INTERNATIONAL ECONOMIC REVIEW
Vol. 63, No. 1, February 2022 DOI: 10.1111/iere.12534

IMPROVING THE MEASUREMENT OF EARNINGS DYNAMICS*

BY MOIRA DALY, DMYTRO HRYSHKO, AND IOURII MANOVSKII

Copenhagen Business School, Denmark; University of Alberta, Canada; University of Pennsylvania, U.S.A.

Empirically, earnings at the start or end of earnings spells are lower and more volatile than in the interior of earnings histories, reflecting mainly the effects of working less than the full year. Ignoring these properties leads to a mismeasurement of the permanent and transitory shock variances and induces the large and widely documented divergence in the estimates of those variances based on fitting the earnings moments in levels or growth rates. Accounting for these effects enables more accurate analysis using quantitative models with permanent and transitory earnings risk and improves empirical estimates of consumption insurance against permanent earnings shocks.

1. INTRODUCTION

The central element of many models in modern quantitative macroeconomics with heterogeneous agents is either an exogenously specified or an endogenously determined stochastic process for individual earnings. For example, in the models with incomplete insurance markets, the properties of the earnings process serve as key determinants of the evolution of consumption, assets, and other economic choices over the life cycle and across individuals.[1] Following the seminal contribution by Friedman (1957), modern consumption theory recognizes that consumption should respond more to the longer lasting or permanent, as opposed to transitory, innovations in earnings. This explains the keen interest in the literature in measuring the variances of these components using variants of the permanent/transitory earnings decomposition[2] written, in its basic form, as:

$$\begin{aligned} y_{it} &= \alpha_i + p_{it} + \tau_{it}, \\ p_{it} &= \phi_p p_{it-1} + \xi_{it}, \\ \tau_{it} &= \theta(L)\epsilon_{it}, \end{aligned} \tag{1}$$

where log-earnings y_{it} of individual i at time t consist of the permanent component, p_{it}, and the transitory component, τ_{it}. If ϕ_p is close to 1, the shocks ξ_{it} are highly persistent (and are truly permanent if ϕ_p is 1), and if $\theta(L) = 1$ (where $\theta(L)$ is a moving average polynomial in the lag operator L), the shocks ϵ_{it} are completely transitory.

*Manuscript received May 2018; revised June 2021.

We benefited from comments of the editor, Jesús Fernández-Villaverde, four anonymous referees, and participants at numerous conferences and seminars. Support from the National Science Foundation Grants No. SES–0922406, 1357903, 1824520, the Danish Social Science Research Council (FSE) grant No. 300279, and SSHRC IG grant 435-2018-1275 is gratefully acknowledged. Please address correspondence to: Iourii Manovskii, Department of Economics, The Ronald O. Perelman Center for Political Science and Economics, University of Pennsylvania, 133 South 36th Street, Philadelphia, PA 19104. E-mail: *manovski@econ.upenn.edu*

[1] See, for example, Deaton (1991), Carroll (1997), and Castaneda et al. (2003).

[2] This decomposition was pioneered by Friedman and Kuznets (1954) and empirically supported by MaCurdy (1982), Abowd and Card (1989), and Meghir and Pistaferri (2004), among others. A prominent alternative, for example, Guvenen (2009), allows for less persistent shocks but individual-specific trends in earnings.

In addition to determining equilibrium consumption and wealth distributions, the variance and persistence of the shocks ξ_{it} and ϵ_{it} have important implications for policy design. For example, they are crucial for determining the optimal design of the bankruptcy code in Livshits et al. (2007), and they govern the impact of the welfare system on household savings in Hubbard et al. (1995), stimulus effects of fiscal policy in Heathcote (2005) and Hagedorn et al. (2019b), the transmission mechanism of monetary policy in Kaplan et al. (2018) and Hagedorn et al. (2019a), as well as the optimal design of the tax system in Banks and Diamond (2010) and Farhi and Werning (2012). Moreover, there is great interest in understanding whether the dramatic increase in earnings dispersion over the last few decades in the U.S. is due to the increase in the variances of persistent or transitory shocks (e.g., Gottschalk and Moffitt, 1994). A better understanding of this increase in earnings dispersion could help determine why consumption inequality did not increase nearly as much (e.g., Krueger and Perri, 2006; Blundell et al., 2008; Heathcote et al., 2010b; Attanasio et al., 2015). Knowing the stochastic nature of earnings is also essential for the design of active labor market policies. For example, Meghir and Pistaferri (2011) suggest that income maintenance policies might be an appropriate response to changes in inequality driven by transitory shocks, whereas training programs are potentially more relevant to counteracting the effects of permanent shocks.

Unfortunately, despite their manifest importance, there is no consensus in the literature on the size of the shocks ϵ_{it} and ξ_{it}. In particular, using the same data, the estimates of the earnings process in Equation (1) when targeting the moments of log-earnings in levels are dramatically different from the estimates obtained when fitting the moments of log-earnings in differences. Although this discrepancy was first documented using survey-based data, it remained undiminished when the focus of the literature has shifted to relying more on administrative data sets.[3] These data sets are typically orders of magnitude larger than survey-based ones; free of sampling issues; do not suffer from the typical issues of attrition; are based on administrative sources, such as tax records, which are considered highly reliable and free of issues of systematic nonresponse or measurement errors that typically plague survey-based data. Yet, despite their numerous attractive properties, these data sets must also have features that lead to the large discrepancy in the estimates based on moments in growth rates and levels.

Such an observation led Heathcote et al. (2010a) to conclude that the widely used model of earnings dynamics in Equation (1) is misspecified. Unfortunately, in the absence of knowledge of the nature of this potential misspecification, one cannot be confident in the conclusions of the models that incorporate this earnings process. Even if this misspecified process is used as a model input because there is no better alternative, whether it is more appropriate to parameterize it using the estimates targeting the data moments in levels or in differences is unclear. Relatedly, in the literature that endogenizes the earnings process,[4] it is unclear whether the process implied by the model should be compared to the one estimated from the data using the specification in levels or differences, given that estimating the reduced-form process (1) on the model-generated data does not give rise to the observed discrepancy.

In this article, we uncover an important source of this misspecification. Estimation of the parameters of the earnings process in the literature is based on fitting the entire set of autocovariance moments for levels or differences of log-earnings. However, even when estimation is based on the same set of observations in the data, computation of the autocovariance moments in levels and differences is effectively based on different information. To clarify with an example, consider an individual with a single earnings observation in the sample. This observation will contribute to the estimated variance of earnings in levels, but it will not contribute to any moment in differences. More generally, earnings observations adjacent to a missing one (e.g., observations at the start or at the end of an individual's earnings history) also contribute differently to the moments in levels and differences. If earnings observations surrounding the

[3] Recent contributions include Blundell et al. (2015), DeBacker et al. (2013), Domeij and Floden (2010), Guvenen et al. (2014), among others.

[4] For example, Huggett et al. (2011) and Postel-Vinay and Turon (2010).

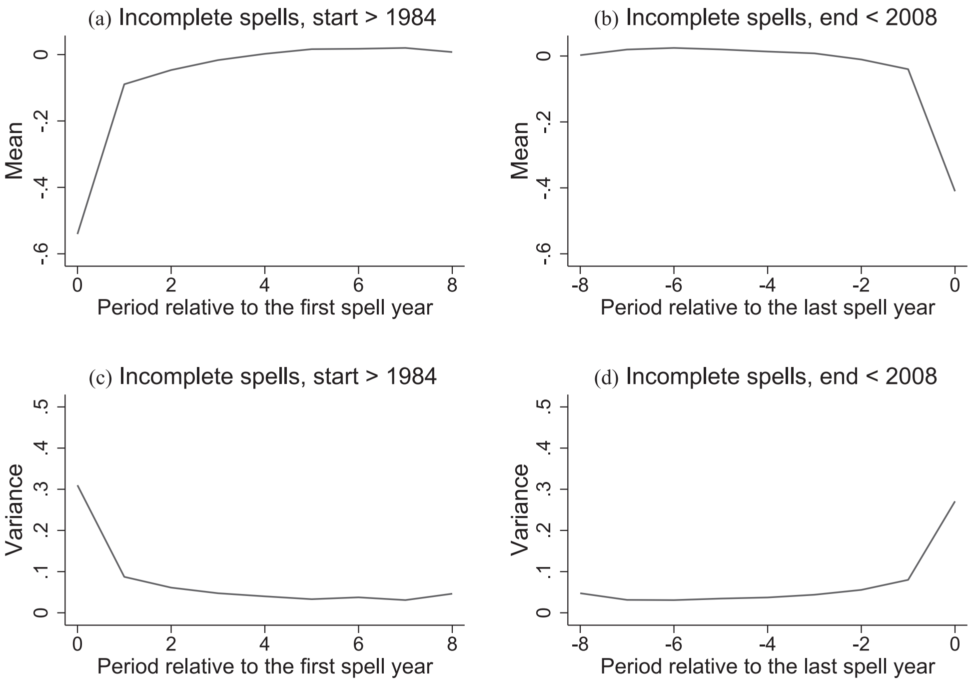

FIGURE 1

PROPERTIES OF RESIDUAL EARNINGS IN INCOMPLETE SPELLS: GERMAN DATA

missing ones were random draws similar to observations from the rest of earnings histories, this would not matter. However, in the data these earnings, observations are much lower and more volatile. We will show formally that this data feature raises the variance of transitory shocks when estimation relies on the moments in levels and raises the variance of permanent shocks recovered by estimation based on the moments in differences.

In the first set of quantitative experiments in the article, we assess the magnitude of these effects using large administrative data sets from Denmark and Germany. The Danish data contain complete earnings histories of each resident of Denmark from 1981 through 2006. The German data are a 2% random sample of social security numbers. For these individuals, we observe the complete earnings history from 1984 through 2008. These samples are sufficiently large to allow analysis at the level of particular age cohorts, making it possible to focus on a parsimonious earnings model in (1), sidestepping the issue of modeling cohort effects. Moreover, the large size of the data enables reliable estimation when replicating the design of samples typically used in the literature. Specifically, we consider a balanced sample spanning 25 (26) years in German (Danish) data, a sample with nine or more consecutive observations, as in, for example, Browning et al. (2010) and Meghir and Pistaferri (2004), and a sample with 20 or more not necessarily consecutive observations as in, for example, Guvenen (2009). Our smallest Danish sample is composed of about 67,000 individuals and 1.7 million observations, whereas our smallest German sample contains about 10,000 individuals with more than 200,000 observations.

Using the unbalanced samples in both data sets, we find, consistent with the literature, a substantially higher estimated variance of permanent (transitory) shocks targeting the moments of earnings in growth rates (levels). In contrast, we find that the discrepancy is nearly absent in balanced samples drawn from the two data sets. To highlight the special nature of the earnings trajectories in unbalanced samples, in Figure 1, we plot means and variances of (residual) earnings for the German data for individuals whose earnings are first observed in the data only after the start of the sample period in 1984, or for those whose earnings are last observed in the data before the end of the sample window in 2008. Panels (a) and (b) show that earnings are considerably lower on average, and Panels (c) and (d) show that they are substantially more volatile than typical earnings observations at the start and end of incomplete earnings spells, respectively.[5] Online Appendix Figure A-1 presents a different visualization of these patterns in unbalanced samples and, in addition, highlights their absence in the balanced sample. What then makes balanced and unbalanced samples different?

For the vast majority of individuals in the balanced sample, their first year in the sample does not coincide with the first year of their earnings history. Similarly, their last year in the sample mechanically truncates earnings histories, implying that it is not the last year of the actual earnings spell. Thus, earnings of these individuals in the first and last sample years are not expected to differ systematically from the neighboring observations in the interior of the sample window.

This stands in sharp contrast to the earnings histories of individuals entering and/or exiting the data in the interior of the sample window. Entering the sample in the interior of the sample window implies that these individuals do not have a valid earnings observation, at least in the first year of the sample window. By construction, their first available earnings observation must be preceded by a missing one. Similarly, the last earnings observation of a person exiting the sample early must be followed by a missing one. The reasons why earnings observations could be missing in administrative data will differ across data sets, but most of them do not include earnings of individuals who are self-employed or work for the government, especially in tenured positions (because in many countries, these individuals do not contribute to the regular Social Security system from which the data typically come). Individuals who die or move abroad and pay taxes there will be missing from the data. Individuals will not typically have earnings records in the years they devote to education. Long unpaid paternity or maternity leave may also result in missing earnings observations. Individuals will typically not appear in the data, whereas they are out of work and are not receiving unemployment or welfare benefits. Some missing observations are created by researchers themselves through setting sample restrictions. For example, in an attempt to guard against errors in the data, it is common to remove earnings outliers by setting them to missing. Most studies also set to missing observations when individuals do not satisfy some restrictions on, for example, the minimum amount of time worked during a year. Except for the observations set to missing by researchers when processing the data, in most data sets, it is not even possible to know why a particular observation is missing.[6]

Importantly, regardless of why earnings might be missing, earnings that follow or precede these periods ought to be relatively low in expectation and vary widely across individuals for the following reasons. The data in most administrative data sets are annual. Individuals who join the sample in year t but had missing earnings in year $t-1$ because they were, for example, unemployed, students, self-employed, or government workers are not expected to become regular employees on January 1 of year t, meaning that their annual earnings (from regular jobs) in the first year of earnings history are going to be relatively low in expectation and vary widely across individuals depending on when they actually started regular employment. Similarly, not all individuals die or move out of the country on December 31, and neither do they necessarily leave regular employment for other activities on that day. Thus, earnings in the last year are also going to be relatively low and more volatile. There are also some wage effects in the first and last years of incomplete earnings spells that we document below, but the transitory effect of incomplete working years is clearly dominant.

[5] Log residual earnings are residuals from cross-sectional regressions of log earnings on educational dummies, a third polynomial in age, and the interactions of the age polynomial with the educational dummies. We subtracted individual fixed effects from residual earnings before taking means and variances.

[6] In Online Appendix B, we exploit the richness of the Danish data to document the relative importance of various reasons for why individuals may not be present in the administrative earnings records.

Our theoretical argument implies that the nonrandomness of earnings that surround missing observations in the unbalanced samples can induce the discrepancy between the estimates in levels and differences in the data from the unbalanced samples. The quantitative question is how large this effect is. To provide an answer, we proceed in three steps. First, we quantify the contribution of the low mean and high variance of earnings surrounding missing observations in the unbalanced samples to the subset of theoretical autocovariance moments on which the identification argument in levels and differences is based and confirm that they induce the observed discrepancy in the estimates. Second, using unbalanced samples, we remove a few observations adjacent to missing records. We find that estimating the earnings process in levels and in differences on the remaining data yields virtually identical estimates of the variances of permanent and transitory shocks. Third, we simulate artificial data based on these estimates of the earnings process while replicating the structure of the unbalanced samples (by design, observations surrounding missing records are not systematically different from observations in the rest of the earnings histories). We find no discrepancy between the estimates in levels and differences when these artificial data are used. We then draw an additional shock at the start and end of the contiguous earnings spells to replicate the mean and the variance of earnings in those periods in the data. We find that in this case, the estimates of the variance of permanent and transitory shocks are very different when moments in levels and differences are used but are very close to those in the data from the corresponding unbalanced samples.

The results of these experiments lead us to conclude that the discrepancy in the estimates of the earnings process (1) in growth rates and levels is indeed driven by its misspecification. The nature of the misspecification is surprisingly simple. It is driven by the high variance and the low mean of the observations surrounding missing records. We show that an extended earnings process that includes these data elements can be estimated in the data. Estimating such an extended process results in similar parameters regardless of whether the moments in levels or differences are used.

Although the focus of this article is on the administrative data on individual earnings, the insights apply more broadly. In the Online Appendix, we show that both individual earnings and hourly wage observations surrounding the missing observations in survey data from the Panel Study of Income Dynamics (PSID) exhibit a lower mean and higher variance than typical observations, and that accounting for the properties of these observations helps reconcile the dramatically different estimates of the variances of permanent and transitory shocks when targeting the moments in levels or differences. Hryshko and Manovskii (2018) show that the same findings also hold for total household earnings and net family incomes in the PSID. They also show that not taking these features of the data into account when estimating the stochastic properties of family earnings and income induces a large bias on the estimated degree of consumption insurance achieved through household saving and borrowing as well as on the estimated insurance role of the tax and transfer system.

Although the mechanism described in this article is powerful in reconciling the estimates of the earnings process in growth rates and levels, it is not the only mechanism that can generate such a discrepancy. For example, Hryshko and Manovskii (2017) show that the discrepancy can also be induced if one restricts the permanent component to a random walk when its true persistence is lower. Importantly, this type of misspecification cannot generate the difference between the theoretical moments that we use to establish identification in levels and differences in this article because they are identically affected by any such misspecification. These theoretical identifying moments can only differ if the underlying autocovariance moments on which they are based disagree, and we show that this is indeed the consequence of the low mean and high variance of observations at the start and end of earnings spells. We find that this accounts for virtually all of the discrepancy between the estimates in growth rates and levels in the data considered in this article.

The rest of the article is organized as follows: In Section 2, we discuss identification of the permanent-transitory decomposition of earnings and derive theoretically the biases in the estimated variances of permanent and transitory shocks when using the moments in levels and differences constructed from an unbalanced panel. In Section 3, we describe the administrative Danish and German data, the estimation procedure and basic results, and document the low mean and high variance of earnings observations surrounding the missing ones. In Section 4, we show that this property of earnings quantitatively accounts for the difference in the estimates of earnings processes in levels and differences in our administrative data. In the Online Appendix, we confirm that our findings based on administrative data also apply to survey data on earnings and wages from the PSID. Section 5 concludes.

2. SOURCES OF THE DIFFERENCES IN ESTIMATES TARGETING EARNINGS GROWTH RATES AND LEVELS

In this section, we first outline identification of the canonical earnings process that consists of a permanent random walk component and a nonpersistent transitory shock. We then derive the biases in the estimated variances of permanent and transitory shocks in unbalanced panels that feature lower on average and more volatile earnings next to the missing records. We close the section by discussing the biases when permanent shocks have finite persistence or transitory component is modeled as a moving-average process.

2.1. *Identifying Moments for the Canonical Earnings Process.* In the literature, estimation of the parameters of the earnings process typically relies on the minimum-distance method. In particular, estimation based on the moments in levels targets the entire set of autocovariance moments $E[y_{it}y_{it+j}]$, where $i \in [1, N]$ denotes individuals in the sample, t denotes time, and j denotes all leads and lags of earnings observed in the data. In differences, estimation targets the full set of autocovariance moments $E[\Delta y_{it}\Delta y_{it+j}]$, where Δ is the difference operator between two consecutive observations, $\Delta y_{it+j} \equiv y_{it+j} - y_{it+j-1}$.

Although all available autocovariance moments are used in estimation, identification is usually established using only a subset of autocovariance moments; see, for example, Meghir and Pistaferri (2004), Blundell et al. (2008), and Heathcote et al. (2014). For example, consider the earnings process that consists of a random walk and an i.i.d. transitory shock, which corresponds to setting $\theta(L)$ and ϕ_p to 1 in Equation (1). This process was considered by Heathcote et al. (2010a), who, assuming that transitory and permanent shocks are not correlated with each other and with initial conditions, proposed the following moments to identify the variances of permanent and transitory shocks at time t:

`Differences:`

(D1) $$\sigma^2_{\xi,t} = E[\Delta y_{it}\Delta y_{it-1}] + E[\Delta y_{it}\Delta y_{it}] + E[\Delta y_{it}\Delta y_{it+1}],$$

(D2) $$\sigma^2_{\epsilon,t} = -E[\Delta y_{it}\Delta y_{it+1}].$$

Note that (D1) and (D2) represent linear combinations of autocovariance moments for earnings growth rates. For clarity, we will refer to individual autocovariance moments as simply "moments," and to a linear combination of autocovariance moments used for identification such as (D1) and (D2) as "identifying moments."

Expanding (D1) and (D2), we obtain the identifying moments for the variances of permanent and transitory shocks, based on autocovariance moments in levels, at time t:

`Levels:`

(L1) $$\sigma^2_{\xi,t} = E[y_{it}y_{it+1}] - E[y_{it+1}y_{it-1}] - E[y_{it}y_{it-2}] + E[y_{it-1}y_{it-2}],$$

(L2) $$\sigma^2_{\epsilon,t} = E[y_{it}y_{it}] - E[y_{it}y_{it+1}] - E[y_{it-1}y_{it}] + E[y_{it-1}y_{it+1}].$$

In a sample of individuals whose earnings are nonmissing for the periods $t-2$ through $t+1$, the identifying moments (D1)–(D2) and (L1)–(L2) are expected to deliver identical estimates of the variance of permanent and transitory shocks at time t, since they are based on exactly the same earnings information. Moreover, as the moments (L1)–(L2) simply represent an expansion of the moments (D1)–(D2), they will be identically affected by other potential misspecifications of the earnings process. This allows us to isolate and measure the importance of the high variance and low mean of the observations at the start and end of contiguous earnings spells, which, as we show below, contribute differently to the autocovariance moments on which (D1)–(D2) and (L1)–(L2) are based.[7]

For example, the presence of omitted idiosyncratic trends in earnings will not induce a wedge between the estimated variances of permanent shocks using the moments (L1) and (D1) (or transitory shocks using the moments (L2) and (D2)). Specifically, suppose that individuals differ in growth rates such that the earnings process is $y_{it} = \alpha_i + \beta_i h_{it} + p_{it} + \epsilon_{it}$, where $\beta_i \sim \text{iid}(0, \sigma^2_\beta)$ and h_{it} counts years of (potential) work experience. In this case, (L1) and (D1) will both deliver $3\sigma^2_\beta + \sigma^2_{\xi_t}$.[8] It follows that both (L1) and (D1) will recover an upward-biased estimate of the variance of the permanent shock, but the bias will be the same in levels and differences. Relatedly, the typical estimates of σ^2_ξ using (D1) imply a much steeper profile of earnings inequality over the life cycle (and time) than that observed in the data. The fit to this profile might be improved if one allows for a negative cross-sectional correlation between initial conditions, α_i, and permanent shocks, ξ_{it}. However, omitting such correlation does not induce a difference in the estimated moments (L1) and (D1). For example, suppose that the correlation is implied by $\xi_{it} = \kappa\alpha_i + \eta_{it}$, where η_{it} is orthogonal to α_i and ϵ_{it}. In this case, (D1) and (L1) will recover identical upward-biased estimate $3\kappa^2\sigma^2_\alpha + \sigma^2_{\xi_t}$, but the bias will once again be the same in levels and differences.

Importantly, each autocovariance moment is measured as the average across all available observations that contribute to it. This implies that, although the identifying moments (D1)–(D2) and (L1)–(L2) are based on the same earnings data, the autocovariance moments used in estimating (D1)–(D2) and (L1)–(L2) are computed using different sets of observations. Returning to the extreme example used in the Introduction, consider an individual who appears in the sample only once, in period t. This individual will contribute to the autocovariance moment $E[y_{it}y_{it}]$, and thus, his only earnings observation will affect the identifying moment (L2), but it will not contribute to any autocovariance moment used to construct the corresponding identifying moment in differences (D2). If earnings of individuals who appear in the sample only once are systematically different, this will induce the difference between identifying moments (L2) and (D2) and lead to different estimates of the variance of transitory shocks using the moments in levels and differences.

[7] Identifying moments in levels can be constructed using fewer autocovariance moments, such as

(L1-Short) $$\sigma^2_{\xi,t} = E[y_{it}y_{it+1}] - E[y_{it}y_{it-1}],$$

(L2-Short) $$\sigma^2_{\epsilon,t} = E[y_{it}y_{it}] - E[y_{it}y_{it+1}].$$

These moments are analogous to those in Heathcote et al. (2010a) if one relies on the annual data, instead of biennial data used in their paper, for identification of the variances. These identifying moments in levels do not, however, use the same information as the identifying moments (D1)–(D2) in differences. For example, the information on earnings in $t-2$ is used in (D1) but not in (L1-Short). Moreover, Hryshko and Manovskii (2017) show that a misspecification of the persistence of the permanent component drives a wedge between the estimates based on identifying moments (D1)–(D2) and (L1-Short)–(L2-Short), but not between identifying moments (D1)–(D2) and (L1)–(L2).

[8] This derivation assumes that $\text{corr}(\alpha_i, \beta_i) = 0$ but a similar expression obtains if this assumption is relaxed.

Similarly, we will now show that earnings observations at the time individuals enter or exit the sample contribute differently to the autocovariance moments on which the identifying moments (D1)–(D2) and (L1)–(L2) are based. Moreover, our empirical analysis below will reveal that these earnings observations are systematically different (they are typically lower and substantially more volatile). In the rest of this section, we formally show that this induces systematic differences in estimated variances of permanent and transitory shocks using the moments in growth rates and levels. In subsequent sections, we quantify the magnitude of the induced difference.

2.2. *Potential Bias Induced by Observations Next to the Missing Ones in Various Samples.* We will consider three types of samples. Suppose a data set containing panel data on individual earnings starts in period t_0 and ends in period T. We refer to the sample as balanced if all individuals in the sample have $T - t_0 + 1$ valid earnings observations. While not part of the formal definition, it is convenient to think that earnings spells of individuals in the balanced samples start before t_0 and end after T. In other words, the boundaries of the balanced sample mechanically truncate continuous earnings spells in progress. We refer to samples that include only uninterrupted earnings spells (i.e., no gaps) but with a duration of less than $T - t_0 + 1$ for at least some individuals as consecutive unbalanced samples. Finally, we refer to unbalanced samples that also include individual earnings spells interrupted by missing observations in any period $t \in (t_0, T)$ as nonconsecutive unbalanced samples.

2.2.1. *Biases in consecutive unbalanced samples.* The nature of consecutive unbalanced samples is that at least some individuals are observed starting or ending their earnings spells inside the sample window.

As mentioned above and documented below, earnings have a lower mean and are highly volatile in the first and last periods of an incomplete earnings history. Consider modeling this through an additional ν-shock that occurs only in the first and last year of an individual's earnings history, that is,

$$y_{it} = \alpha_i + p_{it} + \epsilon_{it} + \nu_{it},$$

where ν_{it} has mean μ_ν (taking a negative value) and variance σ_ν^2, and is uncorrelated with permanent and transitory shocks and initial conditions. We will now show that ignoring ν_{it} and estimating the process (1) instead leads to an upward bias in the estimated variance of permanent shocks using the moments in differences and the estimated variance of transitory shocks using the moments in levels.

For simplicity, assume that there is a set of individuals first entering the sample at time t, in the interior of the sample period $[t_0, T]$, whereas the remaining individuals are continuously observed throughout the sample. Individuals first appearing at time t will contribute to estimation of the autocovariance moments $E[y_{it}y_{it}]$ and $E[y_{it}y_{it+1}]$ in the identifying moment (L2). The estimated moment $E[y_{it}y_{it+1}]$ will be no different for such individuals than for the rest of the sample, and will equal $\sigma_\alpha^2 + \text{var}(p_{it})$. The other moments in (L2), $E[y_{it-1}y_{it}]$ and $E[y_{it-1}y_{it+1}]$, will both equal $\sigma_\alpha^2 + \text{var}(p_{it-1})$. The autocovariance moment $E[y_{it}y_{it}]$ estimated on the full sample, however, will equal $\sigma_\alpha^2 + \text{var}(p_{it}) + \sigma_{\epsilon,t}^2 + s_t(\mu_\nu^2 + \sigma_\nu^2)$, where s_t is the share of individuals, at time t, whose (incomplete) spells start at time t in the total number of individuals at time t with nonmissing earnings. The identifying moment (L2) therefore will recover an estimate of the variance of transitory shocks equal to $\sigma_{\epsilon,t}^2 + s_t(\mu_\nu^2 + \sigma_\nu^2)$, with an upward bias of $s_t(\mu_\nu^2 + \sigma_\nu^2)$.

The variance of permanent shocks at time $t+1$, estimated using the identifying moment (D1), will also be biased upward. Individuals first appearing at t will contribute to estimation of the autocovariance moments $E[\Delta y_{it+1}\Delta y_{it+1}]$ and $E[\Delta y_{it+1}\Delta y_{it+2}]$ in the identifying moment (D1). For such individuals, the autocovariance moment $E[\Delta y_{it+1}\Delta y_{it+2}]$ will be no different from the rest of the sample and will equal $-\sigma_{\epsilon_{t+1}}^2$, whereas the autocovariance moment

$E[\Delta y_{it+1}\Delta y_{it+1}]$ will equal $\sigma^2_{\xi_{t+1}} + s_{t,t+1}(\mu^2_\nu + \sigma^2_\nu) + \sigma^2_{\epsilon_t} + \sigma^2_{\epsilon_{t+1}}$, where $s_{t,t+1}$ is the share of individuals who start (incomplete) earnings spells at time t, with nonmissing earnings at times t and $t+1$, in the number of individuals with nonmissing earnings both at t and $t+1$. Since the autocovariance moment $E[\Delta y_{it+1}\Delta y_{it}]$ will be estimated using information for those individuals whose earnings are nonmissing in periods $t-1$ through $t+1$ and will equal $-\sigma^2_{\epsilon_t}$, the identifying moment (D1) for time $t+1$ will recover an estimate of the permanent shock equal to $\sigma^2_{\xi_{t+1}} + s_{t,t+1}(\mu^2_\nu + \sigma^2_\nu)$, with an upward bias of $s_{t,t+1}(\mu^2_\nu + \sigma^2_\nu)$.

Note that if the ν-shock first appears, say, at time $t+1$, that is, in the interior of an earnings spell for individuals first entering into the sample at time t, it will simply elevate, by the same magnitude, the estimated variance of transitory shocks in levels and differences at time $t+1$, with no differential effect on the identifying moments (L2) and (D1).

Summing up, incomplete earnings spells first appearing in the sample at t will bias upward the estimated variance of transitory shocks at time t when targeting the moments in levels and will bias upward the estimated variance of permanent shocks at time $t+1$ when targeting the moments in differences. They have no effect, at any point in time, on the estimated magnitude of the identifying moments (L1) and (D2).

The same logic extends to the incomplete earnings spells ending at time t, which is different from the last potential sample year T—the presence of such spells will produce upward-biased estimates of permanent variances in differences at t (since these individuals will contribute to the estimation of the moment $E[\Delta y_{it}\Delta y_{it}]$ that is part of the identifying moment D1) and of transitory variances in levels at t.

2.2.2. *Biases in nonconsecutive unbalanced samples.* We now consider the consequences of missing earnings in the interior points of the earnings history. We assume that individual earnings are realizations of the earnings process (1), with some observations missing in any period $t \in (t_0, T)$ but not in periods $t-1$ and $t+1$. We will show below that such periods are often associated in the data with low mean and high variance of earnings in periods $t-1$ and $t+1$. We model this by introducing additional transitory ν-shocks with a negative mean μ_ν at the time before and after earnings are missing (ν_{it-1} and ν_{it+1}, respectively) that are assumed to be uncorrelated with fixed effects, permanent and transitory shocks, and uncorrelated with each other:[9]

$$\begin{gathered} y_{it-1} = \alpha_i + p_{it-1} + \epsilon_{it-1} + \nu_{it-1}, \\ y_{it} \ \text{ missing}, \\ y_{it+1} = \alpha_i + p_{it+1} + \epsilon_{it+1} + \nu_{it+1}. \end{gathered}$$

Assume that there is a set of individuals whose earnings are missing in period $t \in (t_0, T)$, whereas the rest of the individuals have continuously observed earnings throughout the whole sample period.

In this case, the variance of transitory shocks at times $t-1$ and $t+1$ using the moments in levels will be biased upward as the autocovariance moments $E[y_{it-1}y_{it-1}]$ and $E[y_{it+1}y_{it+1}]$ in the identifying moment (L2) are amplified by the variation of the ν-shocks. Similarly, the variance of permanent shocks at times $t-1$ and $t+2$ using the moments in differences will be biased upward as the autocovariance moments $E[\Delta y_{it-1}\Delta y_{it-1}]$ and $E[\Delta y_{it+2}\Delta y_{it+2}]$ in the identifying moment (D1) are amplified by the variation of the ν-shocks. Since the ν-shocks are assumed to be uncorrelated, the identifying moment (D2) will not be affected. Since ν-shocks

[9] For ease of exposition, we assume that the mean and variance of the ν-shock one year before and after earnings are missing are the same, although in the data they slightly differ.

have negative means, there will be a downward bias in the estimated variance of permanent shocks using (L1) due to an amplified moment $E[y_{it-1}y_{it+1}]$.[10]

Thus, incomplete earnings spells with missing earnings at t, in the interior of the sample period, will bias upward the estimated variance of transitory shocks at times $t-1$ and $t+1$ when targeting the moments in levels, and will bias upward the variance of permanent shocks at times $t-1$ and $t+2$ when targeting the moments in differences.

2.3. *Extensions.*

2.3.1. *Biases when ξ_{it} shocks have limited persistence.* If ϕ_p in Equation (1) is less than 1, one must rely on a modified set of identifying moments to recover the permanent and transitory variances. For a given estimate of the persistence ϕ_p, which can be separately identified,[11] the set of identifying moments will amount to

`Differences:`

$$\sigma^2_{\xi,t} = E[\tilde{\Delta} y_{it}\tilde{\Delta} y_{it+1}] + \phi_p E[\tilde{\Delta} y_{it}\tilde{\Delta} y_{it}] + \phi_p^2 E[\tilde{\Delta} y_{it}\tilde{\Delta} y_{it-1}], \qquad \text{(D1 – a)}$$

$$\sigma^2_{\epsilon,t} = -\frac{1}{\phi_p} E[\tilde{\Delta} y_{it}\tilde{\Delta} y_{it+1}] \qquad \text{(D2 – a)},$$

where $\tilde{\Delta} y_{it} \equiv y_{it} - \phi_p y_{it-1}$.

Expanding the above moments results in the following set of moments in levels identifying the variances at time t:

`Levels:`

$$\sigma^2_{\xi,t} = E[y_{it}y_{it+1}] - \phi_p E[y_{it+1}y_{it-1}] - \phi_p^3 E[y_{it}y_{it-2}] + \phi_p^4 E[y_{it-1}y_{it-2}], \qquad \text{(L1 – a)}$$

$$\sigma^2_{\epsilon,t} = E[y_{it}y_{it}] - \frac{1}{\phi_p} E[y_{it}y_{it+1}] - \phi_p E[y_{it-1}y_{it}] + E[y_{it-1}y_{it+1}]. \qquad \text{(L2 – a)}$$

Although the biases for the variance of transitory shocks in levels will be exactly the same as in the random-walk case, the biases for the variance of permanent shocks recovered using the identifying moments in differences will be scaled by the persistence ϕ_p. Note, however, that the size of the bias is unlikely to decline substantially since ϕ_p is typically estimated at high values in various data sets.

2.3.2. *Biases when transitory component and/or ν-shocks are serially correlated.* The transitory component is often estimated to have some persistence. Assume that the transitory component is modeled as $\tau_{it+1} = \epsilon_{it+1} + \theta_\tau \epsilon_{it}$, and that the ν-shock component is modeled as $\chi_{it} = \nu_{it}$, which is nonzero in the beginning and/or end of an incomplete earnings spell, and before/after a missing earnings record, and that $\chi_{it+1} = \theta_\chi \nu_{it}$ – both will be consistent with the autocovariance function for earnings growth rates truncating at the second order, as is often found in the empirical applications.[12] In this case, the moments (L1)–(D2) no longer identify the variances of permanent and transitory shocks. In growth rates, the identifying moment for

[10] A further downward bias in the variance of permanent shocks at time t using (L1) will occur due to missing observations at $t-1$ surrounded by nonmissing records at $t-2$ and t. These downward biases will be absent if one relies on (L1-short) instead of (L1) for estimating the variance of transitory shocks using the moments in levels.

[11] The persistence ϕ_p can be recovered from the moments $\frac{E[y_{it+k+3}y_{it+k}]-E[y_{it+k+2}y_{it+k}]}{E[y_{it+k+2}y_{it+k}]-E[y_{it+k+1}y_{it+k}]}$ for $k \geq 0$. One can also use the moments in growth rates to identify it; see, for example, Hryshko (2012). There is also a large literature, reviewed in MaCurdy (2007) and Arellano and Honoré (2001), that does not rely on fitting the autocovariance function of earnings but exploits various orthogonality conditions in a generalized method of moments setting to recover the persistence.

[12] This formulation assumes that ν- and ϵ-shocks both die out in two periods, with the difference that the ν-shock process does not renew itself in the next period with a new ν-shock.

the variance of permanent shocks should be modified to

$$\sigma^2_{\xi,t} = E[\Delta y_{it}\Delta y_{it+2}] + E[\Delta y_{it}\Delta y_{it+1}] + E[\Delta y_{it}\Delta y_{it}] + E[\Delta y_{it}\Delta y_{it-1}] + E[\Delta y_{it}\Delta y_{it-2}]. \qquad (\text{D1}-\text{b})$$

The variance of permanent shocks at time $t+1$, estimated using (D1-b), will be biased upward by the magnitude $s_{t,t+1}(1-\theta_\chi)^2(\mu^2_\nu + \sigma^2_\nu)$ for a sample with consecutive earnings spells where a fraction of individuals enter the sample at time $t > t_0$, for the first time. Note that the bias will remain large for small positive values of θ_χ. If, instead, individuals exit the sample at some time $t < T$, the bias of the permanent variance using the moments in growth rates will be unaffected by serial correlation of the ν-shocks since the earnings paths for such individuals are unobserved past year t; the bias, in this case, will be the same as in the case of a serially uncorrelated transitory component. The same logic extends to the biases in the nonconsecutive samples. The variance of permanent shocks recovered using the moments in levels will remain unbiased (as can be verified from the identifying moment for permanent shocks in levels obtained by expanding (D1-b)).

Under the assumption of no measurement error in administrative earnings, θ_τ can be identified from the first- and second-order autocovariances in earnings growth rates if the transitory component is serially correlated and there are no ν-shocks; see, for example, Meghir and Pistaferri (2004). One can then identify the variance of transitory shocks dividing (L2) and (D2) by $(1-\theta_\tau)^2$. If the ν-shock is serially correlated, however, θ_τ will be recovered with a bias using the standard moment. We will label this estimate as $\tilde{\theta}_\tau$. Assuming that the variance of transitory shocks does not change much between adjacent periods, for the data with incomplete consecutive spells that start at t, an estimate of the variance of transitory shocks relying on (L2) will yield $(1-\tilde{\theta}_\tau)^{-2}[(1-\theta_\tau)^2\sigma^2_{\epsilon_t} + s(1-\theta_\chi)(\mu^2_\nu+\sigma^2_\nu)]$, whereas an estimate relying on (D2) will yield $(1-\tilde{\theta}_\tau)^{-2}[(1-\theta_\tau)^2\sigma^2_{\epsilon_t} - s\theta_\chi(1-\theta_\chi)(\mu^2_\nu+\sigma^2_\nu)]$ for $t+1$.[13] Clearly, an estimate of the variance of transitory shocks in levels is larger than an estimate using growth rates given θ_χ is nonnegative. This logic extends to other examples of incomplete earnings spells in consecutive and nonconsecutive panels—the estimated variance of transitory shocks using the moments in levels will be higher than the estimated variance of transitory shocks using the moments in growth rates.

2.4. *Summary.* The analysis above yields three major implications if ν-shocks are present in the data. First, estimating the abbreviated earnings process in (1), one may expect to recover fairly well the variance of transitory shocks using the moments in growth rates if the ν-shock is not serially correlated, and the variance of permanent shocks using the moments in levels. Second, the identifying moments in levels tend to produce upward-biased estimates of the variance of transitory shocks, while the identifying moments in differences produce upward-biased estimates of the variance of permanent shocks. The magnitude of the biases depends positively on the variance of the ν-shocks and on the difference between their mean from the mean of the shocks in the rest of earnings histories. Finally, if one's interest extends beyond identifying the variances of permanent and transitory shocks of the abbreviated earnings process in (1), the remaining parameters of the comprehensive earnings process can also be estimated by introducing the moments identifying the properties of ν-shocks.

3. DATA, ESTIMATION DETAILS, AND BASIC RESULTS

In this section, we first describe administrative Danish and German data used for estimation of earnings processes. We then present estimation of the canonical earnings process in Equation (1) on balanced and unbalanced samples from the two data sets. Since the results on balanced samples substantively deviate from the results on unbalanced samples, we further present descriptive regression results for the level and volatility of residual earnings

[13] We assumed that $s_t = s_{t,t+1} = s_{t,t+2} = s$ in the derivation.

confirming the special nature of earnings records surrounding the missing ones in the unbalanced samples.

3.1. *Data.* In this section, we describe the administrative data and construction of the samples that we study. Following the literature, we focus on individuals with a strong attachment to the labor market characterized by sufficiently high earnings and time spent working.[14]

3.1.1. *Danish data.* Several administrative registers provided by Statistics Denmark were used to construct our samples. The Danish Integrated Database for Labor Market Research (IDA), containing earnings from the tax register from 1980–2006, provides panel data on total earnings for more than 99.9% of Danish residents between the ages of 15 and 70. Population and education registers are also merged so that demographic variables such as age and educational status could be appended. Our sample window spans the years from 1981 to 2006. The population we study consists of Danish males born in 1951–55. We first remove all individuals who were ever self-employed and drop records in which an individual was making nonpositive labor market earnings. Next, we drop records for those individuals who have worked less than 10% of the year as a full-time employee.

Annual earnings in a particular year include all earned labor income, taken from tax records, for that calendar year. This variable is considered "high quality" by Statistics Denmark in that it very accurately captures the earnings of individuals. Earnings are expressed in 1981 monetary units (Danish kroner). We calculate the maximum number of consecutive periods in which an individual has nonmissing earnings and use this information to construct two consecutive samples: a sample in which an individual's maximum spell is at least nine consecutive periods (102,825 individuals), and a balanced sample in which the individual's maximum spell covers the entire 26 periods (67,008 individuals). For the sample with nine or more consecutive observations, periods outside of the longest spell are discarded. Within the longest spell, an earnings outlier is defined by an increase in earnings of more than 500% or a fall of more than 80% in adjacent years. Individuals with earnings outliers within their longest spell are dropped. The third sample we consider consists of individuals who have at least 20 not necessarily consecutive periods in which they have nonmissing earnings (90,668 individuals). We also drop individuals from this sample if they have earnings growth outliers. Finally, we drop individuals if their educational status has changed during the spells considered. Table 1 contains basic statistics for selected samples.

3.1.2. *German data.* We use administrative panel data from the Sample of Integrated Labour Market Biographies (SIAB) that follows individuals selected as a 2% random sample of German Social Security records for the years 1974–2008. A detailed description of the data set can be found in Dustmann et al. (2009). We use full-time job spells for German males born in 1951–55, dropping the spells in East Germany. We also drop annual records when an individual was in apprenticeship during any part of the year. Individual real earnings is

[14] The selection rules we adopt are typical of the literature that utilizes survey data as well as administrative data. For example, Guvenen et al. (2014) use U.S. administrative data on individual wage and salary income and make the following sample selection: "For a statistic computed using data for not necessarily consecutive years $t_1, t_2, \dots, t_n$, an individual observation is included if the following three conditions are satisfied for all these years: the individual (i) is between the ages of 25 and 60, (ii) has annual wage/salary earnings that exceed a time-varying minimum threshold, and (iii) is not self-employed (i.e., has self-employment earnings less than the same minimum threshold). This minimum, denoted $Y_{min,t}$, is equal to one-half of the legal minimum wage times 520 hours… This condition allows us to focus on workers with a reasonably strong labor market attachment and avoids issues with taking the logarithm of small numbers. It also makes our results more comparable to the income dynamics literature, where this condition is standard." Similarly, DeBacker et al. (2013) "…exclude earnings (or income) observations below a minimum threshold…" and "…take the relevant threshold to be one-fourth of a full-year, full-time minimum wage." In line with our selection of consecutive unbalanced samples (with the difference that we use at least nine consecutive earnings observations), Blundell et al. (2015) "…restrict the sample to individuals with at least four subsequent observations with positive market income."

TABLE 1
DANISH DATA, 1981–2006: SUMMARY STATISTICS FOR SELECTED YEARS

	9 Consec.	20 Not Nec. Consec.	Balanced
Number of individuals	102825	90668	67008
Number of observations	2367552	2298429	1742208
Education			
Less than high school	0.227	0.222	0.206
High school degree	0.032	0.031	0.029
Vocational training	0.505	0.521	0.542
Two-year university degree	0.046	0.046	0.047
Bachelor's degree	0.125	0.122	0.124
Master or Ph.D.	0.065	0.059	0.051
Earnings			
1985	40157	40227	41383
	(12831)	(12889)	(12278)
1995	48197	48444	50004
	(20562)	(20462)	(19954)
2005	52656	51511	53298
	(26635)	(26279)	(25917)
Spell counts			
Start 1981, end 2006	67008	80787	67008
Start after 1981, end 2006	13439	4376	0
Start in 1981, end before 2006	17723	5210	0
Start after 1981, end before 2006	4655	295	0
Total	102825	90668	67008
Number of spells with 20 or more not nec. consec. observations, by length			
[Fraction of missing to nonmissing observations within spell in square brackets]			
20		1634 [0.144]	
21		2009 [0.119]	
22		2665 [0.096]	
23		3296 [0.079]	
24		4486 [0.054]	
25		9570 [0.030]	
26		67008 [0.00]	

Notes: Earnings are expressed in 2005 Euros; the standard deviation of earnings is given in parentheses. See Section 3.1 for the details on the construction of the samples. "9 consec." sample contains individual earnings spells with nine or more consecutive earnings observations. "20 not nec. consec." sample contains individual earnings spells with at least twenty not necessarily consecutive earnings observations. "Balanced" sample contains individual earnings spells that cover the entire period.

a sum of earnings from the jobs held within a year expressed in 2005 euros. We set individual education to the maximum schooling attained during the sample years and set the number of days worked to the sum of calendar days on all jobs within a year. As individual earnings are right-censored at the highest level subject to social security contributions, we impute earnings exceeding the limit, assuming that daily wages in the upper tail follow a Pareto distribution, the parameters of which differ by year and age group.[15] After 1983, earnings include one-time payments such as bonuses. To make variable definitions consistent throughout, we use only the data since 1984. We also drop individual records on annual earnings if the combined duration of job spells within a year is fewer than 35 calendar days and drop records with

[15] We consider the following eight age groups: those younger than 25, six five-year age groups (25–29, 30–34, 35–39, 40–44, 45–49, and 50–54), and those older than 54. We use a "fixed effects" imputation, keeping a uniform draw for each individual affected by the right-censoring limit fixed when creating a Pareto variate in different years. We also experimented with imputation based on the assumption that truncated log-wage distribution is normal, and a simpler imputation when the daily wage is multiplied by the factor 1.2 if it hits the upper censoring limit. These three imputation methods have been used in Dustmann et al. (2009). Our conclusions below are robust with respect to the choice of the imputation method as well as with respect to limiting the sample to individuals whose earnings histories are not affected by the censoring.

TABLE 2
GERMAN DATA, 1984–2008: SUMMARY STATISTICS FOR SELECTED YEARS

	9 Consec.	20 Not Nec. Consec.	Balanced
Number of individuals	18130	13635	9452
Number of observations	379080	330748	236300
Education			
Middle school or no degree	0.05	0.04	0.04
Vocational training	0.72	0.74	0.76
High school degree	0.06	0.05	0.05
College	0.17	0.17	0.15
Earnings			
1985	33626	33930	34559
	(15876)	(13323)	(12881)
1995	45309	47180	47965
	(24702)	(24295)	(24463)
2005	49121	51289	52457
	(36473)	(37106)	(37666)
Spell counts			
Start 1984, end 2008	9452	11179	9452
Start after 1984, end 2008	3136	1007	0
Start in 1984, end before 2008	4463	1393	0
Start after 1984, end before 2008	1079	56	0
Total	18130	13635	9452
Number of spells with 20 or more not nec. consec. observations, by length			
[Fraction of missing to nonmissing observations within spell in square brackets]			
20		575 [0.054]	
21		509 [0.054]	
22		623 [0.05]	
23		871 [0.037]	
24		1605 [0.027]	
25		9452 [0.00]	

Notes: Earnings are expressed in 2005 Euros; the standard deviation of earnings is given in parentheses. See Section 3.1 for the details on the construction of the samples. "9 consec." sample contains individual earnings spells with nine or more consecutive earnings observations. "20 not nec. consec." sample contains individual earnings spells with at least twenty not necessarily consecutive earnings observations. "Balanced" sample contains individual earnings spells that cover the entire period.

very low daily earnings, defined as the earnings records below 14 euros in 2003 prices. As in the Danish data, we construct three samples—balanced, with nine or more consecutive, and with 20 or more not necessarily consecutive earnings observations—and, as with the Danish samples, drop individuals who have earnings growth outliers. The respective samples contain 9,452, 18,130, and 13,635 individuals with 236,300, 379,080, and 330,748 observations, respectively. Table 2 provides some descriptive details of the samples.

3.2. *Estimation Details.* As is standard in the literature, we estimate the earnings process in Equation (1) using the method of minimum distance, fitting the entire set of autocovariances of log-earnings in levels or first differences to the autocovariance function implied by the model.[16] We allow for an MA(1) transitory component and an unrestricted estimation of the persistence of the permanent component, ϕ_p.[17] Thus, we estimate five parameters in total—the persistence and the variance of permanent shocks, ϕ_p and σ_ξ^2; the persistence and the variance of transitory shocks, θ and σ_ϵ^2; and the variance of individual fixed effects, σ_α^2. We assume that individuals start accumulating permanent and transitory shocks at the age of 25 so that part of the estimated variance of fixed effects captures the accumulated perma-

[16] One of the recent exceptions is Browning et al. (2010) who, apart from selected moments in levels and differences, fit a variety of other data moments studied in the literature on earnings dynamics.

[17] Allowing, instead, for an AR(1) transitory component has no substantive influence on the results.

TABLE 3
ESTIMATES OF THE EARNINGS PROCESS IN ADMINISTRATIVE DATA

	9 Consec.				20 Not Nec. Consec.				Balanced sample			
	German data		Danish data		German data		Danish data		German data		Danish data	
	Levs. (1)	Diffs. (2)	Levs. (3)	Diffs. (4)	Levs. (5)	Diffs. (6)	Levs. (7)	Diffs. (8)	Levs. (9)	Diffs. (10)	Levs. (11)	Diffs. (12)
$\hat{\phi}_p$	0.976	0.992	0.955	0.987	0.993	0.991	0.964	0.982	0.998	0.994	0.969	0.970
	(0.001)	(0.0008)	(0.0008)	(0.0004)	(0.001)	(0.001)	(0.0007)	(0.0006)	(0.001)	(0.002)	(0.0007)	(0.0009)
$\hat{\sigma}_\xi^2$	0.008	0.019	0.008	0.013	0.0048	0.009	0.007	0.012	0.0034	0.0036	0.005	0.005
	(0.0002)	(0.0003)	(0.0001)	(0.0001)	(0.0001)	(0.0002)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.00004)	(0.00005)
$\hat{\theta}$	0.129	0.153	0.204	0.209	0.119	0.192	0.137	0.217	0.193	0.185	0.212	0.209
	(0.005)	(0.009)	(0.002)	(0.003)	(0.008)	(0.008)	(0.003)	(0.003)	(0.008)	(0.009)	(0.003)	(0.003)
$\hat{\sigma}_\epsilon^2$	0.024	0.009	0.019	0.012	0.016	0.009	0.022	0.013	0.0072	0.007	0.009	0.009
	(0.0003)	(0.0002)	(0.0001)	(0.0001)	(0.0003)	(0.0003)	(0.0002)	(0.0001)	(0.0002)	(0.0002)	(0.0001)	(0.0001)
$\hat{\sigma}_\alpha^2$	0.024	—	0.020	—	0.027	—	0.023	—	0.024	—	0.017	—
	(0.002)	—	(0.0004)	—	(0.002)	—	(0.0004)	—	(0.001)	—	(0.0004)	—
χ^2	1024.43	711.47	7104.42	4527.83	1562.51	1393.16	7002.05	5836.32	1297.41	993.44	6950.05	5855.67
(d.f.)	320	296	346	321	320	296	346	321	320	296	346	321

Notes: See Section 3.1 for the details on the construction of the samples. "9 consec." sample contains individual earnings spells with nine or more consecutive earnings observations. "20 not nec. consec." sample contains individual earnings spells with at least twenty not necessarily consecutive earnings observations. "Balanced" sample contains individual earnings spells that cover the entire period. The estimated earnings process is: $y_{it} = \alpha_i + p_{it} + \tau_{it}$, where $p_{it+1} = \phi_p p_{it} + \xi_{it+1}$ and $\tau_{it+1} = \epsilon_{it+1} + \theta\epsilon_{it}$. Models are estimated using the optimally weighted minimum distance method. Asymptotic standard errors are in parentheses. German data span the period 1984–2008, while Danish data span the period 1981–2006.

nent and transitory components prior to that age. We also assume that permanent and transitory shocks are not correlated with each other and with fixed effects. We remove predictable variation in earnings by estimating cross-sectional regressions of log earnings on educational dummies, a third polynomial in age, and the interactions of the age polynomial with the educational dummies. Our measure of idiosyncratic earnings, consistent with the literature, is the residual from those regressions. Since our samples are large, we estimate the model using the optimal weighting matrix, which is an inverse of the variance–covariance matrix of the data moments.

3.3. *Basic Results: Estimates of the Canonical Earnings Process.*

3.3.1. *Samples with nine or more consecutive observations.* Columns (1)–(4) in Table 3 contain estimation results for the samples with nine or more consecutive observations in the German and Danish data.[18] The permanent component is estimated to be close to a random walk using the moments in differences, but slightly less persistent using the moments in levels. Importantly, in both data sets, the variance of the permanent shock is about two times larger in the estimation that uses the moments in growth rates, whereas the estimated variance of the transitory shock is larger using the moments in levels. Thus, our administrative data exhibit the same large discrepancy that is endemic in this literature. The pattern is less pronounced in the Danish data, which is consistent with the mechanism we describe. In the Danish data, 65% of individuals have complete earnings spells, whereas in the German data, this number is only 52%. Consequently, fewer individuals have irregular earnings observations adjacent to the missing ones in the Danish data.[19]

[18] In differences, the variance of fixed effects is not identified. It can be identified using quasi-differences when $\phi_p < 1$ as in, for example, Blundell et al. (2015). We do not pursue this approach as our focus is on the standard estimation in differences (and levels).

[19] Randomly dropping individuals with incomplete earnings histories in the German data to match their share in the Danish data results in similar discrepancies across the two data sets.

3.3.2. *Samples with 20 or more not necessarily consecutive observations.* Columns (5)–(8) in Table 3 contain the results for the samples with 20 or more not necessarily consecutive observations. The variances of persistent shocks are somewhat smaller than those in columns (1)–(4), whereas the variances of transitory shocks are similar in magnitude. Importantly, we still observe that estimations using the moments in differences deliver relatively higher estimates of the variance of permanent shocks, whereas estimations in levels deliver relatively higher estimates of the variance of transitory shocks, once again confirming the widely documented discrepancy.[20]

3.3.3. *Balanced samples.* Estimation results based on the balanced samples are reported in columns (9)–(12) of Table 3. Relative to the estimates on the unbalanced samples discussed above, the use of balanced samples results in a more than 50% reduction of the variance of permanent shocks when using the moments in differences. There is a similarly striking reduction of at least 50% in the variance of transitory shocks when using the moments in levels. It appears that the use of balanced samples largely eliminates the discrepancy between the estimates of the earnings process in levels and differences.

3.4. *A Closer Look at Unbalanced Samples.* In this section, we first show, in the regression setting, that earnings observations around missing ones are lower and more volatile. We then discuss the economic forces behind these data features.

3.4.1. *The means and variances of earnings records surrounding missing observations.* The results of estimation on balanced and unbalanced samples indicate that the discrepancy between the estimates based on the moments in levels and differences is specific to unbalanced samples. A defining feature of unbalanced samples is that some observations are missing. As discussed in Section 2, if earnings observations surrounding the missing ones are systematically different, this can induce the difference in the estimates of the earnings process in growth rates or levels. This data feature has the potential to explain why the discrepancy arises when the estimation is based on unbalanced samples only.

Indeed, Figure 1 in the Introduction revealed a clear pattern that the earnings at the start and/or the end of incomplete earnings spells are considerably lower on average and substantially more volatile than typical earnings observations. To explore these patterns more formally, in columns (1)–(4) of Table 4, we report the estimates from the fixed-effects panel regressions of residual earnings on dummies for the first and last years of individual earnings spells inside the overall sample window. Specifically, the dummies “Year observed: first”–“Year observed: third” equal 1 if an individual’s first earnings record in the sample occurs later than 1984 in the German data (later than 1981 in the Danish data), and 0 otherwise, whereas the dummies “Year observed: second-to-last”–“Year observed: last” equal 1 if an individual’s last earnings record is prior to 2008 in the German data (2006 in the Danish data), and 0 otherwise.[21]

In both samples and both data sets, earnings are about 0.50–0.60 log points lower than an individual’s average in the first year of the spell, whereas the last earnings record is below an individual’s average by about 0.30–0.40 log points. Earnings are still somewhat lower in the two years following the first earnings record as well as in the two years preceding the last

[20] As was the case with nine or more consecutive observations, the discrepancy is less pronounced in the Danish data because the share of individuals with complete earnings spells is larger, and the share of missing earnings observations is smaller than in the German data.

[21] To reinforce the conclusion that patterns in Table 4 are actually driven by starting and ending of the earnings spells, in Table A-3 of Online Appendix C, we repeat the same analysis by focusing on individuals whose earnings spells begin in the first sample year or end in the last sample year. For the vast majority of these individuals, such cutoffs do not represent an actual start or end of their earnings spells; instead, the sample window mechanically truncates earnings spells in progress. Accordingly, the first (last) few dummies equal 1 if an individual’s fist (last) earnings record is in the first (last) sample year, and 0 otherwise.

TABLE 4
RESIDUAL EARNINGS AND SQUARED RESIDUAL EARNINGS SURROUNDING MISSING OBSERVATIONS

	Means				Variances			
	9 Consec.		20 Not Nec. Consec.		9 Consec.		20 not nec. consec.	
	German Data (1)	Danish Data (2)	German Data (3)	Danish Data (4)	German Data (5)	Danish Data (6)	German Data (7)	Danish Data (8)
Year observed: first	–0.57***	–0.48***	–0.65***	–0.47***	0.23***	0.19***	0.29***	0.20***
	(–64.24)	(–119.58)	(–34.49)	(–58.12)	(41.80)	(70.08)	(21.12)	(35.41)
Year observed: second	–0.11***	–0.17***	–0.14***	–0.19***	0.04***	0.08***	0.09***	0.12***
	(–22.57)	(–54.50)	(–11.38)	(–27.24)	(11.54)	(34.16)	(7.91)	(21.04)
Year observed: third	–0.07***	–0.09***	–0.09***	–0.10***	0.02***	0.04***	0.05***	0.05***
	(–16.71)	(–36.26)	(–8.82)	(–16.87)	(7.93)	(21.78)	(5.80)	(12.19)
Year observed: second-to-last	–0.03***	–0.05***	–0.05***	–0.08***	0.01***	0.02***	0.02***	0.03***
	(–8.34)	(–22.93)	(–6.71)	(–15.56)	(5.62)	(11.23)	(4.31)	(7.78)
Year observed: next-to-last	–0.06***	–0.08***	–0.10***	–0.11***	0.03***	0.04***	0.06***	0.06***
	(–13.99)	(–33.78)	(–9.97)	(–20.98)	(10.47)	(20.42)	(7.25)	(12.41)
Year observed: last	–0.43***	–0.28***	–0.47***	–0.32***	0.20***	0.15***	0.25***	0.18***
	(–59.93)	(–85.30)	(–30.92)	(–43.58)	(42.37)	(58.04)	(21.38)	(27.45)
3 years before earn. miss.			–0.03***	–0.03***			0.02***	0.02***
			(–4.59)	(–10.06)			(2.97)	(9.53)
2 years before earn. miss.			–0.05***	–0.04***			0.04***	0.04***
			(–7.76)	(–15.33)			(5.08)	(13.13)
1 year before earn. miss.			–0.27***	–0.26***			0.15***	0.12***
			(–27.16)	(–78.38)			(15.31)	(38.05)
1 year after earn. miss.			–0.39***	–0.43***			0.23***	0.19***
			(–34.56)	(–112.15)			(20.82)	(50.92)
2 years after earn. miss.			–0.15***	–0.14***			0.05***	0.07***
			(–19.29)	(–46.64)			(6.99)	(22.16)
3 years after earn. miss.			–0.13***	–0.11***			0.03***	0.04***
			(–16.97)	(–36.71)			(4.94)	(15.55)
Adj. R^2	0.126	0.057	0.102	0.079	0.053	0.025	0.050	0.033
No. obs.	379080	2367552	330748	2298429	379080	2367552	330748	2298429
No. indiv.	18130	102825	13635	90668	18130	102825	13635	90668

Notes: See Section 3.1 for the details on the construction of the samples. "9 consec." sample contains individual earnings spells with nine or more consecutive earnings observations. "20 not nec. consec." sample contains individual earnings spells with at least twenty not necessarily consecutive earnings observations. Log residual earnings are residuals from cross-sectional regressions of log earnings on educational dummies, a third polynomial in age, and the interactions of the age polynomial with the educational dummies. German data span the period 1984–2008, while Danish data span the period 1981–2006. The dummies "Year observed: first"–"Year observed: third" are equal to one if an individual's first earnings record is later than in 1984 in German data and 1981 in Danish data, zero otherwise; "Year observed: second-to-last"–"Year observed: last" are equal to one if an individual's last earnings record is earlier than in 2008 in German data and 2006 in Danish data, zero otherwise. Standard errors are clustered by individual; t-statistics are in parentheses. ***significant at the 1% level, **significant at the 5% level, *significant at the 10% level.

earnings record. Moreover, earnings are, on average, also lower in the years preceding and following a missing earnings record in the nonconsecutive samples. Interestingly, the dummies for the few first and last earnings records within a spell explain 5–13% of the within-individual variation in residual earnings. This number is quite high taking into account that a variety of observable factors in a typical Mincer-style regression explain less than 30% of within-individual variation in earnings.

Performing the same experiment in reverse, we use our samples with 20 or more not necessarily consecutive observations to assess the predictive power of earnings dynamics for the incidence of missing earnings. Specifically, in Online Appendix Table A-5, the dependent variable is a dummy that equals 100 if individual earnings are missing and 0 otherwise. We find that the predictive power of observables—earnings growth rates before and after missing earnings records, together with education dummies and age—on the incidence of missing earnings is quite low, in line with Fitzgerald et al. (1998) who made a similar observation using PSID data. The strong (weak) earnings growth after (before) missing records lacks high explanatory power for a missing record because there are also many declines and subsequent recoveries of earnings inside uninterrupted earnings spells. Nonetheless, missing observations are associated with positive earnings growth in the periods following a missing record and with negative earnings growth in the periods preceding a missing earnings record, implying that these individual realizations of residual earnings are not random draws from the earnings distribution. As pointed out by Moffitt and Gottschalk (2012), little is known about the effect of attrition on the autocovariance function of earnings, and therefore, on the estimates of the earnings process. Our results indicate that the effect can be large.

In columns (5)–(8) of Table 4, we proceed to explore the *volatility* of idiosyncratic earnings at the start and end of earnings spells. The size of squared residual earnings is mechanically higher in the few first and last earnings records since, as we have just seen, residual earnings are more negative, on average, in those periods. To remove the influence of this mechanical effect, we take the (individually demeaned) residuals from the regressions of columns (1)–(4) and then square them. In the German data, the overall mean of squared residual earnings is about 0.15 in both unbalanced samples, whereas in the Danish data, the corresponding mean in both samples is 0.11. The results imply that earnings are significantly more volatile in the first and last years of individual spells. For example, in the German data, the mean of squared residual earnings in the first year is about 153% ($100 \cdot 0.23/0.15$) larger than the typical size measured by the mean of squared residual earnings in the sample. In the German consecutive sample, about 23% of individuals have their first earnings record after 1984, the first calendar year of the sample, and about 31% of individuals have their last record before 2008, the last year of the sample. The same numbers for Danish data are 18% and 22%, respectively. This is a nontrivial number of individuals with pronounced differences in the level and volatility of residual earnings in the few first and last periods of earnings spells. In the nonconsecutive samples, earnings in the periods preceding and following interior missing earnings records are also highly volatile. In the German data, for instance, the volatility of earnings observations one year before an interior missing record is about 100% ($100 \cdot 0.15/0.15$) larger than the volatility of typical earnings observations. However, the interior missing observations are much less prevalent than missing observations at the start and end of earnings spells. Tables 1 and 2 indicate that less than 1.5% of observations are missing on the interior of the nonconsecutive samples in our Danish and German data.[22,23]

[22] This fraction can be calculated using the numbers reported in the tables as $\frac{\sum_{j=20}^{K} N_j \cdot p_j \cdot j}{\sum_{j=20}^{K} N_j \cdot p_j \cdot j + \sum_{j=20}^{K+1} N_j \cdot j}$, where N_j is the number of spells with j nonmissing observations, $j = 20, \ldots, K+1$, and p_j is the fraction of interior missing to nonmissing observations per typical spell with j nonconsecutive observations. K=25 for the Danish and 24 for the German data.

[23] Browning and Ejrnæs (2013) argue that the first-stage regression (removing predictable variation in earnings due to observables such as time effects, age, education, race, etc.) should be ideally integrated with the second-stage analysis of modeling the earnings dynamics. As a special case, they show that removing time effects in a first-stage

3.4.2. *Forces behind low and volatile earnings surrounding missing records.* Finally, we consider some of the forces leading to low and volatile earnings at the start and end of the earnings spells. One obvious explanation is based on the fact that the data on earnings are typically recorded at an annual frequency. An individual who is, say, entering the sample for the first time is (statistically) expected to enter in the middle of the year but may enter at any point throughout the year. Thus, earnings in that year are expected to be lower and have a larger variance than interior earnings observations from contiguous earnings histories. We can assess this conjecture using our German data, which contain information on the number of days worked on all jobs and the average daily wage from all jobs held during a year. We use these data to decompose earnings cuts in the years around missing earnings records due to a reduction in days worked and wages. As can be seen from Table 5, most of the reduction in earnings in the first or last year of the earnings spell is due to the reduction in days worked.[24] The reduction in wages in those years is nontrivial as well but is relatively less important in inducing earnings fluctuations.[25,26]

In Online Appendix Table A-6, we report that years at the start and at the end of earnings spells are associated with a somewhat elevated probability of occupation (three digits), industry (two digits), and employer change. Individuals continue experiencing elevated mobility rates in the second and third years of incomplete earnings spells, as well as in a few years prior to the end of incomplete earnings spells. Since many individuals take some time off in-between jobs, the elevated mobility rates help explain why the effect on work hours can last beyond the first and last year of the spell, although the effect in those years is much smaller.[27]

4. QUANTITATIVE EVALUATION OF THE IMPORTANCE OF OBSERVATIONS AROUND MISSING ONES FOR THE BIASES

In this section, we first directly verify the contribution of the low mean and high variance of earnings surrounding missing observations in the unbalanced samples to the variances of permanent and transitory shocks estimated using the theoretical moments (L1)–(L2) and

regression could potentially distort the estimated distribution of an autoregressive parameter in a model allowing for parameters to vary across individuals. Although we follow the common practice of separating the first-stage regression from the second stage and focus on the canonical earnings process whose parameters are shared by all individuals, we have obtained very similar results when performing the analysis of Table 4 on raw earnings using the regressions that do not control for education and predictable time variation; see Online Appendix Table A-4.

[24] The extent of the reduction in days worked is directly affected by the sample inclusion restrictions. For example, our sample includes only observations when individuals worked at least 10% of the year. In Online Appendix D, we present the results of Tables 3–4 for alternative cutoffs, ranging from 0% (when we keep all nonzero earnings) to 100% (when we restrict the sample to those working the entire year). As expected, in the sample that includes all nonzero earnings, observations surrounding the missing records have even lower mean and higher variance because they now include observations with the particularly low number of days worked. In contrast, in the sample restricted to observations where individuals work the entire year, the earnings records surrounding missing observations are not very different in mean and variance to the records in the interior of the spells (because the variation in the number of days worked is eliminated). As a consequence, the difference in the estimated variances of permanent and transitory shocks using the moments in levels and differences shrinks as the restriction on the minimum number of days worked becomes tighter. In the extreme, when no variation in the number of days worked is allowed, the results on unbalanced samples become similar to those obtained for the balanced samples.

[25] Consistent with the logic of our argument, we find a much smaller discrepancy in the estimated variances of permanent and transitory shocks when using the moments in levels and differences in *wages*, as much of the variability in earnings in our administrative data sets at the start and end of contiguous spells is due to the variability in hours. This is consistent with the observation of Krueger et al. (2010) that the discrepancy is larger for the estimates of earnings processes than wage processes in a broad cross-section of countries. One may also expect that the discrepancy will be less of an issue if the earnings process is estimated using data at a higher frequency than annual although the bias will remain as long as the earnings in the first or the last periods reflect the effects of working only a part of those periods.

[26] Somewhat relatedly, Hoffmann and Malacrino (2019) study the business-cycle variation in earnings and find that the tails of the distribution of the annual earnings growth are driven by the variation in weeks worked.

[27] To define mobility, we consider how many individuals report more than one occupation, industry, or employer in a given year of the spell. The mobility measures are relatively low in the first and last years of incomplete spells because they are typically based only on the part of the year when individuals are present in our sample.

TABLE 5
EARNINGS, WAGE, AND DAYS WORKED RESIDUALS: GERMAN DATA

	9 Consec.			20 Not Nec. Consec.		
	Earnings (1)	Days (2)	Daily Wages (3)	Earnings (4)	Days (5)	Daily Wages (6)
Year observed: first	–0.57***	–0.43***	–0.14***	–0.67***	–0.49***	–0.17***
	(–64.24)	(–57.32)	(–32.99)	(–35.59)	(–31.91)	(–16.48)
Year observed: second	–0.11***	–0.03***	–0.09***	–0.16***	–0.03***	–0.11***
	(–22.57)	(–8.79)	(–23.34)	(–12.50)	(–4.37)	(–11.56)
Year observed: third	–0.07***	–0.02***	–0.05***	–0.11***	–0.02***	–0.08***
	(–16.71)	(–6.76)	(–16.48)	(–10.04)	(–3.08)	(–8.86)
Year observed: second-to-last	–0.03***	–0.02***	–0.01***	–0.05***	–0.02***	–0.03***
	(–8.34)	(–6.66)	(–5.48)	(–6.59)	(–3.19)	(–6.52)
Year observed: next-to-last	–0.06***	–0.03***	–0.03***	–0.10***	–0.04***	–0.06***
	(–13.99)	(–10.61)	(–9.93)	(–10.28)	(–5.66)	(–9.60)
Year observed: last	–0.43***	–0.38***	–0.05***	–0.48***	–0.38***	–0.09***
	(–59.93)	(–59.94)	(–15.03)	(–31.25)	(–30.14)	(–11.49)
3 years before earn. miss.				–0.03***	–0.03***	–0.00
				(–4.19)	(–6.64)	(–0.22)
2 years before earn. miss.				–0.05***	–0.04***	–0.01***
				(–7.74)	(–8.09)	(–2.91)
1 year before earn. miss.				–0.27***	–0.23***	–0.04***
				(–27.10)	(–27.08)	(–8.23)
1 year after earn. miss.				–0.39***	–0.27***	–0.12***
				(–34.42)	(–29.88)	(–23.85)
2 years after earn. miss.				–0.15***	–0.05***	–0.10***
				(–19.07)	(–9.39)	(–20.62)
3 years after earn. miss.				–0.12***	–0.03***	–0.09***
				(–16.42)	(–7.26)	(–17.54)
Adj. R^2	0.126	0.193	0.013	0.103	0.150	0.021
No. obs.	379080	379080	379080	330748	330748	330748
No. indiv.	18130	18130	18130	13635	13635	13635

Notes: See Section 3.1 for the details on the construction of the samples. "9 consec." sample contains individual earnings spells with nine or more consecutive earnings observations. "20 not nec. consec." sample contains individual earnings spells with at least twenty not necessarily consecutive earnings observations. Log residual earnings (days) [daily wages] are residuals from cross-sectional regressions of log earnings (log days worked) [log daily wages] on educational dummies, a third polynomial in age, and the interactions of the age polynomial with the educational dummies. The dummies "Year observed: first"–"Year observed: third" are equal to one if an individual's first earnings record is later than in 1984, zero otherwise; "Year observed: second-to-last"–"Year observed: last" are equal to one if an individual's last earnings record is earlier than in 2008, zero otherwise. Standard errors are clustered by individual; t-statistics are in parentheses. ***significant at the 1% level, **significant at the 5% level, *significant at the 10% level.

(D1)–(D2). We also highlight the importance of those observations for the estimated earnings process by simply dropping them prior to estimation—an experiment that recovers the same variances of transitory and permanent shocks in levels and differences. We then proceed by augmenting the model in Equaution (1) with extra transitory shocks whose means and variances are estimated by matching the means and variances of earnings records surrounding the missing ones in the data. Since we rely on the full set of the data autocovariance moments in our estimation, as is standard in the literature, our minimum-distance estimator is overidentified and does not offer transparent mapping between the estimated parameters and data moments. We therefore further show in simulations replicating our unbalanced samples that our results based on the overidentified minimum-distance estimation hold for various plausible parameterizations of the extended earnings process. We close this section by offering some thoughts on the implications of our results for quantitative modeling that relies on the external estimation of the income process parameters.

4.1. *Direct Evaluation of the Biases Using the Permanent-Transitory Decomposition Moments.* In this section, we directly verify that irregular observations surrounding the missing ones induce most of the difference between permanent and transitory shock variances implied by identifying moments (L1)–(L2) and (D1)–(D2). This result provides evidence that the identifying moments on which our theoretical argument is based are indeed the relevant ones and largely responsible for the results of our full estimation targeting all available autocovariance moments.

As an example of computing these implied variances, we calculate an estimate of the permanent variance at time t using the identifying moment in levels (L1) as

$$\sigma^2_{\xi,l,t} = \frac{\sum_i y_{i,t}y_{i,t+1}}{\sum_i I^i_{t,t+1}} + \frac{\sum_i y_{i,t-2}y_{i,t-1}}{\sum_i I^i_{t-2,t-1}} - \frac{\sum_i y_{i,t+1}y_{i,t-1}}{\sum_i I^i_{t-1,t+1}} - \frac{\sum_i y_{i,t}y_{i,t-2}}{\sum_i I^i_{t-2,t}}, \tag{2}$$

where the subscript l indicates that we are estimating the variance using information on log-earnings in levels, and $I^i_{t,t'}$ is an indicator function taking the value of 1 if individual earnings observations are nonmissing in both years t and t', and taking the value of 0 otherwise. Note that individual i will not contribute to the estimated variance of the permanent shock at time t only if all of the earnings cross-products for that individual—$y_{it}y_{it+1}$, $y_{it-2}y_{it-1}$, $y_{it+1}y_{it-1}$, and $y_{it}y_{it-2}$—are missing.

Let I^m_{it} be an indicator function that equals 1 if an individual's earnings is missing at one of the periods $t-2,\ldots,t+2$ defined as $1(\sum_{j=-2}^{2}(1-I_{it+j})>0)$, where $1(\cdot)$ equals 1 if the expression in brackets is true and 0 otherwise. We calculate the variance of permanent shocks due to irregular observations surrounding the missing earnings records, $\sigma^2_{\xi,l,o,t}$, as

$$\sigma^2_{\xi,l,o,t} = \frac{\sum_i y_{i,t}y_{i,t+1}I^m_{it}}{\sum_i I^i_{t,t+1}I^m_{it}} + \frac{\sum_i y_{i,t-2}y_{i,t-1}I^m_{it}}{\sum_i I^i_{t-2,t-1}I^m_{it}} - \frac{\sum_i y_{i,t+1}y_{i,t-1}I^m_{it}}{\sum_i I^i_{t-1,t+1}I^m_{it}} - \frac{\sum_i y_{i,t}y_{i,t-2}I^m_{it}}{\sum_i I^i_{t-2,t}I^m_{it}}. \tag{3}$$

An estimate of the permanent variance in levels, net of the effects of irregular observations surrounding the missing ones, $\sigma^2_{\xi,l,n,t}$, can then be calculated as

$$\begin{aligned}\sigma^2_{\xi,l,n,t} = {} & \frac{\sum_i y_{i,t}y_{i,t+1}(1-I^m_{it})}{\sum_i I^i_{t,t+1}(1-I^m_{it})} + \frac{\sum_i y_{i,t-2}y_{i,t-1}(1-I^m_{it})}{\sum_i I^i_{t-2,t-1}(1-I^m_{it})} \\ & - \frac{\sum_i y_{i,t+1}y_{i,t-1}(1-I^m_{it})}{\sum_i I^i_{t-1,t+1}(1-I^m_{it})} - \frac{\sum_i y_{i,t}y_{i,t-2}(1-I^m_{it})}{\sum_i I^i_{t-2,t}(1-I^m_{it})}.\end{aligned} \tag{4}$$

We can similarly define the variances of permanent and transitory shocks in levels and differences for the consecutive unbalanced panels—for example, the permanent variance utilizing all sample information ($\sigma^2_{\xi,l,t}$ for levels and $\sigma^2_{\xi,d,t}$ for differences), the permanent variance due to irregular observations in the first and last few periods of an individual's earnings spell ($\sigma^2_{\xi,l,o,t}$ and $\sigma^2_{\xi,d,o,t}$), and the permanent variance net of their effects ($\sigma^2_{\xi,l,n,t}$ and $\sigma^2_{\xi,d,n,t}$).

We present the estimates of those variances, averaged across all sample years, for both data sets in Table 6. For the German data, in the consecutive sample, the estimates of the variance of permanent shocks in levels and differences using all sample information are 0.013 and 0.024, respectively.[28] Net of the effect of observations surrounding the missing ones, the estimated variances are $\hat{\sigma}^2_{\xi,l,n}=0.010$ in levels and $\hat{\sigma}^2_{\xi,d,n}=0.010$ in differences. The unadjusted variances of transitory shocks in levels and differences are estimated at 0.020 and 0.008, respectively, whereas the variances net of outliers in levels and differences are both estimated at 0.007. The results for the Danish data are qualitatively similar. Clearly, the discrepancy between the estimates of permanent and transitory shock variances in levels and differences is virtually eliminated when netting out the effects of irregular observations surrounding the missing ones.

[28] The estimates deviate from the values in Table 3 because we do not impose the exact permanent-transitory decomposition on the data in the minimum-distance estimation of Table 3. The difference between the estimated variance of permanent shocks in levels and differences is not as drastic as in Table 3 because the estimated persistence of the permanent shocks in levels is lower than in differences in the minimum-distance estimation.

TABLE 6
VARIANCES OF PERMANENT AND TRANSITORY SHOCKS IN THE PERMANENT-TRANSITORY DECOMPOSITION OF EARNINGS

	9 Consec.				20 Not Nec. Consec.			
	German Data		Danish Data		German Data		Danish Data	
	Levs. (1)	Diffs. (2)	Levs. (3)	Diffs. (4)	Levs. (5)	Diffs. (6)	Levs. (7)	Diffs. (8)
Perm. var., full sample, $\hat{\sigma}_{\xi}^2$	0.013	0.024	0.016	0.019	0.0096	0.018	0.013	0.019
Perm. var. due to irregular obs. surround. miss., $\hat{\sigma}_{\xi,o}^2$	0.034	0.158	0.053	0.124	–0.009	0.137	–0.004	0.133
Perm. var. net of effects of irregular obs., $\hat{\sigma}_{\xi,n}^2$	0.010	0.010	0.013	0.013	0.0097	0.0097	0.013	0.013
Trans. var., full sample, $\hat{\sigma}_{\epsilon}^2$	0.020	0.008	0.014	0.009	0.018	0.007	0.019	0.009
Trans. var., due to irregular obs. surround. miss., $\hat{\sigma}_{\epsilon,o}^2$	0.143	0.011	0.104	0.022	0.162	0.011	0.173	0.030
Trans. var. net of effects of irregular obs., $\hat{\sigma}_{\epsilon,n}^2$	0.007	0.007	0.008	0.008	0.007	0.007	0.008	0.008

Notes: The variances are calculated as in Equations (2)–(4). See Section 3.1 for the details on the construction of the samples. "9 consec." sample contains individual earnings spells with nine or more consecutive earnings observations. "20 not nec. consec." sample contains individual earnings spells with at least twenty not necessarily consecutive earnings observations.

Similarly, in the German nonconsecutive sample, the variances of permanent shocks are $\sigma_{\xi,l}^2 = 0.0096$, $\sigma_{\xi,l,n}^2 = 0.0097$, $\sigma_{\xi,d}^2 = 0.018$, and $\sigma_{\xi,d,n}^2 = 0.0097$, whereas the variances of transitory shocks are $\sigma_{\epsilon,l}^2 = 0.018$, $\sigma_{\epsilon,l,n}^2 = 0.007$, $\sigma_{\epsilon,d}^2 = 0.007$, and $\sigma_{\epsilon,d,n}^2 = 0.007$. Netting out the influence of missing observations and the influence of the first and last records in the earnings spells eliminates most of the discrepancy between the variances of permanent and transitory shocks in differences and levels.

4.2. *Restricting Unbalanced Samples.* One approach to eliminating the impact of low mean and high variance of observations at the start and end of earnings spells on the estimates of the permanent/transitory decomposition is to simply drop those observations. Accordingly, in Tables 7 and 8, columns (3) and (4) of Panel A, we repeat our analysis of Table 3 using the German and Danish samples with nine or more consecutive observations after dropping the first three observations for individuals whose earnings spells start after the first sample year and the last three observations for individuals whose earnings spells end before the last sample year.[29] In the same columns in Panel B, we, in addition, drop three observations before and after a missing earnings record in the nonconsecutive samples.

For the sample with nine or more consecutive observations, doing so barely affects the persistence of permanent shocks, although their variance estimated using the moments in differences is reduced by about 70%. The variance of transitory shocks estimated using the moments in levels is reduced by about 60%. We observe a similar pattern in the nonconsecutive sample in Panel B. As a result, the estimated earnings process is virtually identical in estimations utilizing the moments for growth rates and levels.

A comparison with Table 3 also indicates, consistent with the analysis in Section 2, that in both data sets, the variance of the permanent component is more robustly estimated using

[29] In Online Appendix Table A-7, we show that, for our administrative data, dropping one observation around missing earnings records is nearly as good as dropping three observations for reconciling the estimated earnings process in levels and growth rates. In other data sets, dropping the first and last records only may not be enough to eliminate the biases in the estimated variances of permanent and transitory shocks in levels and differences as the records in a few subsequent or preceding years may still be somewhat different from the rest of the earnings observations. For example, it is well known that the hazard of leaving a job is high at low tenure levels. As people who are observed starting earnings spells in the data often have new jobs, they may have an elevated frequency of nonemployment in subsequent couple of years. To the extent this is the case, dropping, say, the first observation only leads to the second irregular observation being next to the missing one and induces the associated biases. Dropping three observations minimizes such concerns.

TABLE 7

ESTIMATES OF THE EARNINGS PROCESS IN UNBALANCED SAMPLES: GERMAN DATA

	Full Sample		Drop First & Last Three Obs.		Model Obs. around Miss.	
	Levs. (1)	Diffs. (2)	Levs. (3)	Diffs. (4)	Levs. (5)	Diffs. (6)
A: 9 Consec. Sample						
$\hat{\phi}_p$	0.976	0.992	0.982	0.994	0.973	1.0
	(0.001)	(0.0008)	(0.001)	(0.001)	(0.001)	(0.002)
$\hat{\sigma}_\xi^2$	0.0078	0.019	0.006	0.005	0.008	0.005
	(0.0002)	(0.0003)	(0.0001)	(0.0001)	(0.0002)	(0.0002)
$\hat{\theta}$	0.129	0.153	0.197	0.186	0.142	0.141
	(0.005)	(0.009)	(0.007)	(0.007)	(0.004)	(0.006)
$\hat{\sigma}_\epsilon^2$	0.024	0.009	0.010	0.009	0.01	0.01
	(0.0003)	(0.0002)	(0.0002)	(0.0002)	(0.0002)	(0.0002)
$\hat{\sigma}_\alpha^2$	0.024	—	0.019	—	0.022	—
	(0.002)	—	(0.002)	—	(0.001)	—
B: 20 Not Nec. Consec. Sample						
$\hat{\phi}_p$	0.993	0.991	0.992	0.995	0.999	0.994
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.002)
$\hat{\sigma}_\xi^2$	0.0048	0.009	0.0047	0.0046	0.0046	0.0057
	(0.0001)	(0.0002)	(0.0001)	(0.0001)	(0.0001)	(0.0002)
$\hat{\theta}$	0.119	0.192	0.204	0.190	0.176	0.167
	(0.008)	(0.008)	(0.007)	(0.008)	(0.007)	(0.006)
$\hat{\sigma}_\epsilon^2$	0.016	0.009	0.009	0.008	0.0095	0.0096
	(0.0003)	(0.0002)	(0.0002)	(0.0002)	(0.0003)	(0.0002)
$\hat{\sigma}_\alpha^2$	0.027	—	0.021	—	0.029	—
	(0.002)	—	(0.001)	—	(0.001)	—

Notes: See Section 3.1 for the details on the construction of the samples. "9 consec." sample contains individual earnings spells with nine or more consecutive earnings observations. "20 not nec. consec." sample contains individual earnings spells with at least twenty not necessarily consecutive earnings observations. Columns (1) and (2) reproduce the corresponding estimates from Table 3. In columns (1)–(4), the estimated earnings process is: $y_{it} = \alpha_i + p_{it} + \tau_{it}$, where $p_{it+1} = \phi_p p_{it} + \xi_{it+1}$ and $\tau_{it+1} = \epsilon_{it+1} + \theta\epsilon_{it}$. In columns (5)–(8), the estimated earnings process is in Equation (5) in the text. Models are estimated using the optimally weighted minimum distance method. Asymptotic standard errors are in parentheses. German data span the period 1984–2008, while Danish data span the period 1981–2006. Full estimation results for the models in columns (5)–(6) are contained in Online Appendix Table A-12.

the moments in levels, whereas the variance of the transitory component, net of the transitory variation in earnings due to ν-shocks, is more robustly estimated using the moments in differences.

4.2.1. *Validation using simulated data.* We now simulate artificial earnings panels, consistent with the properties of the consecutive and nonconsecutive samples in our German data. In this experiment, we know the true parameters of the earnings process, which allows us to assess the performance of the proposed empirical methods and procedures in recovering the true values of these parameters. Specifically, we verify that: (i) the observed low mean and high variance of observations surrounding the missing ones induces a large discrepancy in the estimated variances of shocks when targeting the moments in levels and differences, (ii) the variance of permanent (transitory) shocks is more robustly estimated using the moments in levels (differences), and (iii) dropping observations surrounding the missing ones reconciles the estimates when using the moments in differences and levels and yields accurate estimates of the permanent and transitory shocks.

We replicate our German unbalanced samples in terms of the number of person-year observations and the age distribution and assume that incomes in the spells starting (ending) in the years other than the first (last) year of the sample are, in addition, affected by the ν-shocks. For the consecutive sample, we assume that the persistence of the permanent component is 0.980, the variance of permanent shocks is 0.008, the persistence of the transitory

TABLE 8
ESTIMATES OF THE EARNINGS PROCESS IN UNBALANCED SAMPLES: DANISH DATA

	Full Sample		Drop First & Last Three Obs.		Model Obs. around Miss.	
	Levs. (1)	Diffs. (2)	Levs. (3)	Diffs. (4)	Levs. (5)	Diffs. (6)
A: 9 Consec. Sample						
$\hat{\phi}_p$	0.955	0.987	0.957	0.980	0.954	0.985
	(0.001)	(0.0004)	(0.001)	(0.0006)	(0.0004)	(0.0004)
$\hat{\sigma}_\xi^2$	0.008	0.013	0.007	0.007	0.008	0.006
	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.00004)	(0.00004)
$\hat{\theta}$	0.204	0.209	0.226	0.220	0.233	0.190
	(0.002)	(0.003)	(0.003)	(0.003)	(0.002)	(0.002)
$\hat{\sigma}_\epsilon^2$	0.019	0.012	0.012	0.011	0.013	0.014
	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)
$\hat{\sigma}_\alpha^2$	0.020	—	0.019	—	0.024	—
	(0.0004)	—	(0.0004)	—	(0.0002)	—
B: 20 Not Nec. Consec. Sample						
$\hat{\phi}_p$	0.964	0.982	0.963	0.979	0.963	0.981
	(0.001)	(0.0006)	(0.001)	(0.0007)	(0.0004)	(0.0004)
$\hat{\sigma}_\xi^2$	0.007	0.012	0.006	0.007	0.007	0.008
	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.00004)	(0.0001)
$\hat{\theta}$	0.137	0.217	0.225	0.221	0.219	0.204
	(0.003)	(0.003)	(0.003)	(0.003)	(0.002)	(0.002)
$\hat{\sigma}_\epsilon^2$	0.022	0.013	0.013	0.011	0.014	0.013
	(0.0002)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)
$\hat{\sigma}_\alpha^2$	0.023	—	0.020	—	0.022	—
	(0.0004)	—	(0.0004)	—	(0.0002)	—

Notes: See Section 3.1 for the details on the construction of the samples. "9 consec." sample contains individual earnings spells with nine or more consecutive earnings observations. "20 not nec. consec." sample contains individual earnings spells with at least twenty not necessarily consecutive earnings observations. Columns (1) and (2) reproduce the corresponding estimates from Table 3. In columns (1)–(4), the estimated earnings process is: $y_{it} = \alpha_i + p_{it} + \tau_{it}$, where $p_{it+1} = \phi_p p_{it} + \xi_{it+1}$ and $\tau_{it+1} = \epsilon_{it+1} + \theta\epsilon_{it}$. In columns (5)–(8), the estimated earnings process is in Equation (5) in the text. Models are estimated using the optimally weighted minimum distance method. Asymptotic standard errors are in parentheses. German data span the period 1984–2008, while Danish data span the period 1981–2006. Full estimation results for the models in columns (5)–(8) are contained in Online Appendix Table A-12.

component is 0.170, the variance of transitory shocks is 0.010, and the variance of fixed effects is 0.025. These values are similar to the estimates of the transitory component using the moments in growth rates and of the permanent component using the moments in levels in Table 3, columns (1)–(2). We assume that the shocks and fixed effects are drawn from Student's t-distributions with four degrees of freedom, since our samples have high excess kurtosis.[30] We take the means and variances of the ν-shocks in the first three and last three periods from columns (1) and (5) of Table 4. Following the same strategy for the nonconsecutive sample, we assume that the persistence of the permanent component is 0.999, the variance of permanent shocks is 0.005, the persistence of the transitory component is 0.20, the variance of transitory shocks is 0.01, and the variance of fixed effects is 0.025. This is in line with the estimated permanent component in column (5) and transitory component in column (6) of Table 3. We take the means and variances of ν-shocks in the first three and last three periods, and three years before and three years after missing earnings records, from columns (3) and (7) of Table 4, and assume that the shocks follow the moving-average structure of order 1 with the persistence equal to 0.20.

The results for estimations fitting the entire set of autocovariances of the (simulated) data in levels or growth rates are in Table 9. The results are averages across 100 simulations.

[30] Assuming normal shocks instead has no impact on our findings. The choice of degrees of freedom for a Student's t-distribution of the shocks is consistent with the data; see footnote (35).

TABLE 9
ESTIMATES OF THE EARNINGS PROCESS IN UNBALANCED SAMPLES: SIMULATED "GERMAN" DATA

	9 Consec.				20 Not Nec. Consec.			
	Full Sample		Drop		Full Sample		Drop	
	Levs.	Diffs.	Levs.	Diffs.	Levs.	Diffs.	Levs.	Diffs.
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
$\hat{\phi}_p$	0.979	0.988	0.980	0.980	0.997	0.995	0.999	0.999
	(0.001)	(0.001)	(0.0009)	(0.001)	(0.0007)	(0.001)	(0.0006)	(0.0007)
$\hat{\sigma}_\xi^2$	0.008	0.016	0.008	0.008	0.005	0.009	0.005	0.005
	(0.0001)	(0.0006)	(0.0001)	(0.0001)	(0.0001)	(0.0003)	(0.0001)	(0.0001)
$\hat{\theta}$	0.133	0.143	0.170	0.170	0.152	0.189	0.20	0.20
	(0.003)	(0.005)	(0.004)	(0.004)	(0.007)	(0.004)	(0.006)	(0.003)
$\hat{\sigma}_\epsilon^2$	0.018	0.009	0.01	0.01	0.014	0.01	0.01	0.01
	(0.0005)	(0.0001)	(0.0001)	(0.0001)	(0.0003)	(0.0001)	(0.0002)	(0.0001)
$\hat{\sigma}_\alpha^2$	0.025	—	0.025	—	0.024	—	0.024	—
	(0.002)	—	(0.002)	—	(0.002)	—	(0.003)	—

Notes: The true earnings process is in Equation (5). In columns (1)–(4), $\sigma_\alpha^2 = 0.025$, $\phi_p = 0.98$, $\sigma_\xi^2 = 0.008$, $\theta = 0.170$, $\sigma_\epsilon^2 = 0.01$, while in columns (5)–(8), $\sigma_\alpha^2 = 0.025$, $\phi_p = 0.999$, $\sigma_\xi^2 = 0.005$, $\phi_\tau = 0.20$, $\sigma_\epsilon^2 = 0.01$. In columns (3)–(4) the first 3 (last 3) observations are dropped if an individual's earnings spell starts (ends) later (earlier) than in 1984 (2008); in columns (7) and (8), in addition, three observations before and after missing earnings records are dropped. The results are the averages across 100 simulations. Our simulated samples replicate the respective samples from German data in terms of the number of individuals, the number of individual-year observations, and the age distribution. "9 consec." sample contains individual earnings spells with nine or more consecutive earnings observations. "20 not nec. consec." sample contains individual earnings spells with at least twenty not necessarily consecutive earnings observations. The model is estimated using the optimal weighting minimum distance method. Standard errors, calculated as the standard deviations of the estimates across simulations, are in parentheses.

Utilizing the full sample results in an overestimation of the variance of the permanent (transitory) shock in differences (levels) as we observed in the actual data in Table 3. The permanent component is recovered fairly well utilizing the moments in levels, whereas the transitory component is closer to the truth utilizing the moments in differences.[31] Dropping the three observations adjacent to missing records aligns the results of estimations in levels and differences and correctly recovers the parameters of the underlying earnings process.

4.3. *Modeling Earnings Records Surrounding the Missing Ones.* In columns (5)–(8) of Tables 7 and 8, instead of dropping observations surrounding the missing ones, we estimate an extended earnings process that explicitly models them.[32] Specifically, we estimate the following model:[33]

$$
\begin{aligned}
y_{it} &= \alpha_i + p_{it} + \tau_{it} + \chi_{it}, \quad t = t_0, \ldots, T, \\
p_{it} &= \phi_p p_{it-1} + \xi_{it}, \\
\tau_{it} &= \epsilon_{it} + \theta \epsilon_{it-1}, \\
\chi_{it+j} &= \begin{cases} \nu_{it} & \text{if } y_{it-1} \text{ or } y_{it+1} \text{ is missing and } t-1 \geq t_0, t+1 \leq T, \; j = 0 \\ \theta \nu_{it} & j = 1 \\ 0 & \text{otherwise.} \end{cases}
\end{aligned}
\tag{5}
$$

[31] We present additional simulations in Online Appendix F, in which we vary the size of the true variance and persistence of permanent and transitory shocks and find that this pattern remains robust.

[32] An alternative exposition of this earnings process, cast in terms of a selection model, is presented in Online Appendix H.

[33] Since most of the effects of irregular observations are due to the variation in hours, simultaneous modeling of hours and earnings along the lines of Altonji et al. (2013) could be an appealing approach for measuring the risk to earnings. However, besides modeling the variety of reasons for missing earnings records, which is challenging in itself, to be successful, this approach should also present a mechanism generating low and volatile earnings before and after missing earnings observations.

Columns (5) and (6) of Tables 7 and 8 contain the estimates of this earnings process, where we model the means and variances of the immediate observations around missing records by matching the regression coefficients from Table 4. We assume that ν_{it}'s are drawn from distributions with means and variances that depend on whether an individual has missing observations in the interior of a nonconsecutive earnings spell, at the beginning of a consecutive earnings spell, or at the end of a consecutive earnings spell (the corresponding means and variances have superscripts m, f, and l, respectively, in Online Appendix Table A-12, which contains full estimation results). We further assume that the persistence of the shock ν_{it}, θ, is the same as the persistence of the ϵ_{it} shock because the estimated persistence of the transitory component barely changes when we drop observations surrounding the missing ones, which can be verified by comparing the results in columns (3)–(4) with the results in columns (1)–(2).[34] As before, the other moments used for estimating the model are the autocovariance moments in either levels or differences. We rely on the simulated minimum distance method, assuming that observations are missing at random, all of the innovations are i.i.d. normal, and utilizing the optimal weighting matrix estimated by block-bootstrap.[35] As expected, estimating the extended earnings process results in a substantial reduction of the estimated variance of permanent shocks in differences and transitory shocks in levels.[36,37]

4.4. *Implications.* The experiments described in this section suggest that no other mechanisms but the presence of irregular earnings observations around missing earnings records are responsible for the discrepancy in the estimated variances of the shocks in levels and differences in our administrative earnings data. The practical implications of these findings depend on the objective of the analysis. If one is interested in the properties of permanent and transitory components as well as in the detailed analysis of earnings at the start and end of employment spells, one can estimate the extended process in Equation (5) where the mean and the variance of the shocks at the beginning and the end of contiguous earnings histories are readily identified from the mean and the variance of earnings in those periods. Many macromodels are too stylized, however, to incorporate explicit treatment of these observations.[38]

[34] We allow the persistence for symmetry with the modeling of the transitory shock, but imposing that it is zero (or another plausible value) does not affect the results.

[35] Because it is well known that earnings shocks are non-Gaussian, we have also tried estimations that assume that the shocks are drawn from a Student's t-distribution with the degrees of freedom estimated from the data by matching kurtosis of the growth in earnings observed in the data. We found that the point estimates in Online Appendix Table A-12 were virtually the same (with the estimated degrees of freedom of the Student's t-distribution equal to about 4, implying a leptokurtic distribution of the shocks). This is not surprising since the discrepancy in the estimated variances is the feature of the second moments of the data—and not the higher-order moments—as is highlighted in (D1)–(L2).

[36] It is clear from the precision of our estimates that allowing for the means and variances of ν-shocks is not redundant in fitting the data moments; for example, the quasi-likelihood ratio test's p-values for excluding $\sigma^2_{\nu_t^f}$, $\sigma^2_{\nu_t^l}$, $\mu_{\nu_t^f}$, and $\mu_{\nu_t^l}$ in the estimation of column (1) Online Appendix Table A-12 are all well below 1%.

[37] In the previous version of the article, we allowed for modeling of three observations adjacent to the missing ones, with little difference to our results on the variances of permanent and transitory earnings shocks and their persistence. This is not surprising. When estimating the extended earnings process, it is only essential to account for the observations immediately adjacent to the missing ones. The other irregular—but much less different—observations will be subsumed in the estimated variance of transitory shocks (as the data include earnings before and after those earnings records that allow to detect mean reversion of those shocks if the properties of the observation immediately adjacent to the missing one are controlled for).

[38] The approach in the macroliterature can be partially justified by noting that these properties of observations adjacent to the missing ones are largely induced by the measurement timing and frequency in the data that are not actually relevant in the model. To see this, consider an example. An individual graduates from college and starts working on July 1 of some year, works at the same job making $100 a day for 40 years, and retires on June 31. If a researcher could observe daily earnings, she would conclude that this individual faces no risk to earnings. Even if the data were annual, but the year was defined to run from July 1 to June 31, a researcher will reach the same conclusion. But if a researcher can only observe annual income based on a calendar year running from January 1 to December 31, she will conclude that there is high risk in the first and last year when the individual is observed with positive earnings. Clearly, this "risk" is just an artifact of available data and not the risk that needs to be modeled.

They use as an input only the permanent/transitory components of the earnings process, as in Equation (1), so that it becomes crucial to obtain the correct estimates of the stochastic properties of these components in the data. These components can be estimated using the extended earnings process in Equation (5), although simpler alternatives are also available. First, we have shown theoretically and verified empirically that the variance of the transitory shock is estimated with no bias when estimation is based on the moments for earnings growth rates if the ν-shock is not serially correlated and the variance of the permanent shock is recovered well when estimation targets the moments in levels. One could therefore use the estimated permanent component from targeting the moments in levels and the estimated transitory component from targeting the moments in growth rates. An alternative approach is to estimate the earnings process in Equation (1) on the data that do not include the observations surrounding the missing ones. As we have shown, this solution recovers the true parameters of this abbreviated process quite well.

In Online Appendix H, we assess the performance of these approaches using the standard calibrated incomplete markets model. Specifically, we first calibrate a model using the canonical earnings process with permanent and transitory shocks extended to include missing observations and low mean/high variance of observations surrounding the missing ones. The parameters of this process come from our estimates based on the German administrative data. We then simulate a data set from this model and treat it as if true data available to a researcher. We ask this hypothetical researcher first to estimate the canonical earnings process (which includes only permanent and transitory shocks) using the standard approach that does not consider the presence of low-mean and high-variance observations surrounding the missing ones. Next, we ask her to use the approach proposed in this article. After that, the researcher is asked to use the two estimated earnings processes as inputs into a standard incomplete markets model that does not feature missing observations. Finally, we ask which of the two calibrated models serves as a good guide to the objects of interest in the original model that generated the data. We find that if the researcher followed the strategy proposed in this article by measuring the variances of permanent and transitory shocks while accounting for the properties of observations next to the missing ones, she would have reached largely correct conclusions. In contrast, had she followed the standard practice in the literature and ignored the features of earnings observations surrounding the missing ones when estimating the earnings process, some of her conclusions would be grossly erroneous. This exercise suggests that obtaining correct variances of permanent and transitory shocks in the data is of first-order importance, whereas the loss from not modeling missing observations and the properties of observations surrounding them in incomplete markets models is much smaller.

We, of course, recognize that ignoring missing observations in quantitative models is a shortcut. A precise assessment of the amount of risk that individuals face does depend on what happens to earnings when they are missing in the data. However, it seems very difficult to know this because of the great variety of reasons for why earnings could be missing in administrative data, as we discuss in Online Appendix B. Depending on the reason, true unobserved earnings may plausibly be the same or higher/lower than the observed ones, and one would have to separately model these pathways into and out of missing observations (along the lines of, e.g., Altonji et al. 2013). This would, in theory, allow one to obtain more precise estimates for the variances of permanent and transitory shocks to earnings and a more accurate assessment in a quantitative model. Even if we knew why a particular observation is missing, such an exercise would be extremely challenging.

We do not expect the literature to overcome these challenges in the foreseeable future. Until that happens, however, we offer a constructive approach that follows the literature in utilizing in estimation all computable individual-level moments that do not involve missing values and not modeling missing observations in quantitative incomplete markets models. Researchers taking this currently unavoidable shortcut need to be careful. As we show, ignoring the unique properties of earnings observations around the missing observations would lead to severely biased estimates of the variances of permanent and transitory shocks in growth rates and levels and erroneous inference in quantitative models that use these shock variances as an input.

5. CONCLUSION

Properties of the earnings process play an important role in various areas of macro and labor economics. Different specifications of this process have been explored in the literature. The most widely used specification is based on decomposing earnings into the sum of persistent and transitory components and often assumes that the persistent component follows a random walk. The parameters of such a process can be identified using the moments based on earnings growth rates (first difference in log earnings) or the moments based on log earnings levels. Historically, the former approach is more common in labor economics, whereas the latter is more common in the macroeconomics literature. Unfortunately, these two approaches lead to dramatically different estimates of the variances of permanent and transitory components. In particular, using the same set of observations in the data, the variance of persistent shocks is typically estimated to be much higher when the moments in growth rates are targeted, whereas the variance of transitory shocks is found to be much higher when the estimation is based on fitting the moments in levels. This discrepancy has important implications for substantive economic analysis. For example, the earnings process drives the heterogeneity in Bewley-type models with incomplete markets, and the variances of earnings components determine not only economic choices, such as consumption and savings, but also the optimal design of policies, such as taxes and transfers. Moreover, the standard approach to estimating the amount of insurance that individuals have against permanent and transitory shocks in the data relies on the estimated variances of permanent and transitory components. The uncertainty about the size of these variances translates into uncertainty regarding the right amount of insurance that should be generated by the widely used incomplete markets models and the associated uncertainty about the results of welfare analyses using those models.

In this article, we uncovered the data features that can quantitatively account for the large difference in the estimates based on earnings growth rates and levels, at least in the administrative data from Denmark and Germany. In particular, we found that earnings are lower on average and more volatile at the start and end of continuous earnings spells. We have shown theoretically that these irregular earnings observations, which are either preceded or followed by a missing observation, induce an upward bias in the estimates of the variance of permanent shocks based on the moments in differences and the variance of transitory shocks when estimation is based on the moments in levels. Thus, even when working with very large administrative data sets with highly reliable information, one must remain vigilant because natural features of the data sets, like the low mean and high variance of earnings at the start and end of earnings spells, can induce extremely large biases in the estimated earnings processes.

Although the primary focus of this article is on estimating earnings processes on large administrative data sets that are becoming central in the literature, the mechanism we describe also applies to survey-based PSID data on hourly wages, individual earnings, combined earnings of husbands and wives, and net family income.[39] This is important, in part, because the PSID data are the primary source of information on the extent of consumption insurance achieved by U.S. households and on the sources of insurance, such as household savings and borrowing or the public transfer and tax system. Hryshko and Manovskii (2018) find that correcting the measurement of family income dynamics for the effects of the low mean and high variance of the observations surrounding the missing ones as proposed in this article leads to a very different assessment of the degree and sources of consumption insurance relative to the key empirical benchmark for incomplete markets models provided by Blundell et al. (2008).

Our punchline is that there clearly must be a tight correspondence between the components of the earnings process used in the model and the ones estimated in the data. If the model

[39] See Online Appendix G and Hryshko and Manovskii (2018) for details.

utilizes a version of the permanent/transitory process as in Equation (1), the measurement of these permanent/transitory components in the data must not be biased due to the special nature of observations adjacent to the missing ones. If, on the other hand, one wishes to incorporate the features of these observations into the model, one needs to estimate the extended process in Equation (5) in the data and use it as an input into the model. The article provides a way to implement both of these approaches.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Supporting Information

REFERENCES

ABOWD, J., AND D. CARD, "On the Covariance Structure of Earnings and Hours Changes," *Econometrica* 57 (1989), 411–45.

ALTONJI, J. G., A. A. SMITH JR., AND I. VIDANGOS, "Modeling Earnings Dynamics," *Econometrica* 81 (2013), 1395–454.

ARELLANO, M., AND B. HONORÉ, "Panel Data Models: Some Recent Developments," in J. J. Heckman and E. E. Leamer, eds., *Handbook of Econometrics*, Volume 5, Chapter 53 (Amsterdam: Elsevier, 2001), 3229–96.

ATTANASIO, O., E. HURST, AND L. PISTAFERRI, "The Evolution of Income, Consumption, and Leisure Inequality in the United States, 1980–2010," in C. D. Carroll, T. F. Crossley, and J. Sabelhaus, eds., *Improving the Measurement of Consumer Expenditures*, Chapter 4 (Chicago: University of Chicago Press, 2015), 100–40.

BANKS, J., AND P. DIAMOND, "The Base for Direct Taxation," in The Institute for Fiscal Studies (IFS), ed., *Dimensions of Tax Design: The Mirrlees Review*, Chapter 6 (Oxford: Oxford University Press for The Institute for Fiscal Studies, 2010), 548–648.

BLUNDELL, R., M. GRABER, AND M. MOGSTAD, "Labor Income Dynamics and the Insurance from Taxes, Transfers, and the Family," *Journal of Public Economics* 127 (2015), 58–73.

———, L. PISTAFERRI, AND I. PRESTON, "Consumption Inequality and Partial Insurance," *American Economic Review* 98 (2008), 1887–921.

BROWNING, M., AND M. EJRNÆS, "Heterogeneity in the Dynamics of Labor Earnings," *Annual Review of Economics* 5 (2013), 219–45.

———, ———, AND J. ALVAREZ, "Modelling Income Processes with Lots of Heterogeneity," *Review of Economic Studies* 77 (2010), 1353–81.

CARROLL, C. D., "Buffer-Stock Saving and the Life Cycle/Permanent Income Hypothesis," *Quarterly Journal of Economics* 112 (1997), 1–55.

CASTAÑEDA, A., J. DÍAZ-GIMÉNEZ, AND J.-V. RÍOS-RULL, "Accounting for the U.S. Earnings and Wealth Inequality," *Journal of Political Economy* 111 (2003), 818–57.

DEATON, A., "Saving and Liquidity Constraints," *Econometrica* 59 (1991), 1121–48.

DEBACKER, J., B. HEIM, V. PANOUSI, S. RAMNATH, AND I. VIDANGOS, "Rising Inequality: Transitory or Persistent? New Evidence from a Panel of U.S. Tax Returns," *Brookings Papers on Economic Activity* 2013 (2013), 67–142.

DOMEIJ, D., AND M. FLODÉN, "Inequality Trends in Sweden 1978–2004," *Review of Economic Dynamics* 13 (2010), 179–208.

DUSTMANN, C., J. LUDSTECK, AND U. SCHÖNBERG, "Revisiting the German Wage Structure," *Quarterly Journal of Economics* 124 (2009), 843–81.

FARHI, E., AND I. WERNING, "Insurance and Taxation over the Life Cycle," *Review of Economic Studies* 80 (2012), 596–635.

FITZGERALD, J., P. GOTTSCHALK, AND R. MOFFITT, "An Analysis of Sample Attrition in Panel Data: The Michigan Panel Study of Income Dynamics," *Journal of Human Resources* 33 (1998), 251–99.

FRIEDMAN, M., *A Theory of the Consumption Function* (Princeton, NJ: Princeton University Press, 1957).

———, AND S. KUZNETS, *Income from Independent Professional Practice* (New York: National Bureau of Economic Research, 1954).

GOTTSCHALK, P., AND R. A. MOFFITT, "The Growth of Earnings Instability in the U.S. Labor Market," *Brookings Papers on Economic Activity* 2 (1994), 217–54.

Guvenen, F., "An Empirical Investigation of Labor Income Processes," *Review of Economic Dynamics* 12 (2009), 58–79.

———, S. Ozcan, and J. Song, "The Nature of Countercyclical Income Risk," *Journal of Political Economy* 122 (2014), 621–60.

Hagedorn, M., J. Luo, I. Manovskii, and K. Mitman, "Forward Guidance," *Journal of Monetary Economics* 102 (2019a), 1–23.

———, I. Manovskii, and K. Mitman, "The Fiscal Multiplier," National Bureau of Economic Research, Working Paper 25571, 2019b.

Heathcote, J., "Fiscal Policy with Heterogeneous Agents and Incomplete Markets," *Review of Economic Studies* 72 (2005), 161–88.

———, F. Perri, and G. L. Violante, "Unequal We Stand: An Empirical Analysis of Economic Inequality in the United States: 1967–2006," *Review of Economic Dynamics* 13 (2010a), 15–51.

———, K. Storesletten, and ———, "The Macroeconomic Implications of Rising Wage Inequality in the United States," *Journal of Political Economy* 118 (2010b), 681–722.

———, ———, and ———, "Consumption and Labor Supply with Partial Insurance: An Analytical Framework," *American Economic Review* 104 (2014), 2075–126.

Hoffmann, E. B., and D. Malacrino, "Employment Time and the Cyclicality of Earnings Growth," *Journal of Public Economics* 169 (2019), 160–71.

Hryshko, D., "Labor Income Profiles Are Not Heterogeneous: Evidence From Income Growth Rates," *Quantitative Economics* 3 (2012), 177–209.

———, and I. Manovskii, "How Much Consumption Insurance in the U.S.?," Mimeo, University of Alberta and University of Pennsylvania, 2017.

———, and ———, "Income Dynamics and Consumption Insurance," Mimeo, University of Alberta and University of Pennsylvania, 2018.

Hubbard, R. G., J. Skinner, and S. P. Zeldes, "Precautionary Saving and Social Insurance," *Journal of Political Economy* 103 (1995), 360–99.

Huggett, M., G. Ventura, and A. Yaron, "Sources of Lifetime Inequality," *American Economic Review* 101 (2011), 2923–54.

Kaplan, G., B. Moll, and G. Violante, "Monetary Policy According to HANK," *American Economic Review* 108 (2018), 697–743.

Krueger, D., and F. Perri, "Does Income Inequality Lead to Consumption Inequality? Evidence and Theory," *Review of Economic Studies* 73 (2006), 163–93.

———, and ———, L. Pistaferri, and G. L. Violante, "Cross-sectional Facts for Macroeconomists," *Review of Economic Dynamics* 13 (2010), 1–14.

Livshits, I., J. MacGee, and M. Tertilt, "Consumer Bankruptcy: A Fresh Start," *American Economic Review* 97 (2007), 402–18.

MaCurdy, T. E., "The Use of Time Series Processes to Model the Error Structure of Earnings in a Longitudinal Data Analysis," *Journal of Econometrics* 18 (1982), 83–114.

———, "A Practitioner's Approach to Estimating Intertemporal Relationships Using Longitudinal Data: Lessons from Applications in Wage Dynamics," in J. J. Heckman and E. E. Leamer, eds., *Handbook of Econometrics*, Volume 6A, Chapter 62 (Amsterdam: Elsevier, 2007), 4057–167.

Meghir, C., and L. Pistaferri, "Income Variance Dynamics and Heterogeneity," *Econometrica* 72 (2004), 1–32.

———, and ———, "Earnings, Consumption, and Life Cycle Choices," in D. Card and O. Ashenfelter, eds., *Handbook of Labor Economics*, Volume 4B, Chapter 9 (Amsterdam: Elsevier, 2011), 773–854.

Moffitt, R. A., and P. Gottschalk, "Trends in the Transitory Variance of Male Earnings: Methods and Evidence," *Journal of Human Resources* 47 (2012), 204–36.

Postel-Vinay, F., and H. Turon, "On-the-Job Search, Productivity Shocks, and the Individual Earnings Process," *International Economic Review* 51 (2010), 599–629.