0.46)	(1.40)	(1.34)
Senegal; 2006-2013	0.15	0.70	1.89	-2.44
	(0.84)	(0.27)	(1.65)	(1.44)
Sierra Leone; 2016-2022	-1.21	-5.41	8.38	-4.18
	(0.86)	(1.99)	(3.82)	(4.26)
South Africa; 2006-2019	1.21	7.74	21.17	-27.70
	(0.20)	(5.27)	(11.29)	(12.46)
Tanzania; 2005-2012	0.02	-3.74	5.12	-1.36
	(1.04)	(2.24)	(2.49)	(3.23)
Uganda; 2005-2012	1.09	-2.37	0.25	3.21
	(4.53)	(2.03)	(6.51)	(4.54)
Zambia; 2006-2018	-0.48	-2.25	-12.79	14.56
	(1.45)	(0.97)	(3.67)	(3.29)

Notes: Table reports HT decomposition estimates for a sample of Sub-Saharan African countries using TFP estimates from the WBES. Column two reports the estimated change in aggregate productivity within each country over the relevant period. The last three columns report estimates of the within-, between-, and selection-effects, respectively. Estimates are expressed in log changes. Standard errors are in parentheses below each term.

of Zambia. As in Ghana, aggregate productivity in Zambia was essentially unchanged over its sample period. However, the selection effect in Zambia is 14.56%, suggesting that net entry significantly increased aggregate productivity. However, as the estimates presented in columns three and four indicate, aggregate productivity in Zambia was stagnant primarily due to a reduction in productivity from the reallocation of economic activity to lower continuing firms that were relatively low productivity, reflected in the between effect of -12.79%. When viewed in their entirety, the estimates reported in the Table 4 indicate that for many countries in Sub-Saharan Africa, existing firms have been getting less productive, and economic activity has been reallocated from relatively productive to relatively unproductive firms.

As a final step in our analysis, we compare the estimates of the within, between and selection effects from Table 4 with those from a "naive" approach in which Equation (2) is directly applied to the WBES data. For the sake of brevity, the results from this comparison are presented in Supplemental Appendix D. As these results show, applying the naive decomposition produces very little variation in the within, between, and selection effect estimates, with most hovering quite close to zero. In contrast, the estimates from our adjusted HT estimator are starkly different, highlighting how our estimator can improve our understanding of the sources of productivity growth in settings where the direct application of an accounting decomposition would yield little useful insight.

5 Conclusion

Decompositions are a widely used method for quantifying the sources of aggregate fluctuations in economic outcomes. The typical decomposition starts from an accounting identity that explicitly links within-agent changes and re-allocations of economic activity across agents to changes in an aggregate outcome. In this paper, we show that the common practice of applying such decompositions to study changes in aggregate outcomes using sample data leads to biased estimates of the sources of these aggregate changes. We also make progress in addressing these biases by reformulating the decomposition as an estimation problem and proposing a new estimator for the class of survey designs in which all agents from the population of interest have a non-zero probability of being sampled in each period, as this type of design underpins many of the datasets used in previous work. We further illustrate the utility of our estimator through two applications studying how firm dynamics have contributed to aggregate productivity growth in India and in a set of seventeen countries in Sub-Saharan Africa.

While this paper makes initial progress in the use of decompositions with sample

data, there are several avenues for future research. Estimating decomposition components in settings with other survey designs, particularly those with reporting thresholds, is one potential fruitful path. The study of sample bias for alternative decomposition approaches, such as that of Petrin and Levinsohn (2012), would also be worthwhile. We leave these avenues to future work.

References

Acemoglu, D., U. Akcigit, H. Alp, N. Bloom, and W. Kerr (2018). Innovation, Reallocation, and Growth. *American Economic Review 108*(11), 3450–3491.

Ackerberg, D. A., K. Caves, and G. Frazer (2015). Identification Properties of Recent Production Function Estimators. *Econometrica 83*(6), 2411–2451.

Allcott, H., A. Collard-Wexler, and S. D. O'Connell (2016). How Do Electricity Shortages Affect Industry? Evidence from India. *American Economic Review 106*(3), 587–624.

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

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

Aw, B. Y., X. Chen, and M. J. Roberts (2001). Firm-Level Evidence on Productivity Differentials and Turnover in Taiwanese Manufacturing. *Journal of Development Economics 66*(1), 51–86.

Axtell, R. L. (2001). Zipf Distribution of US Firm Sizes. *Science 293*(5536), 1818–1820.

Backus, M. (2020). Why is Productivity Correlated with Competition? *Econometrica 88*(6), 2415–2444.

Baily, M., C. Hulten, and D. Campbell (1992). Productivity Dynamics Manufacturing in Plants. *Brookings Papers on Economic Activity: Microeconomics*, 187–267.

Baily, M. N., E. J. Bartelsman, and J. Haltiwanger (2001). Labor Productivity: Structural Change and Cyclical Dynamics. *Review of Economics and Statistics 83*(3), 420–433.

Baldwin, J. R. and W. Gu (2006). Plant Turnover and Productivity Growth in Canadian Manufacturing. *Industrial and Corporate Change 15*(3), 417–465.

Bartelsman, E., J. Haltiwanger, and S. Scarpetta (2009). Measuring and Analyzing Cross-Country Differences in Firm Dynamics. In T. Dunne, J. B. Jensen, and M. J. Roberts (Eds.), *Producer Dynamics: New Evidence from Micro Data*, pp. 15–76. University of Chicago Press.

Bartelsman, E. J. and M. Doms (2000). Understanding Productivity: Lessons from Longitudinal Microdata. *Journal of Economic Literature 38*(3), 569–594.

Barwick, P. J., L. Chen, S. Li, and X. Zhang (2025). Entry Deregulation, Market Turnover, and Efficiency: China's Business Registration Reform. *Review of Economics and Statistics*, 1–46.

Bau, N. and A. Matray (2023). Misallocation and Capital Market Integration: Evidence from India. *Econometrica 91*(1), 67–106.

Bernard, A. B., J. Eaton, J. B. Jensen, and S. Kortum (2003). Plants and Productivity in International Trade. *American Economic Review 93*(4), 1268–1290.

Bloom, N., P. Bunn, P. Mizen, P. Smietanka, and G. Thwaites (2025). The Impact of COVID-19 on Productivity. *Review of Economics and Statistics 107*(1), 28–41.

Boehm, J. and E. Oberfield (2020). Misallocation in the Market for Inputs: Enforcement and the Organization of Production. *Quarterly Journal of Economics 135*(4), 2007–2058.

Bollard, A., P. J. Klenow, and G. Sharma (2013). India's Mysterious Manufacturing Miracle. *Review of Economic Dynamics 16*(1), 59–85.

Brandt, L., J. Van Biesebroeck, and Y. Zhang (2012). Creative Accounting or Creative Destruction? Firm-Level Productivity Growth in Chinese Manufacturing. *Journal of Development Economics 97*(2), 339–351.

Cerulli, G. (2015). *Econometric Evaluation of Socio-Economic Programs: Theory and Applications*, Volume 49. Berlin: Springer.

Chandra, A., A. Finkelstein, A. Sacarny, and C. Syverson (2016). Health Care Exceptionalism? Performance and Allocation in the US Health Care Sector. *American Economic Review 106*(8), 2110–2144.

Chandra, A., P. Kakani, and A. Sacarny (2024). Hospital Allocation and Racial Disparities in Health Care. *Review of Economics and Statistics 106*(4), 924–937.

Chari, A., E. M. Liu, S.-Y. Wang, and Y. Wang (2021). Property Rights, Land Misallocation, and Agricultural Efficiency in China. *Review of Economic Studies 88*(4), 1831–1862.

Cherniwchan, J., B. R. Copeland, and M. S. Taylor (2017). Trade and the Environment: New Methods, Measurements, and Results. *Annual Review of Economics 9*, 59–85.

Demnati, A. and J. N. Rao (2004). Linearization Variance Estimators for Model Parameters from Complex Survey Data. *Survey Methodology 30*(1), 193–201.

Disney, R., J. Haskel, and Y. Heden (2003). Restructuring and Productivity Growth in UK Manufacturing. *Economic Journal 113*(489), 666–694.

Foster, L., J. Haltiwanger, and C. Syverson (2008). Reallocation, Firm Turnover, and Efficiency: Selection on Productivity or Profitability? *American Economic Review 98*(1), 394–425.

Foster, L., J. C. Haltiwanger, and C. J. Krizan (2001). Aggregate Productivity Growth: Lessons From Microeconomic Evidence. In C. R. Hulten, E. R. Dean, and M. J. Harper (Eds.), *New Developments in Productivity Analysis*, pp. 303–372. University of Chicago Press.

Frazer, G. (2005). Which Firms Die? A Look at Manufacturing Firm Exit in Ghana. *Economic Development and Cultural Change 53*(3), 585–617.

Gomez, M. (2023). Decomposing the Growth of Top Wealth Shares. *Econometrica 91*(3), 979–1024.

Griliches, Z. and H. Regev (1995). Firm Productivity in Israeli Industry 1979–1988. *Journal of Econometrics 65*(1), 175–203.

Hájek, J. (1971). Comment on a Paper by D. Basu. In V. P. Godambe and D. A. Sprott (Eds.), *Foundations of Statistical Inference*, pp. 236. Toronto: Holt, Rinehart and Winston.

Haltiwanger, J. (1997). Measuring and Analyzing Aggregate Fluctuations: The Importance of Building From Microeconomic Evidence. *Federal Reserve Bank of St. Louis Review 79*(3), 55–78.

Harrison, A. E., L. A. Martin, and S. Nataraj (2013). Learning Versus Stealing: How Important are Market-Share Reallocations to India's Productivity Growth? *World Bank Economic Review 27*(2), 202–228.

Horvitz, D. G. and D. J. Thompson (1952). A Generalization of Sampling Without Replacement From A Finite Universe. *Journal of the American Statistical Association 47*(260), 663–685.

Hsieh, C.-T. and P. J. Klenow (2009). Misallocation and Manufacturing TFP in China and India. *Quarterly Journal of Economics 124*(4), 1403–1448.

India's Ministry of Commerce and Industry (2025). Office of the Economic Advisor. `https://eaindustry.nic.in/`.

Lentz, R. and D. T. Mortensen (2008). An Empirical Model of Growth Through Product Innovation. *Econometrica 76*(6), 1317–1373.

Li, Y. and M. Rama (2015). Firm Dynamics, Productivity Growth, and Job Creation in Developing Countries: The Role of Micro-and Small Enterprises. *The World Bank Research Observer 30*(1), 3–38.

Martin, L. A., S. Nataraj, and A. E. Harrison (2017). In With The Big, Out With The Small: Removing Small-Scale Reservations in India. *American Economic Review 107*(2), 354–386.

McMillan, M., D. Rodrik, and Í. Verduzco-Gallo (2014). Globalization, Structural Change, and Productivity Growth, with an Update on Africa. *World Development 63*, 11–32.

McMillan, M. and A. Zeufack (2022). Labor Productivity Growth and Industrialization in Africa. *Journal of Economic Perspectives 36*(1), 3–32.

Melitz, M. J. and S. Polanec (2015). Dynamic Olley-Pakes Productivity Decomposition With Entry and Exit. *RAND Journal of Economics 46*(2), 362–375.

Najjar, N. and J. Cherniwchan (2021). Environmental Regulations and the Cleanup of Manufacturing: Plant-Level Evidence. *Review of Economics and Statistics 103*(3), 476–491.

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

Pavcnik, N. (2002). Trade Liberalization, Exit, and Productivity Improvements: Evidence from Chilean Plants. *Review of Economic Studies 69*(1), 245–276.

Petrin, A. and J. Levinsohn (2012). Measuring Aggregate Productivity Growth Using Plant-Level Data. *The Rand Journal of Economics 43*(4), 705–725.

Söderbom, M., F. Teal, and A. Harding (2006). The Determinants of Survival Among African Manufacturing Firms. *Economic Development and Cultural Change 54*(3), 533–555.

Syverson, C. (2011). What Determines Productivity? *Journal of Economic Literature 49*(2), 326–365.

Särndal, C.-E., B. Swensson, and J. Wretman (1992). *Model Assisted Survey Sampling*. New York: Springer.

Tybout, J. R. (2000). Manufacturing Firms in Developing Countries: How Well Do They Do, and Why? *Journal of Economic Literature 38*(1), 11–44.

Van Biesebroeck, J. (2003). Productivity Dynamics with Technology Choice: An Application to Automobile Assembly. *Review of Economic Studies 70*(1), 167–198.

Van Biesebroeck, J. (2005a). Exporting Raises Productivity in Sub-Saharan African Manufacturing Firms. *Journal of International Economics 67*(2), 373–391.

Van Biesebroeck, J. (2005b). Firm Size Matters: Growth and Productivity Growth in African Manufacturing. *Economic Development and Cultural Change 53*(3), 545–583.

Van Biesebroeck, J. (2008a). Aggregating and Decomposing Productivity. *Review of Business and Economics 53*(2), 122–146.

Van Biesebroeck, J. (2008b). The Sensitivity of Productivity Estimates: Revisiting Three Important Debates. *Journal of Business and Economic Statistics 26*(3), 311–328.

World Bank (2025a). Enterprise Surveys Sampling Methodology. `https://www.enterprisesurveys.org/content/dam/enterprisesurveys/documents/methodology/Sampling_Note-Consolidated-2-16-22.pdf`. Accessed Feb 22, 2026.

World Bank (2025b). World Bank Enterprise Surveys. `www.enterprisesurveys.org`. Accessed March 18, 2025.

Supplemental Appendix

A Additional Theoretical Results

A.1 Sample Variability

As discussed in Section 2 of the main paper, with sample data, different samples may produce different stylized facts, a point we now formalize:

Proposition A.1. *Suppose the researcher has access to two samples of data given by the sets* $\boldsymbol{D}^1_{t_0} \cup \boldsymbol{D}^1_{t_1}$ *and* $\boldsymbol{D}^2_{t_0} \cup \boldsymbol{D}^2_{t_1}$ *where* $\boldsymbol{D}^1_{t_0} \subset \boldsymbol{U}_{t_0}$, $\boldsymbol{D}^1_{t_1} \subset \boldsymbol{U}_{t_1}$, $\boldsymbol{D}^2_{t_0} \subset \boldsymbol{U}_{t_0}$, *and* $\boldsymbol{D}^2_{t_1} \subset \boldsymbol{U}_{t_1}$. *Furthermore, suppose* $\boldsymbol{D}^1_{t_0} \cup \boldsymbol{D}^1_{t_1} \neq \boldsymbol{D}^2_{t_0} \cup \boldsymbol{D}^2_{t_1}$. *Then applying Equation* (2) *to each sample does not necessarily yield the same relative magnitudes of the within, between, entry, and exit effects.*

Proof. To prove Proposition A.1, let the set of continuing, entering and exiting observations in $\mathbf{D}^m_t$ be denoted by $\tilde{\mathbf{C}}^m$, $\tilde{\mathbf{E}}^m$, and $\tilde{\mathbf{L}}^m$, respectively, for $m = \{1, 2\}$ and let k_{t_0} and j_{t_1} represent arbitrary observations in $\mathbf{U}_{t_0}$ and $\mathbf{U}_{t_1}$ such that $\mathbf{D}^1_{t_0} = \{\mathbf{D}^2_{t_0}, k_{t_0}\}$ and $\mathbf{D}^1_{t_1} = \{\mathbf{D}^2_{t_1}, j_{t_1}\}$. Given this setup, there are two possible cases.

1. $k_{t_0}, j_{t_1} \in \tilde{\mathbf{C}}^1$. In this case, Equation (2) implies the difference in the within effect computed across the two samples will be $s_{k_{t_0}} \Delta x_{j_{t_1}}$ and the difference in the between effect computed across the two samples will be $\Delta s_{j_{t_1}} x_{j_{t_1}}$. The two samples will produce the same entry and exit effects as $\tilde{\mathbf{E}}^1 = \tilde{\mathbf{E}}^2$ and $\tilde{\mathbf{L}}^1 = \tilde{\mathbf{L}}^2$. Hence, the relative magnitudes of the within, between, entry, and exit effects from the two samples will be different provided $s_{k_{t_0}} \Delta x_{j_{t_1}} \neq 0$ and $\Delta s_{j_{t_1}} x_{j_{t_1}} \neq 0$.

2. $k_{t_0} \in \tilde{\mathbf{L}}^1$ and $j_{t_1} \in \tilde{\mathbf{E}}^1$. In this case, Equation (2) implies the difference in the entry effect computed across the two samples will be $s_{j_{t_1}} x_{j_{t_1}}$ and the difference in the exit effect computed across the two samples will be $-s_{k_{t_0}} x_{k_{t_0}}$. The two samples will produce the same within and between effects as $\tilde{\mathbf{C}}^1 = \tilde{\mathbf{C}}^2$. Hence, the relative magnitudes of the within, between, entry, and exit effects from the two samples will be different provided $s_{j_{t_1}} x_{j_{t_1}} \neq 0$ and $-s_{k_{t_0}} x_{k_{t_0}} \neq 0$.

□

A.2 The Biases from a Random Sample

In Section 2 of the main text, we noted that the bias of the within and between effects can be determined if the researcher possesses a random sample, but the bias of the entry and

exit effects can only be determined if the sample is subject to misweighting. Formally, we have:

Proposition A.2. *Suppose the researcher applies Equation* (2) *to sample data given by the set* $\boldsymbol{D}_{t_0} \cup \boldsymbol{D}_{t_1}$ *where* $\boldsymbol{D}_{t_0} \subset \boldsymbol{U}_{t_0}$ *and* $\boldsymbol{D}_{t_1} \subset \boldsymbol{U}_{t_1}$.

1. *If* $\boldsymbol{D}_{t_0}$ *and* $\boldsymbol{D}_{t_1}$ *are random samples from* $\boldsymbol{U}_{t_0}$ *and* $\boldsymbol{U}_{t_1}$, *that is, if* $Pr(i \in \boldsymbol{D}_t) = \gamma_t < 1 \; \forall \; i$ *and the sample data are subject to misweighting, meaning* $\tilde{\boldsymbol{C}} \subset \boldsymbol{C}$, $\tilde{\boldsymbol{E}} \subset \boldsymbol{E}$, *and* $\tilde{\boldsymbol{L}} \subset \boldsymbol{L}$, *then the sample within and between effects obtained by the researcher understate their true values by* $\gamma_{t_0}\gamma_{t_1}$, *and the sample entry and exit effects understate their true values by* γ_{t_0} *and* γ_{t_1}, *respectively.*

2. *If* $\boldsymbol{D}_{t_0}$ *and* $\boldsymbol{D}_{t_1}$ *are random samples from* $\boldsymbol{U}_{t_0}$ *and* $\boldsymbol{U}_{t_1}$, *that is, if* $Pr(i \in \boldsymbol{D}_t) = \gamma_t < 1 \; \forall \; i$ *and the sample data are subject to misweighting and misclassification, meaning* $\tilde{\boldsymbol{C}} \subset \boldsymbol{C}$, $\tilde{\boldsymbol{E}} \subset \boldsymbol{E} \cup \{\boldsymbol{C} \backslash \tilde{\boldsymbol{C}}\}$, *and* $\tilde{\boldsymbol{L}} \subset \boldsymbol{L} \cup \{\boldsymbol{C} \backslash \tilde{\boldsymbol{C}}\}$, *then the sample within and between effects obtained by the researcher understate their true values by* $\gamma_{t_0}\gamma_{t_1}$. *In this case, the bias for both the sample entry and exit effects cannot be signed.*

Proof. To prove Part 1 of the proposition, first note that when $\mathbf{D}_{t_0}$ and $\mathbf{D}_{t_1}$ are random samples with sampling probabilities $Pr(i \in \mathbf{D}_{t_0}) = \gamma_{t_0}$ and $Pr(i \in \mathbf{D}_{t_1}) = \gamma_{t_1}$, Equation (5) can be rewritten as:

$$\mathbb{E}\left[\Delta Y_{t_1}^{D}\right] = \gamma_{t_0}\gamma_{t_1} \sum_{i \in \mathbf{C}} s_{it_0} \Delta x_{it_1} + \gamma_{t_0}\gamma_{t_1} \sum_{i \in \mathbf{C}} \Delta s_{it_1} x_{it_1} + \gamma_{t_1} \sum_{i \in \mathbf{E}} s_{it_1} x_{it_1} - \gamma_{t_0} \sum_{i \in \mathbf{L}} s_{it_0} x_{it_0} \quad \text{(A.1)}$$

because random sampling implies $Pr(i \in \mathbf{D}_{t_0} \cap \mathbf{D}_{t_1}) = Pr(i \in \mathbf{D}_{t_0}) Pr(i \in \mathbf{D}_{t_1})$. The magnitude of bias for the sample within, between, entry and exit effects can be obtained directly via comparison of Equation (A.1) with Equation (2).

For Part 2 of the proposition, first note that Proposition 1 and Proposition 2 show that the sample within and between effects have the same bias regardless of whether the sample is subject to misweighting or misweighting and misclassification. As such, the biases for the sample within and between effects when the sample is subject to misweighting and misclassification are given in Equation (A.1). The bias of the entry effect can be written as:

$$\begin{aligned} \sum_{i \in \mathbf{E}} [1 - \Pr(i \in \mathbf{D}_{t_1})] s_{it_1} x_{it_1} - \sum_{i \in \mathbf{C}} [\Pr(i \in \mathbf{D}_{t_1}) - \Pr(i \in \mathbf{D}_{t_0} \cap \mathbf{D}_{t_1})] s_{it_1} x_{it_1} & \\ = [1 - \gamma_{it_1}] \sum_{i \in \mathbf{E}} s_{it_1} x_{it_1} - [\gamma_{it_1} - \gamma_{it_0}\gamma_{it_1}] \sum_{i \in \mathbf{C}} s_{it_1} x_{it_1} & \quad \text{(A.2)} \end{aligned}$$

Since $\sum_{i\in\mathbf{C}} s_{it_1}x_{it_1}$ is unobserved by the researcher, it follows that the bias of the entry effect can not be signed. The proof for the exit effect follows similarly. □

A.3 Decomposition with Sample Shares

Recall from the main text that $s_{it} = q_{it}/Q_t$, where $Q_t = \sum_{i\in\mathbf{U}_t} q_{it}$. Now suppose Q_t is not observed and the sample decomposition is performed using shares computed from the data available in the sample, where the sample is again defined by the sets $\mathbf{D}_{t_0} \in \mathbf{U}_{t_0}$ and $\mathbf{D}_{t_1} \in \mathbf{U}_{t_1}$. Let $Q_t^{\mathbf{D}_t} = \sum_{i\in\mathbf{D}_t} q_{it}$, so $\tilde{s}_{it} = q_{it}/Q_t^{\mathbf{D}_t}$. In this case applying Equation (2) to the sample captures the following:

$$\Delta\tilde{Y}_{t_1}^D = \sum_{i\in\tilde{\mathbf{C}}} \tilde{s}_{it_0}\Delta x_{it_1} + \sum_{i\in\tilde{\mathbf{C}}} \Delta\tilde{s}_{it_1}x_{it_1} + \sum_{i\in\tilde{\mathbf{E}}} \tilde{s}_{it_1}x_{it_1} - \sum_{i\in\tilde{\mathbf{L}}} \tilde{s}_{it_0}x_{it_0}.$$

This decomposition will also be biased. To see this, assume the sample is subject to misweighting and misclassification, and note that $\tilde{s}_{it} = [Q_t/Q_t^{\mathbf{D}_t}]s_{it}$. Following the same logic as outlined in Section 2 of the main text, the bias of each decomposition term is given by:

1. Within:

$$\sum_{i\in\mathbf{C}} \left[1 - \mathbb{E}\left[\frac{a_{it_0}a_{it1}}{Q_{t_0}^{\mathbf{D}_{t_0}}}\right] Q_{t_0}\right] s_{it_0}\Delta x_{it_1}$$

2. Between:

$$\sum_{i\in\mathbf{C}} \left[1 - \mathbb{E}\left[\frac{a_{it_0}a_{it1}}{Q_{t_1}^{\mathbf{D}_{t_1}}}\right] Q_{t_1}\right] s_{it_1}x_{it_1} - \sum_{i\in\mathbf{C}} \left[1 - \mathbb{E}\left[\frac{a_{it_0}a_{it1}}{Q_{t_0}^{\mathbf{D}_{t_0}}}\right] Q_{t_0}\right] s_{it_0}x_{it_1}$$

3. Entry:

$$\sum_{i\in\mathbf{E}} \left[1 - \mathbb{E}\left[\frac{a_{it1}}{Q_{t_1}^{\mathbf{D}_{t_1}}}\right] Q_{t_1}\right] s_{it_1}x_{it_1} - Q_{t_1}\sum_{i\in\mathbf{C}} \mathbb{E}\left[\frac{a_{it_1} - a_{it_0}a_{it_1}}{Q_{t_1}^{\mathbf{D}_{t_1}}}\right] s_{it_1}x_{it_1}$$

4. Exit:

$$-\sum_{i\in\mathbf{L}} \left[1 - \mathbb{E}\left[\frac{a_{it_0}}{Q_{t_0}^{\mathbf{D}_{t_0}}}\right] Q_{t_0}\right] s_{it_0}x_{it_0} + Q_{t_0}\sum_{i\in\mathbf{C}} \mathbb{E}\left[\frac{a_{it_0} - a_{it_0}a_{it_1}}{Q_{t_1}^{\mathbf{D}_{t_0}}}\right] s_{it_0}x_{it_0}$$

A.4 The Bias From Using Sample Data: The Dynamic Olley-Pakes Decomposition

In the main text, we noted that results analogous to Proposition 1 and Proposition 2 can be obtained for accounting decompositions based on Olley and Pakes (1996). Here, we show this formally for the dynamic Olley and Pakes decomposition suggested by Melitz and Polanec (2015).

Starting from Equation (1), following Melitz and Polanec the change in the outcome of interest across any two periods t_0 and t_1 can be decomposed as:

$$\begin{aligned}\Delta Y_{t_1} = &\left[\frac{1}{n^{\mathbf{C}}}\sum_{i\in\mathbf{C}}\Delta x_{it_1}\right] + \sum_{i\in\mathbf{C}}\Delta\left[\left[\hat{s}^{\mathbf{C}}_{it_1} - \frac{1}{n^{\mathbf{C}}}\sum_{i\in\mathbf{C}}\hat{s}^{\mathbf{C}}_{it_1}\right]\left[x_{it_1} - \frac{1}{n^{\mathbf{C}}}\sum_{i\in\mathbf{C}}x_{it_1}\right]\right] \\ &+ s^{\mathbf{E}}_{t_1}\left[\sum_{i\in\mathbf{E}}\left[\frac{s_{it_1}}{s^{\mathbf{E}}_{t_1}}\right]x_{it_1} - \sum_{i\in\mathbf{C}}\left[\frac{s_{it_1}}{s^{\mathbf{C}}_{t_1}}\right]x_{it_1}\right] + s^{\mathbf{L}}_{t_0}\left[\sum_{i\in\mathbf{C}}\left[\frac{s_{it_0}}{s^{\mathbf{C}}_{t_0}}\right]x_{it_0} - \sum_{i\in\mathbf{L}}\left[\frac{s_{it_0}}{s^{\mathbf{L}}_{t_0}}\right]x_{it_0}\right] \quad \text{(A.3)}\end{aligned}$$

where $n^{\mathbf{C}}$ is the number of agents in the continuing set C, $s^{\mathbf{G}}_t = \sum_{i\in\mathbf{G}} s_{it}$ is the aggregate market share of agents in set $\mathbf{G}$, and $\hat{s}^{\mathbf{G}}_{it} = s_{it}/s^{\mathbf{G}}_t$. The first term of Equation (A.3) is the change in Y due to a shift in the distribution of continuing agent outcomes, while the second is the change due to reallocations of economic activity across continuing agents (measured as the change in the covariance between agent shares and agent outcomes). The third term captures the change in Y due to entering agents, measured as the difference in the weighted average of outcomes between entering and continuing agents, whereas the fourth term captures the change in Y due to leaving agents, similarly measured as the difference in the weighted average of outcomes for continuing and leaving agents. For convenience, we refer to these four terms as the within, reallocation, entry and exit effects, respectively.

As with the BHC decomposition presented in the main text, when the complete set of agents in both $\mathbf{U}_{t_1}$ and $\mathbf{U}_{t_0}$ are observed, computing each component of Equation (A.3) is straightforward. If instead, the researcher has access to sample data defined by the sets $\mathbf{D}_{t_0} \subset \mathbf{U}_{t_0}$ and $\mathbf{D}_{t_1} \subset \mathbf{U}_{t_1}$, the decomposition captures the following:

$$\begin{aligned}\Delta Y^{\mathbf{D}}_{t_1} = &\left[\frac{1}{n^{\tilde{\mathbf{C}}}}\sum_{i\in\tilde{\mathbf{C}}}\Delta x_{it_1}\right] + \sum_{i\in\tilde{\mathbf{C}}}\Delta\left[\left[\hat{s}^{\tilde{\mathbf{C}}}_{it_1} - \frac{1}{n^{\tilde{\mathbf{C}}}}\sum_{i\in\tilde{\mathbf{C}}}\hat{s}^{\tilde{\mathbf{C}}}_{it_1}\right]\left[x_{it_1} - \frac{1}{n^{\tilde{\mathbf{C}}}}\sum_{i\in\tilde{\mathbf{C}}}x_{it_1}\right]\right] \\ &+ s^{\tilde{\mathbf{E}}}_{t_1}\left[\sum_{i\in\tilde{\mathbf{E}}}\left[\frac{s_{it_1}}{s^{\tilde{\mathbf{E}}}_{t_1}}\right]x_{it_1} - \sum_{i\in\tilde{\mathbf{C}}}\left[\frac{s_{it_1}}{s^{\tilde{\mathbf{C}}}_{t_1}}\right]x_{it_1}\right] + s^{\tilde{\mathbf{L}}}_{t_0}\left[\sum_{i\in\tilde{\mathbf{C}}}\left[\frac{s_{it_0}}{s^{\tilde{\mathbf{C}}}_{t_0}}\right]x_{it_0} - \sum_{i\in\tilde{\mathbf{L}}}\left[\frac{s_{it_0}}{s^{\tilde{\mathbf{L}}}_{t_0}}\right]x_{it_0}\right] \quad \text{(A.4)}\end{aligned}$$

where as in the main text, $\tilde{\mathbf{C}} = \{i | i \in \mathbf{D}_{t_0} \text{ and } i \in \mathbf{D}_{t_1}\}$ is the set of continuing agents in the sample, $\tilde{\mathbf{E}} = \{i | i \notin \mathbf{D}_{t_0} \text{ and } i \in \mathbf{D}_{t_1}\}$ is the set of agents that enter the sample at t_1 and $\tilde{\mathbf{L}} = \{i | i \in \mathbf{D}_{t_0} \text{ and } i \notin \mathbf{D}_{t_1}\}$ is the set of agents that leave the sample after t_0. Thus, $n^{\tilde{\mathbf{C}}}$ is the number of continuing agents in the sample. The biases that arise from the use of sample data can be determined by comparing the equivalent terms from Equation (A.3) and Equation (A.4), and will again depend on the underlying properties of the sample.

To start, first suppose that the sample is such that $\tilde{\mathbf{C}} \subset \mathbf{C}$, $\tilde{\mathbf{E}} \subset \mathbf{E}$, and $\tilde{\mathbf{L}} \subset \mathbf{L}$. Let $\eta_t^{\tilde{\mathbf{C}}} = n^{\tilde{\mathbf{C}}}/n^{\mathbf{C}}$, $\Xi_t^{\tilde{\mathbf{C}}} = s_t^{\tilde{\mathbf{C}}}/s_t^{\mathbf{C}}$, $\Xi_t^{\tilde{\mathbf{E}}} = s_t^{\tilde{\mathbf{E}}}/s_t^{\mathbf{E}}$, $\Xi_t^{\tilde{\mathbf{L}}} = s_t^{\tilde{\mathbf{L}}}/s_t^{\mathbf{L}}$, and a_{it} be an indicator for any observations sampled at time t such that $a_{it} = 1$ if and only if $i \in \mathbf{D}_t$. It is useful to then rewrite Equation (A.4) as:

$$\begin{aligned}
\Delta Y_{t_1}^{\mathbf{D}} = \frac{1}{n^{\mathbf{C}}} \sum_{i \in \mathbf{C}} \frac{a_{it_0} a_{it_1}}{\eta_t^{\tilde{\mathbf{C}}}} \Delta x_{it_1} \\
+ \sum_{i \in \mathbf{C}} \left[\frac{a_{it_0} a_{it_1}}{\Xi_{t_1}^{\tilde{\mathbf{C}}}} \right] \hat{s}_{it_1}^{\mathbf{C}} x_{it_1} - \sum_{i \in \mathbf{C}} \left[\frac{a_{it_0} a_{it_1}}{\Xi_{t_0}^{\tilde{\mathbf{C}}}} \right] \hat{s}_{it_0}^{\mathbf{C}} x_{it_0} - \frac{1}{n^{\mathbf{C}}} \sum_{i \in \mathbf{C}} \frac{a_{it_0} a_{it_1}}{\eta_t^{\tilde{\mathbf{C}}}} \Delta x_{it_1} \\
+ \sum_{i \in \mathbf{E}} a_{it_1} s_{it_1} x_{it_1} - \left[\frac{s_{t_1}^{\mathbf{E}}}{s_{t_1}^{\mathbf{C}}} \right] \sum_{i \in \mathbf{C}} \left[\left[\frac{\Xi_{t_1}^{\tilde{\mathbf{E}}}}{\Xi_{t_1}^{\tilde{\mathbf{C}}}} \right] a_{it_0} a_{it_1} \right] \hat{s}_{it_1}^{\mathbf{C}} x_{it_1} \\
+ \left[\frac{s_{t_0}^{\mathbf{L}}}{s_{t_0}^{\mathbf{C}}} \right] \sum_{i \in \mathbf{C}} \left[\left[\frac{\Xi_{t_0}^{\tilde{\mathbf{L}}}}{\Xi_{t_0}^{\tilde{\mathbf{C}}}} \right] a_{it_0} a_{it_1} \right] \hat{s}_{it_0}^{\mathbf{C}} x_{it_0} - \sum_{i \in \mathbf{L}} a_{it_0} s_{it_0} x_{it_0} \quad \text{(A.5)}
\end{aligned}$$

where we have made use of the fact that the sample reallocation effect (the second term on the right-hand side of Equation (A.4)) can be simplified to $\sum_{i \in \tilde{\mathbf{C}}} \Delta[\hat{s}_{it_1}^{\tilde{\mathbf{C}}} x_{it_1}] - [1/n^{\tilde{\mathbf{C}}}] \sum_{i \in \tilde{\mathbf{C}}} \Delta x_{it_1}$. In Equation (A.5), the a's, Ξ's, and η's are the only random variables in this expression because they are the only elements determined by sampling. All other elements are fixed characteristics of observations.[1]

As such, after taking expectations and noting $\mathbb{E}[a_{it_0}] = \Pr(i \in \mathbf{D}_{t_0})$ and $\mathbb{E}[a_{it_1}] = \Pr(i \in$

[1]Though the sample size is fixed, the share of sampled firms in each group is not fixed, which makes $n^{\tilde{\mathbf{C}}}$ random.

$\mathbf{D}_{t_1}$), this equation can be rewritten as:

$$\begin{aligned}
\mathbb{E}\left[\Delta Y_{t_1}^{\mathbf{D}}\right] = \frac{1}{n^{\mathbf{C}}}\sum_{i\in\mathbf{C}}\mathbb{E}\left[\frac{a_{it_0}a_{it_1}}{\eta_t^{\tilde{\mathbf{C}}}}\right]\Delta x_{it_1} \\
+\sum_{i\in\mathbf{C}}\mathbb{E}\left[\frac{a_{it_0}a_{it_1}}{\Xi_{t_1}^{\tilde{\mathbf{C}}}}\right]\hat{s}_{it_1}^{\mathbf{C}}x_{it_1} - \sum_{i\in\mathbf{C}}\mathbb{E}\left[\frac{a_{it_0}a_{it_1}}{\Xi_{t_0}^{\tilde{\mathbf{C}}}}\right]\hat{s}_{it_0}^{\mathbf{C}}x_{it_0} - \frac{1}{n^{\mathbf{C}}}\sum_{i\in\mathbf{C}}\mathbb{E}\left[\frac{a_{it_0}a_{it_1}}{\eta_t^{\tilde{\mathbf{C}}}}\right]\Delta x_{it_1} \\
+\sum_{i\in\mathbf{E}}\Pr(i\in\mathbf{D}_{t_1})s_{it_1}x_{it_1} - \left[\frac{s_{t_1}^{\mathbf{E}}}{s_{t_1}^{\mathbf{C}}}\right]\sum_{i\in\mathbf{C}}\mathbb{E}\left[\left[\frac{\Xi_{t_1}^{\tilde{\mathbf{E}}}}{\Xi_{t_1}^{\tilde{\mathbf{C}}}}\right]a_{it_0}a_{it_1}\right]s_{it_1}x_{it_1} \\
+\left[\frac{s_{t_0}^{\mathbf{L}}}{s_{t_0}^{\mathbf{C}}}\right]\sum_{i\in\mathbf{C}}\mathbb{E}\left[\left[\frac{\Xi_{t_0}^{\tilde{\mathbf{L}}}}{\Xi_{t_0}^{\tilde{\mathbf{C}}}}\right]a_{it_0}a_{it_1}\right]s_{it_0}x_{it_0} - \sum_{i\in\mathbf{L}}\Pr(i\in\mathbf{D}_{t_0})s_{it_0}x_{it_0} \quad \text{(A.6)}
\end{aligned}$$

Equation (A.6) indicates that, as with the BHC-type decomposition presented in the main text, the expected value of the dynamic Olley and Pakes decomposition depends on the probability with which agents in the true sets of entering, exiting, and continuing agents are observed in the sample the decomposition is applied to. With some algebra, subtracting Equation (A.6) from Equation (A.3) yields:

Proposition A.3. *Suppose the researcher has access to sample data given by the set $\boldsymbol{D}_{t_0} \cup \boldsymbol{D}_{t_1}$ where $\boldsymbol{D}_{t_0} \subset \boldsymbol{U}_{t_0}$ and $\boldsymbol{D}_{t_1} \subset \boldsymbol{U}_{t_1}$, and suppose the sample design is such that $\tilde{\boldsymbol{C}} \subset \boldsymbol{C}$, $\tilde{\boldsymbol{E}} \subset \boldsymbol{E}$, and $\tilde{\boldsymbol{L}} \subset \boldsymbol{L}$.*

1. *If the sample is random, that is if $Pr(i \in \boldsymbol{D}_t) = \gamma_t < 1 \ \forall \ i$, the within, reallocation, entry, and exit effects are biased. The biases of these terms are given by:*

   (a) *Within:*

$$\frac{1}{n^{\boldsymbol{C}}}\sum_{i\in\boldsymbol{C}}\left[1-\mathbb{E}\left[\frac{a_{it_0}a_{it_1}}{\eta_t^{\tilde{\boldsymbol{C}}}}\right]\right]\Delta x_{it_1}$$

   (b) *Reallocation:*

$$\begin{aligned}
\sum_{i\in\boldsymbol{C}}\left[1-S_{t_1}^{\boldsymbol{C}}\mathbb{E}\left[\frac{a_{it_0}a_{it_1}}{\sum_{i\in\boldsymbol{C}}a_{it_0}a_{it_1}s_{it_1}}\right]\right]\hat{s}_{it_1}^{\boldsymbol{C}}x_{it_1} \\
-\sum_{i\in\boldsymbol{C}}\left[1-S_{t_0}^{\boldsymbol{C}}\mathbb{E}\left[\frac{a_{it_0}a_{it_1}}{\sum_{i\in\boldsymbol{C}}a_{it_0}a_{it_1}s_{it_0}}\right]\right]\hat{s}_{it_0}^{\boldsymbol{C}}x_{it_0} \\
-\frac{1}{n^{\boldsymbol{C}}}\sum_{i\in\boldsymbol{C}}\left[1-\mathbb{E}\left[\frac{a_{it_0}a_{it_1}}{\eta_t^{\tilde{\boldsymbol{C}}}}\right]\right]\Delta x_{it_1}
\end{aligned}$$

(c) *Entry:*

$$[1-\gamma_{t_1}]\sum_{i\in\boldsymbol{E}} s_{it_1}x_{it_1} - \left[\frac{s^{\boldsymbol{E}}_{t_1}}{s^{\boldsymbol{C}}_{t_1}}\right]\sum_{i\in\boldsymbol{C}}\left[1-\mathbb{E}\left[\left[\frac{\Xi^{\tilde{\boldsymbol{E}}}_{t_1}}{\Xi^{\tilde{\boldsymbol{C}}}_{t_1}}\right]a_{it_0}a_{it_1}\right]\right]\hat{s}^{\boldsymbol{C}}_{it_1}x_{it_1}$$

(d) *Exit:*

$$\left[\frac{s^{\boldsymbol{L}}_{t_0}}{s^{\boldsymbol{C}}_{t_0}}\right]\sum_{i\in\boldsymbol{C}}\left[1-\mathbb{E}\left[\left[\frac{\Xi^{\tilde{\boldsymbol{L}}}_{t_0}}{\Xi^{\tilde{\boldsymbol{C}}}_{t_0}}\right]a_{it_0}a_{it_1}\right]\right]\hat{s}^{\boldsymbol{C}}_{it_0}x_{it_0} - [1-\gamma_{t_0}]\sum_{i\in\boldsymbol{L}} s_{it_0}x_{it_0}$$

2. *If the sample is non-random, the within, reallocation, entry and exit effects computed by applying Equation* (A.3) *to the sample are biased. The biases of each effect are given by:*

(a) *Within:*

$$\frac{1}{n^{\boldsymbol{C}}}\sum_{i\in\boldsymbol{C}}\left[1-\mathbb{E}\left[\frac{a_{it_0}a_{it_1}}{\eta^{\tilde{\boldsymbol{C}}}_t}\right]\right]\Delta x_{it_1}$$

(b) *Reallocation:*

$$\begin{aligned}\sum_{i\in\boldsymbol{C}}\left[1-S^{\boldsymbol{C}}_{t_1}\mathbb{E}\left[\frac{a_{it_0}a_{it_1}}{\sum_{i\in\boldsymbol{C}}a_{it_0}a_{it_1}s_{it_1}}\right]\right]\hat{s}^{\boldsymbol{C}}_{it_1}x_{it_1}&\\ -\sum_{i\in\boldsymbol{C}}\left[1-S^{\boldsymbol{C}}_{t_0}\mathbb{E}\left[\frac{a_{it_0}a_{it_1}}{\sum_{i\in\boldsymbol{C}}a_{it_0}a_{it_1}s_{it_0}}\right]\right]\hat{s}^{\boldsymbol{C}}_{it_0}x_{it_0}&\\ -\frac{1}{n^{\boldsymbol{C}}}\sum_{i\in\boldsymbol{C}}\left[1-\mathbb{E}\left[\frac{a_{it_0}a_{it_1}}{\eta^{\tilde{\boldsymbol{C}}}_t}\right]\right]\Delta x_{it_1}&\end{aligned}$$

(c) *Entry:*

$$\sum_{i\in\boldsymbol{E}}[1-\Pr(i\in\boldsymbol{D}_{t_1})]\,s_{it_1}x_{it_1} - \left[\frac{s^{\boldsymbol{E}}_{t_1}}{s^{\boldsymbol{C}}_{t_1}}\right]\sum_{i\in\boldsymbol{C}}\left[1-\mathbb{E}\left[\left[\frac{\Xi^{\tilde{\boldsymbol{E}}}_{t_1}}{\Xi^{\tilde{\boldsymbol{C}}}_{t_1}}\right]a_{it_0}a_{it_1}\right]\right]\hat{s}^{\boldsymbol{C}}_{it_1}x_{it_1}$$

(d) *Exit:*

$$\left[\frac{s^{\boldsymbol{L}}_{t_0}}{s^{\boldsymbol{C}}_{t_0}}\right]\sum_{i\in\boldsymbol{C}}\left[1-\mathbb{E}\left[\left[\frac{\Xi^{\tilde{\boldsymbol{L}}}_{t_0}}{\Xi^{\tilde{\boldsymbol{C}}}_{t_0}}\right]a_{it_0}a_{it_1}\right]\right]\hat{s}^{\boldsymbol{C}}_{it_0}x_{it_0} - \sum_{i\in\boldsymbol{L}}[1-\Pr(i\in\boldsymbol{D}_{t_0}]\,s_{it_0}x_{it_0}$$

Proof. Part 1 of the proposition follows from subtracting Equation (A.6) from Equation (A.3) and using $\Pr(i\in\mathbf{D}_{t_0})=\gamma_{t_0}$ and $\Pr(i\in\mathbf{D}_{t_1})=\gamma_{t_1}$ under random sampling. Part

2 of the proposition follows from subtracting Equation (A.6) from Equation (A.3). □

Proposition A.3 shows that the application of a Melitz and Polanec decomposition to sample data results in biased estimates of the within, reallocation, entry, and exit effects when only a subset of continuing, entering, and exiting agents are observed, respectively. This bias is present regardless of whether the sample is random or not.[2] These biases again arise due to misweighting; with sample data, agents that are not observed are "misweighted" and effectively assigned a weight of zero in the calculation of the relevant sum.

We now turn to consider the case where the sample possessed by the researcher is also subject to misclassification, such that the true status of entering and exiting agents is not observed. That is, the case where the sample is such that $\tilde{\mathbf{C}} \subset \mathbf{C}$, with $\tilde{\mathbf{E}} \subset \mathbf{E} \cup \{\mathbf{C} \backslash \tilde{\mathbf{C}}\}$, and $\tilde{\mathbf{L}} \subset \mathbf{L} \cup \{\mathbf{C} \backslash \tilde{\mathbf{C}}\}$. In this case, Equation (A.4) can be rewritten as:

$$
\begin{aligned}
\Delta Y_{t_1}^{\mathbf{D}} = & \frac{1}{n^{\mathbf{C}}} \sum_{i \in \mathbf{C}} \frac{a_{it_0} a_{it_1}}{\eta_t^{\tilde{\mathbf{C}}}} \Delta x_{it_1} \\
& + \sum_{i \in \mathbf{C}} \left[\frac{a_{it_0} a_{it_1}}{\Xi_{t_1}^{\tilde{\mathbf{C}}}} \right] \hat{s}_{it_1}^{\mathbf{C}} x_{it_1} - \sum_{i \in \mathbf{C}} \left[\frac{a_{it_0} a_{it_1}}{\Xi_{t_0}^{\tilde{\mathbf{C}}}} \right] \hat{s}_{it_0}^{\mathbf{C}} x_{it_0} - \frac{1}{n^{\mathbf{C}}} \sum_{i \in \mathbf{C}} \frac{a_{it_0} a_{it_1}}{\eta_t^{\tilde{\mathbf{C}}}} \Delta x_{it_1} \\
& + \sum_{i \in \mathbf{E}} a_{it_1} s_{it_1} x_{it_1} + \sum_{i \in \mathbf{C}} a_{it_1} [1 - a_{it_0}] s_{it_1} x_{it_1} - \left[\frac{s_{t_1}^{\mathbf{E}}}{s_{t_1}^{\mathbf{C}}} \right] \sum_{i \in \mathbf{C}} \left[\left[\frac{\Xi_{t_1}^{\tilde{\mathbf{E}}}}{\Xi_{t_1}^{\tilde{\mathbf{C}}}} \right] a_{it_0} a_{it_1} \right] \hat{s}_{it_1}^{\mathbf{C}} x_{it_1} \\
& + \left[\frac{s_{t_0}^{\mathbf{L}}}{s_{t_0}^{\mathbf{C}}} \right] \sum_{i \in \mathbf{C}} \left[\left[\frac{\Xi_{t_0}^{\tilde{\mathbf{L}}}}{\Xi_{t_0}^{\tilde{\mathbf{C}}}} \right] a_{it_0} a_{it_1} \right] \hat{s}_{it_0}^{\mathbf{C}} x_{it_0} - \sum_{i \in \mathbf{L}} a_{it_0} s_{it_0} x_{it_0} - \sum_{i \in \mathbf{L}} [1 - a_{it_1}] a_{it_0} s_{it_0} x_{it_0} \quad \text{(A.7)}
\end{aligned}
$$

[2] An important note here is that under random sampling, although the naive application of the Melitz and Polanec yields a biased within effect, this bias will be approximately zero as it occurs because the within-effect is a ratio estimator.

Taking expectations yields:

$$\begin{aligned}
\mathbb{E}\left[\Delta Y_{t_1}^{\mathbf{D}}\right] = \frac{1}{n^{\mathbf{C}}}\sum_{i\in\mathbf{C}}\mathbb{E}\left[\frac{a_{it_0}a_{it_1}}{\eta_t^{\tilde{\mathbf{C}}}}\right]\Delta x_{it_1} \\
+\sum_{i\in\mathbf{C}}\mathbb{E}\left[\frac{a_{it_0}a_{it_1}}{\Xi_{t_1}^{\tilde{\mathbf{C}}}}\right]\hat{s}_{it_1}^{\mathbf{C}}x_{it_1} - \sum_{i\in\mathbf{C}}\mathbb{E}\left[\frac{a_{it_0}a_{it_1}}{\Xi_{t_0}^{\tilde{\mathbf{C}}}}\right]\hat{s}_{it_0}^{\mathbf{C}}x_{it_0} - \frac{1}{n^{\mathbf{C}}}\sum_{i\in\mathbf{C}}\mathbb{E}\left[\frac{a_{it_0}a_{it_1}}{\eta_t^{\tilde{\mathbf{C}}}}\right]\Delta x_{it_1} \\
+\sum_{i\in\mathbf{E}}\Pr(i\in\mathbf{D}_{t_1})s_{it_1}x_{it_1} + \sum_{i\in\mathbf{C}}\left[\Pr(i\in\mathbf{D}_{t_1}) - \Pr(i\in\mathbf{D}_{t_0}\cap\mathbf{D}_{t_1})\right]s_{it_1}x_{it_1} \\
-\left[\frac{s_{t_1}^{\mathbf{E}}}{s_{t_1}^{\mathbf{C}}}\right]\sum_{i\in\mathbf{C}}\mathbb{E}\left[\left[\frac{\Xi_{t_1}^{\tilde{\mathbf{E}}}}{\Xi_{t_1}^{\tilde{\mathbf{C}}}}\right]a_{it_0}a_{it_1}\right]s_{it_1}x_{it_1} \\
+\left[\frac{s_{t_0}^{\mathbf{L}}}{s_{t_0}^{\mathbf{C}}}\right]\sum_{i\in\mathbf{C}}\mathbb{E}\left[\left[\frac{\Xi_{t_0}^{\tilde{\mathbf{L}}}}{\Xi_{t_0}^{\tilde{\mathbf{C}}}}\right]a_{it_0}a_{it_1}\right]s_{it_0}x_{it_0} - \sum_{i\in\mathbf{L}}\Pr(i\in\mathbf{D}_{t_0})s_{it_0}x_{it_0} \\
-\sum_{i\in\mathbf{L}}\left[\Pr(i\in\mathbf{D}_{t_0}) - \Pr(i\in\mathbf{D}_{t_0}\cap\mathbf{D}_{t_1})\right]s_{it_0}x_{it_0} \quad \text{(A.8)}
\end{aligned}$$

Equation (A.8) highlights how misclassification is also an issue for Melitz and Polanec-type decompositions. For example, consider the sample entry effect given by the third and fourth lines of Equation (A.8). The sample effect now includes observations from the set of continuing agents that are only sampled in the second period and are misclassified as entering agents, given by the third term of this expression. A similar issue arises for the sample exit effect. Formally, we have:

Proposition A.4. *Suppose the researcher has access to sample data given by the set $\boldsymbol{D}_{t_0}\cup\boldsymbol{D}_{t_1}$ where $\boldsymbol{D}_{t_0}\subset\boldsymbol{U}_{t_0}$ and $\boldsymbol{D}_{t_1}\subset\boldsymbol{U}_{t_1}$, and suppose the sample design is such that $\tilde{\boldsymbol{C}}\subset\boldsymbol{C}$, $\tilde{\boldsymbol{E}}\subset\boldsymbol{E}\cup\{\boldsymbol{C}\backslash\tilde{\boldsymbol{C}}\}$, and $\tilde{\boldsymbol{L}}\subset\boldsymbol{L}\cup\{\boldsymbol{C}\backslash\tilde{\boldsymbol{C}}\}$.*

1. *If the sample is random, that is if $Pr(i\in\boldsymbol{D}_t)=\gamma_t<1\ \forall\ i$, the within, reallocation, entry, and exit effects are biased. The biases of these terms are given by:*

   (a) *Within:*

$$\frac{1}{n^{\boldsymbol{C}}}\sum_{i\in\boldsymbol{C}}\left[1-\mathbb{E}\left[\frac{a_{it_0}a_{it_1}}{\eta_t^{\tilde{\boldsymbol{C}}}}\right]\right]\Delta x_{it_1}$$

(b) *Reallocation:*

$$\sum_{i\in\boldsymbol{C}}\left[1-S_{t_1}^{\boldsymbol{C}}\mathbb{E}\left[\frac{a_{it_0}a_{it_1}}{\sum_{i\in\boldsymbol{C}}a_{it_0}a_{it_1}s_{it_1}}\right]\right]\hat{s}_{it_1}^{\boldsymbol{C}}x_{it_1} - \sum_{i\in\boldsymbol{C}}\left[1-S_{t_0}^{\boldsymbol{C}}\mathbb{E}\left[\frac{a_{it_0}a_{it_1}}{\sum_{i\in\boldsymbol{C}}a_{it_0}a_{it_1}s_{it_0}}\right]\right]\hat{s}_{it_0}^{\boldsymbol{C}}x_{it_0} - \frac{1}{n^{\boldsymbol{C}}}\sum_{i\in\boldsymbol{C}}\left[1-\mathbb{E}\left[\frac{a_{it_0}a_{it_1}}{\eta_t^{\tilde{\boldsymbol{C}}}}\right]\right]\Delta x_{it_1}$$

(c) *Entry:*

$$[1-\gamma_{t_1}]\sum_{i\in\boldsymbol{E}}s_{it_1}x_{it_1} - [\gamma_{t_1}-\gamma_{t_0}\gamma_{t_1}]\sum_{i\in\boldsymbol{C}}s_{it_1}x_{it_1} - \left[\frac{s_{t_1}^{\boldsymbol{E}}}{s_{t_1}^{\boldsymbol{C}}}\right]\sum_{i\in\boldsymbol{C}}\left[1-\mathbb{E}\left[\left[\frac{\Xi_{t_1}^{\tilde{\boldsymbol{E}}}}{\Xi_{t_1}^{\tilde{\boldsymbol{C}}}}\right]a_{it_0}a_{it_1}\right]\right]\hat{s}_{it_1}^{\boldsymbol{C}}x_{it_1}$$

(d) *Exit:*

$$\left[\frac{s_{t_0}^{\boldsymbol{L}}}{s_{t_0}^{\boldsymbol{C}}}\right]\sum_{i\in\boldsymbol{C}}\left[1-\mathbb{E}\left[\left[\frac{\Xi_{t_0}^{\tilde{\boldsymbol{L}}}}{\Xi_{t_0}^{\tilde{\boldsymbol{C}}}}\right]a_{it_0}a_{it_1}\right]\right]\hat{s}_{it_0}^{\boldsymbol{C}}x_{it_0} - [1-\gamma_{t_0}]\sum_{i\in\boldsymbol{L}}s_{it_0}x_{it_0} + [\gamma_{t_0}-\gamma_{t_0}\gamma_{t_1}]\sum_{i\in\boldsymbol{C}}s_{it_0}x_{it_0}$$

2. *If the sample is non-random, the within, reallocation, entry and exit effects computed by applying Equation* (A.3) *to the sample are biased. The biases of each effect are given by:*

(a) *Within:*

$$\frac{1}{n^{\boldsymbol{C}}}\sum_{i\in\boldsymbol{C}}\left[1-\mathbb{E}\left[\frac{a_{it_0}a_{it_1}}{\eta_t^{\tilde{\boldsymbol{C}}}}\right]\right]\Delta x_{it_1}$$

(b) Reallocation:

$$\sum_{i\in \boldsymbol{C}}\left[1-S_{t_1}^{\boldsymbol{C}}\mathbb{E}\left[\frac{a_{it_0}a_{it_1}}{\sum_{i\in \boldsymbol{C}}a_{it_0}a_{it_1}s_{it_1}}\right]\right]\hat{s}_{it_1}^{\boldsymbol{C}}x_{it_1} - \sum_{i\in \boldsymbol{C}}\left[1-S_{t_0}^{\boldsymbol{C}}\mathbb{E}\left[\frac{a_{it_0}a_{it_1}}{\sum_{i\in \boldsymbol{C}}a_{it_0}a_{it_1}s_{it_0}}\right]\right]\hat{s}_{it_0}^{\boldsymbol{C}}x_{it_0} - \frac{1}{n^{\boldsymbol{C}}}\sum_{i\in \boldsymbol{C}}\left[1-\mathbb{E}\left[\frac{a_{it_0}a_{it_1}}{\eta_t^{\tilde{\boldsymbol{C}}}}\right]\right]\Delta x_{it_1}$$

(c) Entry:

$$\sum_{i\in \boldsymbol{E}}\left[1-\Pr(i\in \boldsymbol{D}_{t_1})\right]s_{it_1}x_{it_1} - \sum_{i\in \boldsymbol{C}}[\Pr(i\in \boldsymbol{D}_{t_1})-\Pr(i\in \boldsymbol{D}_{t_1}\cap \boldsymbol{D}_{t_0})]s_{it_1}x_{it_1} - \left[\frac{s_{t_1}^{\boldsymbol{E}}}{s_{t_1}^{\boldsymbol{C}}}\right]\sum_{i\in \boldsymbol{C}}\left[1-\mathbb{E}\left[\left[\frac{\Xi_{t_1}^{\tilde{\boldsymbol{E}}}}{\Xi_{t_1}^{\tilde{\boldsymbol{C}}}}\right]a_{it_0}a_{it_1}\right]\right]\hat{s}_{it_1}^{\boldsymbol{C}}x_{it_1}$$

(d) Exit:

$$\left[\frac{s_{t_0}^{\boldsymbol{L}}}{s_{t_0}^{\boldsymbol{C}}}\right]\sum_{i\in \boldsymbol{C}}\left[1-\mathbb{E}\left[\left[\frac{\Xi_{t_0}^{\tilde{\boldsymbol{L}}}}{\Xi_{t_0}^{\tilde{\boldsymbol{C}}}}\right]a_{it_0}a_{it_1}\right]\right]\hat{s}_{it_0}^{\boldsymbol{C}}x_{it_0} - \sum_{i\in \boldsymbol{L}}\left[1-\Pr(i\in \boldsymbol{D}_{t_0})\right]s_{it_0}x_{it_0} + \sum_{i\in \boldsymbol{C}}[\Pr(i\in \boldsymbol{D}_{t_0})-\Pr(i\in \boldsymbol{D}_{t_1}\cap \boldsymbol{D}_{t_0})]s_{it_0}x_{it_0}$$

Proof. To prove Part 1 of the proposition, first note that the bias of the within and reallocation effects follow directly from the proof of Proposition A.3. The bias of the entry and exit effects follow from subtracting Equation (A.8) from Equation (A.3) and using the fact that random sampling implies $\Pr(i \in \mathbf{D}_{t_0}) = \gamma_{t_0}$, $\Pr(i \in \mathbf{D}_{t_1}) = \gamma_{t_1}$, and $\Pr(i \in \mathbf{D}_{t_0} \cap \mathbf{D}_{t_1}) = \gamma_{t_0}\gamma_{t_1}$.

Part 2 of the proposition follows from subtracting Equation (A.8) from Equation (A.3). □

Propositions A.3 and A.4 demonstrate that the results contained in Propositions 1 and 2 are not specific to the BHC-type decompositions considered in the main text.

A.5 Consistency of the HT and Adjusted HT estimators

Here we examine the consistency of the HT and Adjusted HT estimators. We first establish conditions under which $\widehat{WE}_{t_1}^{HT}$ and $\widehat{BE}_{t_1}^{HT}$ are consistent.

Lemma 1. *Let $\rho^c_{ijt_1} = \Pr(i, j \in \boldsymbol{D}_{t_0} \cap \boldsymbol{D}_{t_1} | i \in \boldsymbol{C})$ denote the probability with which units i and j are sampled at both t_0 and t_1 conditional on surviving. Under Assumptions 1 and 2:*

1. $\widehat{WE}^{HT}_{t_1}$ *is a consistent estimator of* WE_{t_1}.
2. $\widehat{BE}^{HT}_{t_1}$ *is a consistent estimator of* BE_{t_1}.

provided that the sample design satisfies the following regulatory conditions:

A) $\rho^c_{it_1} \to 1 \ \forall i$ *as* $n \to \infty$.

B) $\rho^c_{ijt_1} \to 1 \ \forall i \neq j$ *as* $n \to \infty$.

Proof. To begin note that the consistency of $\widehat{WE}^{HT}_{t_1}$ requires that:

$$\lim_{n\to\infty} \Pr\left(|\widehat{WE}^{HT}_{t_1} - WE_{t_1}| \geqslant \epsilon\right) = 0$$

for all $\epsilon > 0$ where n denotes sample size. By Chebyshev's Inequality:

$$\begin{aligned}\Pr\left(|\widehat{WE}^{HT}_{t_1} - WE_{t_1}| \geqslant \epsilon\right) &= \Pr\left(\left[\widehat{WE}^{HT}_{t_1} - WE_{t_1}\right]^2 \geqslant \epsilon^2\right) \\ &\leqslant \frac{\mathbb{E}\left[\left[\widehat{WE}^{HT}_{t_1} - WE_{t_1}\right]^2\right]}{\epsilon^2}.\end{aligned}$$

By definition, the Mean Squared Error of $\widehat{WE}^{HT}_{t_1}$ is

$$MSE\left(\widehat{WE}^{HT}_{t_1}\right) = \mathbb{E}\left[\widehat{WE}^{HT}_{t_1} - WE_{t_1}\right]^2.$$

Therefore, it suffices to show $\lim_{n\to\infty} MSE\left(\widehat{WE}^{HT}_{t_1}\right) = 0$. Note that:

$$\begin{aligned}MSE\left(\widehat{WE}^{HT}_{t_1}\right) &= Var\left(\widehat{WE}^{HT}_{t_1}\right) + \left[Bias\left(\widehat{WE}^{HT}_{t_1}\right)\right]^2 \\ &= Var\left(\widehat{WE}^{HT}_{t_1}\right)\end{aligned}$$

by Proposition 3. Thus, a sufficient condition is $\lim_{n\to\infty} Var\left(\widehat{WE}^{HT}_{t_1}\right) = 0$.

By definition,

$$Var\left(\widehat{WE}_{t_1}^{HT}\right) = Var\left(\sum_{i\in\mathbf{C}} \frac{s_{it_0}\Delta x_{it_1} a^c_{it_1}}{\rho^c_{it_1}}\right)$$
$$= \sum_{i\in\mathbf{C}} \frac{s_{it_0}\Delta x_{it_1} Var(a^c_{it_1})}{\rho^c_{it_1}} + \sum_{i\in\mathbf{C}}\sum_{\substack{j\in\mathbf{C}\\ j\neq i}} \left[\frac{s_{it_0}\Delta x_{it_1}}{\rho^c_{it_1}}\right]\left[\frac{w_{jt_0}\Delta x_{jt_1}}{\rho^c_{jt_1}}\right] Cov(a^c_{it_1}, a^c_{jt_1}).$$

Given that $a^c_{it_1}$ is a Bernoulli distributed random variable, $Var(a^c_{it_1}) = \rho^c_{it_1}[1-\rho^c_{it_1}]$ and $Cov(a^c_{it}, a^c_{jt}) = \rho^c_{ijt_1} - \rho^c_{it_1}\rho^c_{jt_1}$. Thus:

$$Var\left(\widehat{WE}_{t_1}^{HT}\right) = \sum_{i\in\mathbf{C}} s^2_{it_0}[\Delta x_{it_1}]^2 \frac{1-\rho^c_{it_1}}{\rho^c_{it_1}} + \sum_{i\in\mathbf{C}}\sum_{\substack{j\in\mathbf{C}\\ j\neq i}} \left[s_{it_0}\Delta x_{it_1}\right]\left[s_{jt_0}\Delta x_{jt_1}\right] \frac{\rho^c_{ijt_1} - \rho^c_{it_1}\rho^c_{jt_1}}{\rho^c_{it_1}\rho^c_{jt_1}}.$$

It follows that $\lim_{n\to\infty} Var(\widehat{WE}_{t_1}^{HT}) = 0$, and $\widehat{WE}_{t_1}^{HT}$ is a consistent estimator of WE_{t_1}, under the regulatory conditions in the proposition.

The conditions for the consistency of $\widehat{BE}_{t_1}^{HT}$ can be derived analogously. □

In essence, the two regulatory conditions in Lemma 1 ensure that the sample design is such that the sample coverage uniformly improves across both periods as the sample size increases. This rules out repeat sampling in any given period, oversampling of one period relative to another, or any sample design that systematically omits units, as in the limit all units must be observed.

The adjusted HT estimators of both the within and between effects will also be consistent under the regulatory conditions in Lemma 1.

Lemma 2. *If Assumption 2 fails but Assumption 1 and conditions A) and B) in Lemma 1 hold, then* $\widehat{WE}_{t_1}^{AHT}$ *and* $\widehat{BE}_{t_1}^{AHT}$ *are consistent estimators of* WE_{t_1} *and* BE_{t_1}*, respectively.*

Proof. We start by showing $\widehat{WE}_{t_1}^{AHT}$ is a consistent estimator of WE_{t_1}. First, following Proposition 3 and Lemma 1, it can be shown $\widehat{WE}_{t}^{HT,S}$ converges to $\frac{1}{S^S_{t_0}} WE_{t_1}$.

Note that as $\widehat{S}^s_{t_0}$ is a HT estimator using inclusion probability weights from period t_0, this will be a consistent estimator under weaker conditions than the previous corollary (recall that regulatory conditions A) and B) imply that the limits of ρ_{it_0} and ρ_{ijt_0} are 1).

By the Continuous Mapping Theorem, $\widehat{WE}_t^{AHT}$ converges in probability to:

$$S_{t_0}^S \frac{1}{S_{t_0}^S} WE_t = WE_t.$$

The consistency of $\widehat{BE}_{t_1}^{AHT}$ can be shown analogously. □

B Additional Monte-Carlo Results

B.1 HT Estimator Performance

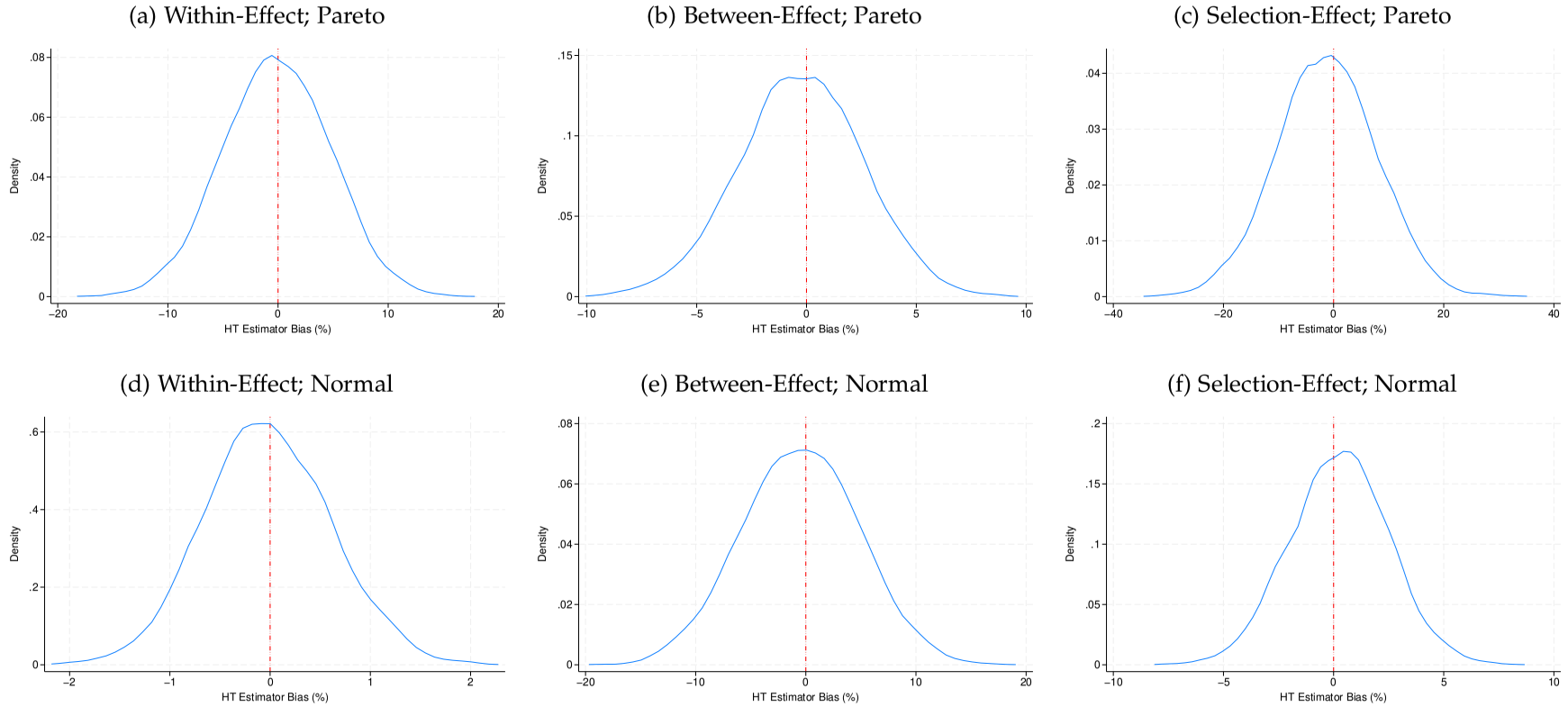
This section shows results of the simulations assessing the performance of the non-adjusted HT estimator. We use the simulation environment described in Section 3.4 to assess the non-adjusted HT estimator. For each simulation run, we compute the percentage difference between the true decomposition term and the term computed by the HT estimator. The mean and median "errors" are shown in Table B.5, with the full distributions shown in Figure B.5. As with the adjusted HT estimator, the HT estimator performs well in these simulations.

Table B.5: HT Estimator Simulation Results

	Mean (1)	Median (2)
	Panel (a): Pareto Distribution	
Within-Effect	-0.12	-0.14
Between-Effect	-0.26	-0.21
Selection-Effect	-1.25	-1.22
	Panel (b): Bivariate Normal Distribution	
Within-Effect	-0.02	-0.04
Between-Effect	-0.48	-0.46
Selection-Effect	0.26	0.31

Notes: Table reports results from two sets of Monte Carlo simulations (each with 5,000 repetitions) testing the bias in our proposed HT estimators of the decomposition terms of Equation (3). Mean and median estimation biases (in %) are reported for each term. The first simulation (Panel (a)) generates population variables from a Pareto distribution with a fixed shape and scale parameter (N=50,000). The second simulation (Panel (b)) generates population variables from a bivariate normal distribution with a fixed mean and correlation structure (N=50,000). In both simulations, a sample of observations are drawn from the population following a size-based sampling design with fixed sample probabilities.

Figure B.5: HT Estimator Simulation Results



Notes: Figure shows BHC Decomposition estimation bias (in %) from a Monte Carlo Simulation (5,000 repetitions) for our proposed HT estimator. Results from two separate simulations are shown. The first simulation (panels (a) through (c)) generates population variables from a Pareto distribution with a fixed shape and scale parameter (N=50,000). The second simulation (panels (d) through (f)) generates population variables from a multivariate normal distribution with a fixed mean and correlation structure (N=50,000). In both simulations, a sample of observations are drawn from the population following a size-based sampling design with fixed sample probabilities.

B.2 Log-Normal Distribution

This section shows results of the log-normal Monte Carlo simulation. The parameters used to generate the population data are shown in Table B.6. The simulation setup mirrors the approach taken in the main text.

Table B.6: Log-Normal Distribution Parameters

	Period 1	Period 2
$E[l_{it}]$	148.4	403.4
$E[z_{it}]$	20.1	7.4
$Cov(l_{it}, z_{it})$	0.4	0.6

Notes: Parameters for the multivariate log-normal distribution used to generate simulated data for the Monte Carlo exercise.

The results of the log-normal simulation are shown in Table B.7 and Figure B.6. Table B.7 shows the mean and median biases of each decomposition effect for all three estimators. Figure B.6 plots kernel density estimates of the distribution of each decomposition effect for all three estimators.

Table B.7: Decomposition Simulation Results - Log Normal Distribution

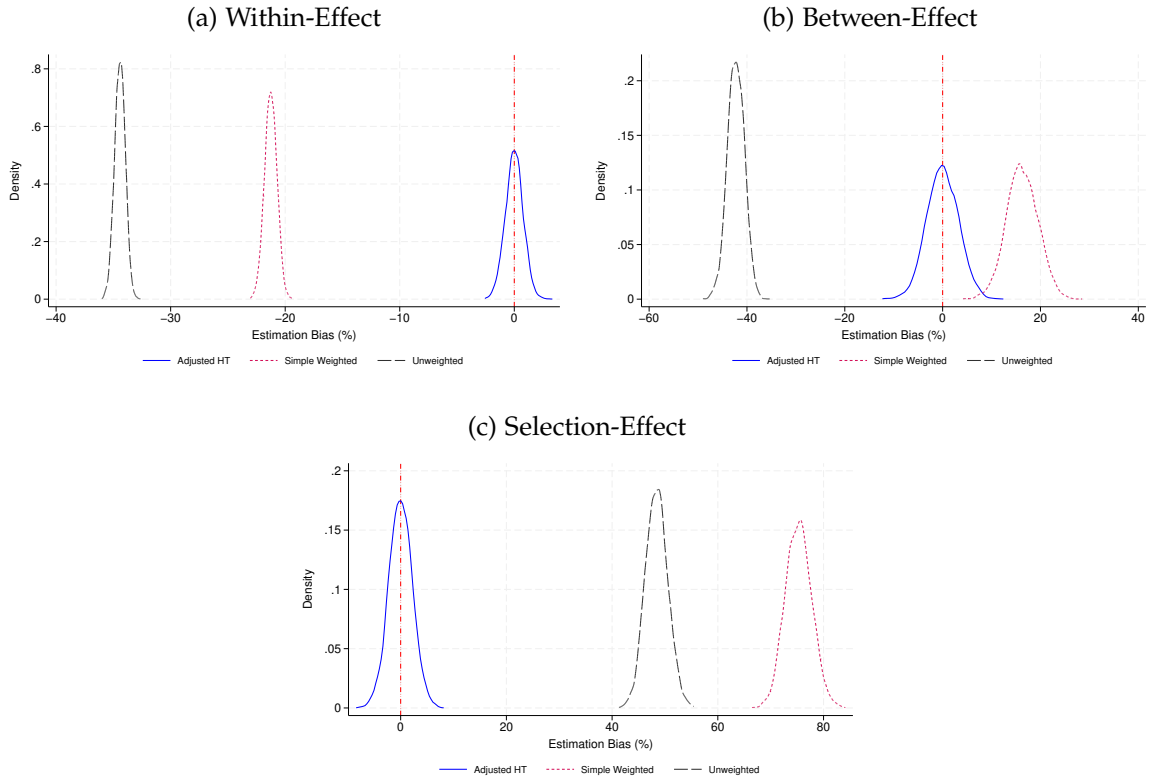
	Adjusted HT		Simple Weighted		Unweighted	
	Mean	Median	Mean	Median	Mean	Median
	(1)	(2)	(3)	(4)	(5)	(6)
Within-Effect	0.00	0.00	-21.26	-21.27	-34.43	-34.42
Between-Effect	0.00	-0.01	16.27	16.19	-42.25	-42.25
Selection-Effect	-0.01	-0.02	75.28	75.27	48.43	48.41

Notes: Table shows results of a Monte Carlo Simulation (5,000 repetitions) testing the bias in three BHC estimators: adjusted Horvitz-Thompson, simple weighted, and unweighted. Mean and median estimation biases (in %) are shown. The simulation generates population variables from a multivariate log-normal distribution with a fixed mean and correlation structure (N=50,000). In the simulation, a sample of observations are drawn from the population following a size-based sampling design with fixed sample probabilities.

B.3 Robustness of the Adjusted HT Estimator to Small Samples

Here we probe the robustness of our proposed adjusted HT estimator when the sample is a small share of the true population. Our simulations in Section 3.4 sampled 70% of the population. Here we sample just 20% of the population in each period. As in

Figure B.6: Decomposition Simulation Results - Log-Normal Distribution



Notes: Figure shows BHC Decomposition estimation bias (in %) from a Monte Carlo Simulation (5,000 repetitions) for three estimators: adjusted Horvitz-Thompson, simple weighted, and unweighted. The simulation generates population variables from a log-normal multivariate distribution with a fixed mean and correlation structure (N=50,000). In the simulation, a sample of observations are drawn from the population following a size-based sampling design with fixed sample probabilities.

Section 3.4, we perform two distinct simulations, one where the data is drawn from a Pareto distribution and one where data is drawn from a bivariate normal distribution. We impose the same size based sampling regime, where we have reduced the sample rate in each size quartile proportionately such that 20% of the full population is sampled.

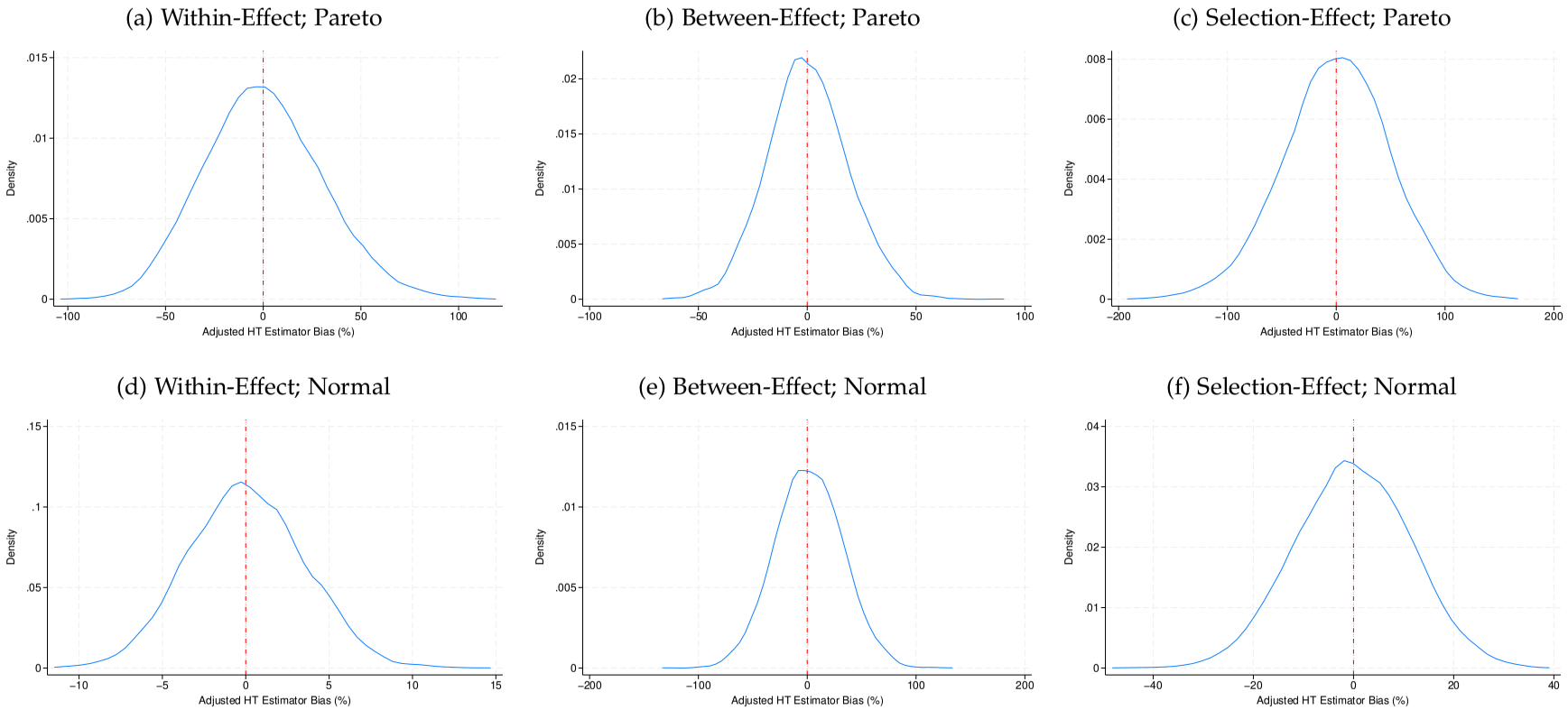
The mean and median "errors" are shown in Table B.8, with the full distributions shown in Figure B.7. As these results show, the adjusted HT estimator still performs well.

B.4 The Performance of Variance Estimators

Here we examine the performance of our proposed estimators of the within, between, and selection effects estimated using the adjusted HT approach. To begin, we derive explicit formulas for each term under a stratified random sample design. We focus on this design as it is the design underlying the data we use in our empirical application.

To start, recall Proposition 5 in the main text provides a general formula for the

Figure B.7: Adjusted HT Estimator Simulation Results



Notes: Figure shows BHC Decomposition estimation bias (in %) from a Monte Carlo Simulation (5,000 repetitions) for our proposed adjusted HT estimator. Results from two separate simulations are shown. The first simulation (panels (a) through (c)) generates population variables from a Pareto distribution with a fixed shape and scale parameter (N=50,000). The second simulation (panels (d) through (f)) generates population variables from a multivariate normal distribution with a fixed mean and correlation structure (N=50,000). In both simulations, a sample of observations are drawn from the population following a size-based sampling design with fixed sample probabilities. A total of 20% of the population is sampled each period.

Table B.8: HT Estimator Simulation Results

	Mean (1)	Median (2)
	Panel (a): Pareto Distribution	
Within-Effect	0.22	-1.00
Between-Effect	0.36	-0.20
Selection-Effect	0.06	1.44
	Panel (b): Bivariate Normal Distribution	
Within-Effect	0.04	-0.04
Between-Effect	1.59	1.23
Selection-Effect	-0.16	-0.25

Notes: Table reports results from two sets of Monte Carlo simulations (each with 5,000 repetitions) testing the bias in our proposed adjusted HT estimators of the decomposition terms of Equation (3). Mean and median estimation biases (in %) are reported for term. The first simulation (Panel (a)) generates population variables from a Pareto distribution with a fixed shape and scale parameter (N=50,000). The second simulation (Panel (b)) generates population variables from a bivariate normal distribution with a fixed mean and correlation structure (N=50,000). In both simulations, a sample of observations are drawn from the population following a size-based sampling design with fixed sample probabilities. A total of 20% of the population is sampled each period.

variances of the HT within-, between-, and selection-effect estimators. Under stratified random sampling, the variance of the within-effect is given by:

$$Var\left(\widehat{WE}_{t_1}^{HT}\right) = \sum_{h=1}^{H} N_h^{c^2}\left[1 - f_h\right] \frac{Var_h(s_{it_0}\Delta x_{it_1})}{\widetilde{N}_h^c}, \tag{B.9}$$

where h indexes strata 1 through H, N_h^c is the number of continuers in strata h in the population, $\widetilde{N}_h^c$ is the number of continuers in strata h in the sample, f_h is the probability with which a continuer in strata h is sampled, and $Var_h(s_{it_0}\Delta x_{it_1})$ is the variance of $s_{it_0}\Delta x_{it_1})$ in stratum h. Substituting estimates of N_h^c and $Var_h(s_{it_0}\Delta x_{it_1})$ into Equation (B.9) gives an estimate of $Var(\widehat{WE}_t^{HT})$.

The variance of the between-effect is given by:

$$Var\left(\widehat{BE}_{t_1}^{HT}\right) = \sum_{h=1}^{H} N_h^{c^2}\left[1 - f_h\right] \frac{Var_h(\Delta s_{it_1} x_{it_1})}{\widetilde{N}_h^c}, \tag{B.10}$$

Substituting estimates of N_h^c and $Var_h(\Delta s_{it_1} x_{it_1})$ into Equation (B.10) gives an estimate of $Var(\widehat{BE}_{t_1}^{HT})$.

The covariance between $\widehat{WE}_t^{HT}$ and $\widehat{BE}_t^{HT}$ is required to estimate the variance of the

selection-effect. That covariance is given by:

$$Cov\left(\widehat{WE}_{t_1}^{HT}, \widehat{BE}_{t_1}^{HT}\right) = \sum_{h=1}^{H} N_h^{c^2} \left[1 - f_h\right] \frac{Cov_h(s_{it_0}\Delta x_{it_1}, \Delta s_{it_1} x_{it_1})}{\widetilde{N}_h^c} \tag{B.11}$$

Substituting estimates of N_h^c and $Cov_h(s_{it_0}\Delta x_{it_1}, \Delta s_{it_1} x_{it_1})$ into Equation (B.11) gives an estimate of $Cov(\widehat{WE}_t^{HT}, \widehat{BE}_t^{HT})$.

The variance of the selection-effect is given by $Var\left(\widehat{SE}_{t_1}\right) = Var\left(\widehat{WE}_{t_1}^{HT}\right) + Var\left(\widehat{BE}_{t_1}^{HT}\right) + 2Cov\left(\widehat{WE}_{t_1}^{HT}, \widehat{BE}_{t_1}^{HT}\right)$. Replacing the within- and between-effect variances and covariance with their estimators gives an estimator for the selection-effect.[3]

If one does not observe the true population shares and instead estimates adjusted HT estimators, then the above variance estimators will be subject to bias. One can either approximate the true errors with the Taylor linearization approach of Demnati and Rao (2004) or live with the bias.[4]

We use simulations to compare the relative size of the bias from using the adjusted HT estimator and the approximation error from using the Taylor linearization approach of Demnati and Rao. We again use the simulated data generated in Section 2.1, and estimate both analytical and linearized variances in each simulated sample. Table B.9 shows the mean and median difference between the true variance and the estimated variance for each decomposition term estimated via the adjusted HT estimator. The analytical variance estimators outperform the linearized variance estimators for both the Pareto and Normal distributions. The results for the Pareto distribution, shown in Panel (a), indicate an average bias for the analytical variance estimators of between 0.45 and 2.12%, depending on decomposition term. The linearized variance estimators have approximation errors of up to 12%. The results for the Normal distribution, shown in Panel (b), show average biases of the analytical variance estimators of up to 6% and average approximation errors of up to 42% for the linearized estimators.

[3]In practice, it is possible for $\widehat{Var}(\widehat{SE}_{t_1})$ to be negative, which is clearly an invalid estimate of $Var(\widehat{SE}_{t_1})$. In these cases, one can replace $\widehat{Var}(\widehat{SE}_{t_1})$ with an upper bound given by $Var(\widehat{SE}_{t_1}) = Var(\widehat{WE}_{t_1}^{HT}) + Var(\widehat{BE}_{t_1}^{HT}) + 2\sqrt{Var(\widehat{WE}_{t_1}^{HT})}\sqrt{Var(\widehat{BE}_{t_1}^{HT})}$. This upper bound follows from the Cauchy-Schwarz inequality.

[4]Note that the formula for the variance of the adjusted Horvitz-Thompson estimator of the selection effect must account for the fact that ΔY_{t_1} is also estimated. In this case, $Var(\widehat{SE}_{t_1}) = Var(\Delta\widehat{Y}_{t_1}^{HT}) + Var(\widehat{WE}_{t_1}^{HT}) + Var(\widehat{BE}_{t_1}^{HT}) - 2Cov(\Delta\widehat{Y}_{t_1}^{HT}, \widehat{WE}_{t_1}^{HT}) - 2Cov(\Delta\widehat{Y}_{t_1}^{HT}, \widehat{BE}_{t_1}^{HT}) + 2Cov(\widehat{WE}_{t_1}^{HT}, \widehat{BE}_{t_1}^{HT})$.

Table B.9: Decomposition Simulation Results - Variances

	Analytical Variances		Linearized Variances	
	Mean (1)	Median (2)	Mean (3)	Median (4)
Panel (a): Pareto Distribution				
Within-Effect	0.45	0.73	-1.15	-0.91
Between-Effect	1.18	1.26	0.31	0.30
Selection-Effect	2.12	1.95	12.08	12.13
Panel (b): Normal Distribution				
Within-Effect	-0.23	-0.21	42.39	42.45
Between-Effect	2.11	2.18	-3.15	-3.13
Selection-Effect	5.81	5.79	3.49	3.49

Notes: Table shows results of a Monte Carlo Simulation (5,000 repetitions) testing the bias of variance estimators of the adjusted Horvitz-Thompson decomposition. The performance of two variance estimators is assessed: analytical variances and linearized variances. Mean and median estimation biases (in %) are shown. Results from two separate simulations are shown. The first simulation (panel (a)) generates population variables from a Pareto distribution with a fixed shape and scale parameter (N=50,000). The second simulation (panel (b)) generates population variables from a multivariate normal distribution with a fixed mean and correlation structure (N=50,000). In both simulations, a sample of observations are drawn from the population following a stratified random size-based sampling design with fixed sample probabilities.

C Data Appendix

C.1 India

We received the ASI panel data by special request from the government of India's Ministry of Statistics and Programme Implementation. This panel runs from the 1998/99 fiscal year to the 2017/18 fiscal year. We only use data from 1998/99 to 2015/16, as described below.

We adopt the following data cleaning and sample trimming procedures.

- Concord NIC 2008 (ISIC Rev 4) and NIC 2004 (ISIC Rev 3.1) to NIC 98 (ISIC Rev 3) industry definitions. For TFP estimation purposes, we combine the following industries to ensure at least 100 observations per industry: 222 and 223, 332 and 333, and 352 and 353.

- Calculate the user cost of capital using the perpetual inventory method, following the approach in Boehm and Oberfield (2020). Doing so imposes depreciation rates of 0%, 5%, 10%, 20%, and 40% for land, buildings, machinery, transportation equipment, and computers & software, respectively. We deflate each type of capital (land, buildings, machinery, transportation equipment, and computers) using

the relevant deflator provided by Boehm and Oberfield (2020) in their replication package. This uses data originally derived from India's Ministry of Statistics and Programme Implementation wholesale price index. The nominal interest rate we use is the India Bank Lending Rate, from the IMF's International Financial Statistics series. We end our analysis in 2015/16 to align with the availability of this data.

- Deflate sales and material costs using industry-specific deflators from the Wholesale Price Index (India's Ministry of Commerce and Industry, 2025).
- Deflate wages by the Consumer Price Index for industrial workers produced by India's Ministry of Labour and Employment.
- Drop establishment-years for which industry definitions are not available.
- Drop establishment-years that are not coded as open according to ASI status.
- Drop establishment-years with no sales, capital, or employees.
- Drop outliers in reported labor or material spending following the approach adopted by Allcott et al. (2016). Labor spending outliers are those with either reported compensation double sales values, compensation assuming 1,000 rupees per worker double sales values, or more than 200,000 workers. Material spending outliers are those with material costs double sales.

C.2 WBES

Although the WBES samples are small, they are intended to be representative of the formal, private manufacturing sector. Moreover, both sample weights and detailed information on the sample design are provided for all years and countries. As a result, data from the WBES can be used to estimate an aggregate decomposition using the adjusted Horvitz-Thompson approach. We do so for 17 countries in Sub-Saharan Africa.

For this exercise we rely on the WBES-provided estimates of TFP. In the WBES, TFP is estimated by OLS at the two-digit ISIC-code level, pooling data from all countries available (including those sampled outside of Sub-Saharan Africa). The regressions include country and year fixed effects. For industries with more than 500 units (across all countries), the production function parameters are allowed to vary by country income group.[5]

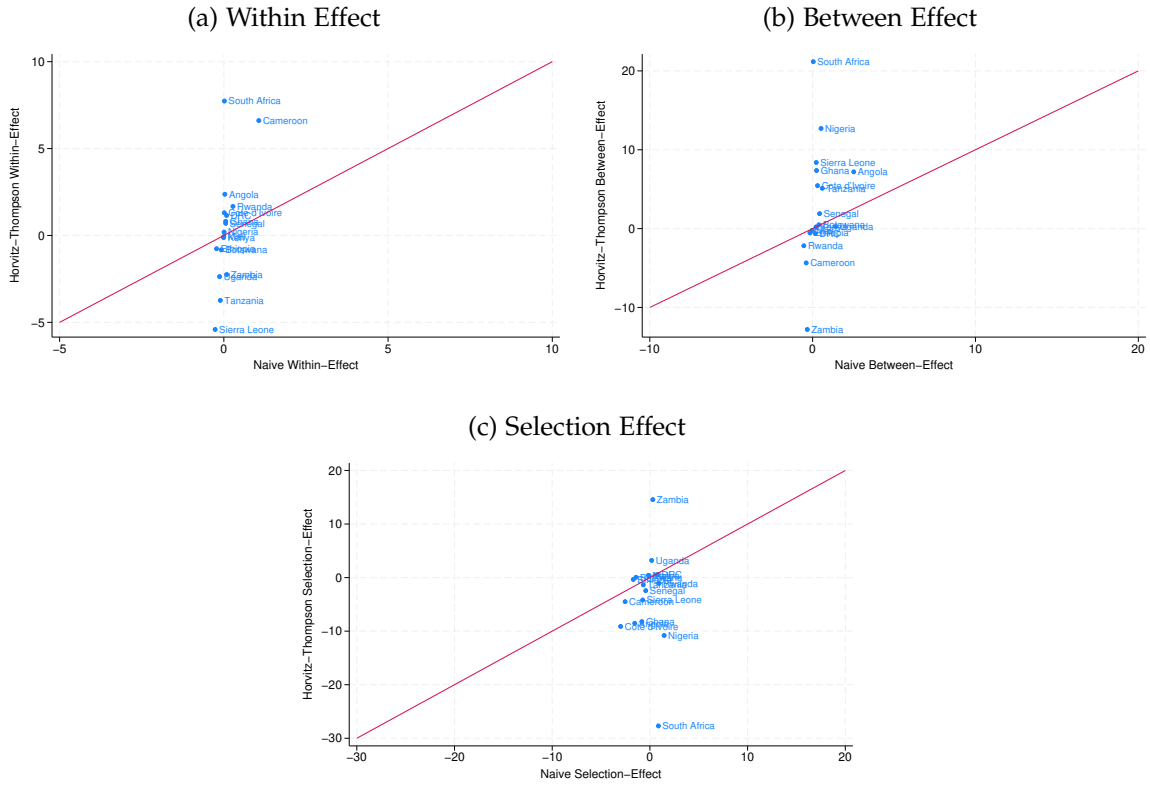
[5]The input elasticities are allowed to vary across income groups for ISIC 10, 13, 14, 16 ,18, 20, 22, 23, 25, 27, 28 and 31. The elasticities are common across countries for ISICs 11, 15, 17, 21, 24, 26, 29.

D Additional Application Results

D.1 Sub-Saharan Africa

As we noted in the main text, as a final step in our analysis of firm dynamics and aggregate productivity in selected Sub-Saharan countries in Africa, we compare the estimates of the within, between and selection effects from Table 4 with those from a "naive" approach in which Equation (2) is directly applied to the WBES data. The results from this comparison are presented in the three panels of Figure D.8. Panel (a) of the figure reports the two sets of estimates for the within effect, Panel (b) reports the two sets of estimates for the between effect, and Panel (c) reports the two sets of estimates for the selection effect.

Figure D.8: Comparison of Decomposition Estimates for Sub-Saharan Africa



Notes: Figure compares adjusted Horvitz-Thompson decomposition estimates to naive decomposition estimates for 17 countries in Sub-Saharan Africa. Panel (a) plots estimates of the within effect, Panel (b) plots estimates of the between effect, and Panel (c) plots estimates of the selection-effect. All estimates are expressed in log changes. A 45 degree line is shown for reference.

If both decomposition approaches yielded similar estimates, then all points on the figure would lie on the 45-degree line in each panel of Figure D.8. However, as the three

panels of the figure show, this is clearly not the case. Applying the naive decomposition produces very little variation in the within, between, and selection effect estimates, with most hovering quite close to zero. The estimates from our adjusted HT estimator are starkly different, highlighting how the application of the estimator can improve our understanding of the sources of productivity growth in settings where the direct application of an accounting decomposition would yield little useful insight.