CENTRE for ECONOMIC PERFORMANCE

CEP Discussion Paper No 653

September 2004

Globalisation, ICT and the Nitty Gritty of Plant Level Datasets

Ralf Martin

Abstract
The net entry contribution to aggregate productivity growth has increased dramatically in the UK over 1990s according to calculations based on data from the Annual Respondents Database (ARD). Some recent studies have tried to link this to other structural changes over the same period such as increased globalisation and usage of ICT. I argue that the increase might equally have been caused by a systematic bias that is introduced to growth decompositions through random survey sampling of the underlying plant or firm panel datasets. This bias – despite being a general problem of growth decompositions does not seem to have been noticed in the literature yet. In the 1990s the Office for National Statistics (ONS) has successively increased the share of plants in the population of the ARD that are subject to random sampling. I show that this could cause the bias to spuriously increase the net entry contribution. My results show that correcting for the bias makes a substantial difference: the net entry contribution is about 10 percentage points lower on the corrected series in the 1990s. Surprisingly however, the positive correlation between ICT and net entry share – a main result of earlier studies – becomes more significant.

JEL classification: C1 F00 L25 L6
Keywords: Productivity Growth Decomposition, Micro Data, Random Sampling, Globalisation, ICT

This paper was produced as part of the Centre's Productivity and Innovation Programme. The Centre for Economic Performance is financed by the Economic and Social Research Council.

Acknowledgements
This work contains statistical data from ONS which is Crown copyright and reproduced with the permission of the controller of HMSO and Queen's Printer for Scotland. The use of the ONS statistical data in this work does not imply the endorsement of the ONS in relation to the interpretation or analysis of the statistical data. Thanks to Eric Bartelsman, Chiara Criscuolo and Jonathan Haskel for comments and discussions. This research was supported by the ESRC/EPSRC Advanced Institute of Management Research under grant number RES-331-25-0030.

Ralf Martin is an Occasional Research Assistant at the Centre for Economic Performance, London School of Economics. He is also a Researcher at the Centre for Research into Business Activity (CeRiBa).
Contact: r.martin@lse.ac.uk

Published by
Centre for Economic Performance
London School of Economics and Political Science
Houghton Street
London WC2A 2AE

ISBN 0 7530 1783 0

1 Introduction

An important contribution of the research using plant or firm level micro data sets has been the decomposition of aggregate productivity growth into components attributable to continuing, exiting and entering plants[1]. Initially research in this area had to rely on datasets spanning only a few years so that productivity growth decompositions could be undertaken for only one time period. Now increasingly longer time series are available allowing decomposition calculations for various time periods and it becomes increasingly possible to analyse changes in, for example, the net entry contribution to productivity over time. Criscuolo et al. (2004) have recently completed a study which suggests that the contribution of net entry had dramatically increased in Britain in the 1990s relative to the 1980s. Figure 1 which shows the contribution of exit and entry as well as net entry to productivity growth restate that point.[2] That study also shows evidence that the changes in the decomposition figures might be linked to globalisation and usage of ICT. Increased globalisation and the arrival of new ICT technology are clearly the major structural changes over the 1980 to 2000 period. However, over long time periods not only the economic environment is subject to major changes. Less noticed by the general public, economic statistics and surveys are subject to major changes in definitions, sampling rules, etc. Plant level datasets such as the Annual Respondents Database (ARD) are no exception. Over the 1980-2000 period the Office of National Statistics not only changed its name[3] but also revised the coding of ARD reference numbers entirely, did a major revision of the ARD register, changed the sector definitions completely, successively included more sectors of the economy, changed questions asked, included ever smaller plants into the survey and continously changed the rules for random sampling. All these changes and modifications might cause changes in measured productivity decomposition figures[4] on top of changes induced by genuine changes in the fundamental economic variables.

[1] examples of earlier studies include Bartelsman and Dhrymes (1998), Foster et al. (1998)

[2] Below I discuss in more detail the construction of these figures

[3] until 1995 it was known as Central Statistical Office

[4] ...as well as most other statistics we might wish to compute from the ARD

The main contribution of this paper is to examine how random survey sampling affects productivity decomposition calculations. More precisely, I show that random survey sampling introduces a systematic upward bias in the estimated contribution of net entry to productivity growth. While being a general, and so far unnoticed[5], problem of productivity growth decompositions it might be of particular relevance in the context of the apparent increase in the net entry contribution in the UK in the 1990s. Over this period the ARD was also subject to increased random sampling. Among other things the threshold for random sampling has successively been increased from plants with less than 100 employees to plants with less than 250 employees[6] Below I show that the bias – which I dub *Continuer's Bias* – might spuriously increase the measured net entry contribution in such a situation so that at least part of the apparent increase in net entry shares might simply be explained by increased random sampling.

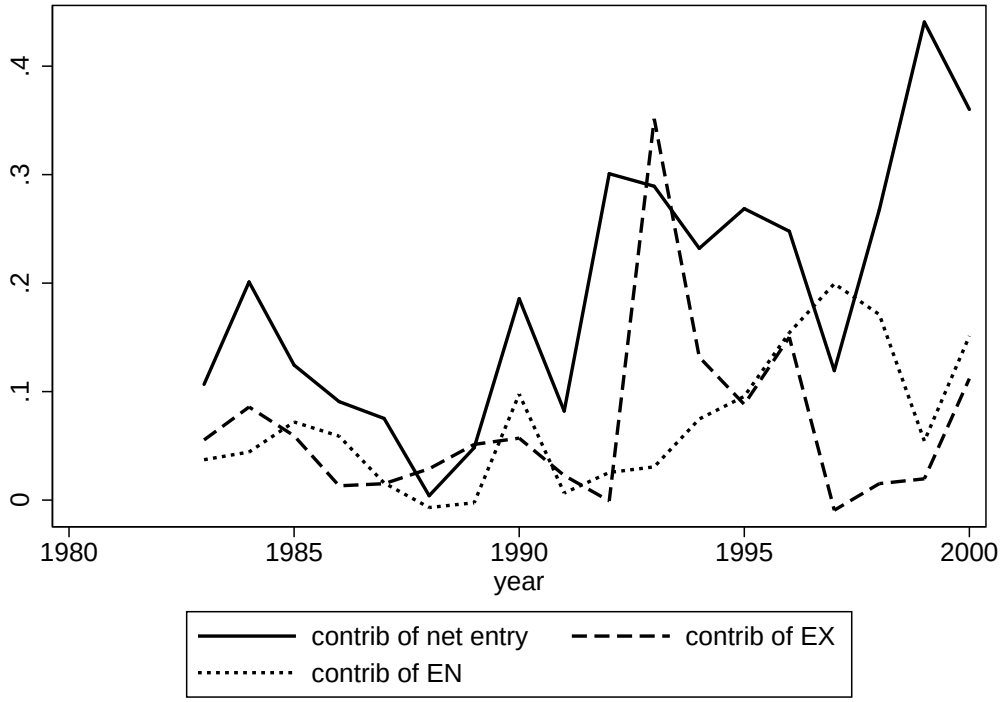
The intuition for the *Continuer's Bias* is as follows: In order to consider a continuing plant in any decomposition method we need to observe it in two consecutive time periods. Thus with random sampling it needs not only survive for these two periods but also be sampled in each of them. Because exit on the other hand is a piece of information which can be derived from the underlying register population – the complete population for all practical purposes – it is enough to be sampled only in the base year – as a matter of fact an exiting plant can *only* be observed in the base year. As a consequence of that exiting plants are over-represented in the sample which we can use for productivity decompositions relative to the underlying complete population. The same is true for entering plants which only need to – and can – be observed in the end year.

The potential importance of increased random sampling as an explanation for the increase in measured net entry shares is illustrated by figure 2 which shows the share of sampled plants in any given year that are affected by random sampling; i.e. those which are not sampled in every year they are alive. The figure reports raw and employment weighted shares. We see that the share of randomly sampled plants increases dramatically in the 1990s

[5]to the best of my knowledge

[6]For details see table 4.

Figure 1: Contribution of net entry to labour productivity growth
(Median over two digit manufacturing sectors)

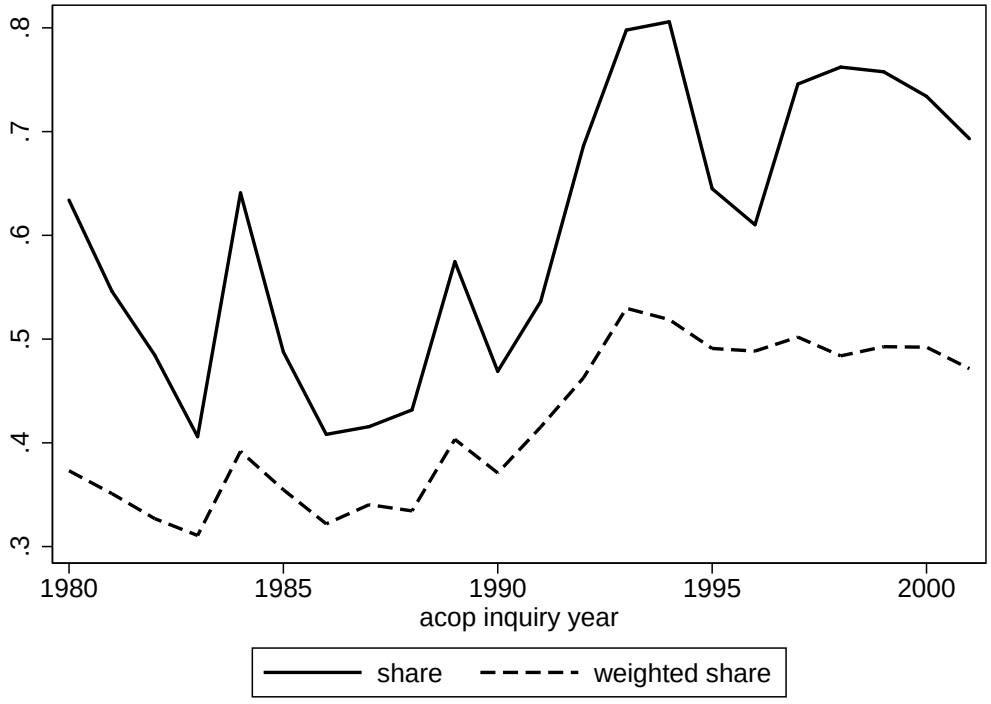


Source: Author's calculations based on ARD.

Notes: Labour productivity is valued over employment (headcount). Growth rates and contributions are calculated over three year intervals. Sectors 23, 36 and 37 have been dropped because of data problems. The median is used to ensure that the series is not dominated by outliers.

from somewhere around 60 percent to around 80 percent. Because randomly sampled plants are naturally smaller, the employment weighted series reports lower shares but there is still a dramatic increase in the 1990s. Below I show that this fundamental change in how the ONS conducts its surveys translates indeed in a significant change of the measured net entry contribution. My findings suggest that in the 1990s uncorrected measured net entry shares are about 10 percent too high.

Figure 2: Share of plants affected by random sampling
(Raw and employment shares)



Source: Author's calculations based on ARD.
Notes: Ratio of plants that are sampled in a given year but not sampled in all the years that they are alive over the total number sampled in that year. 'weighted' reports the same share in terms of employment.

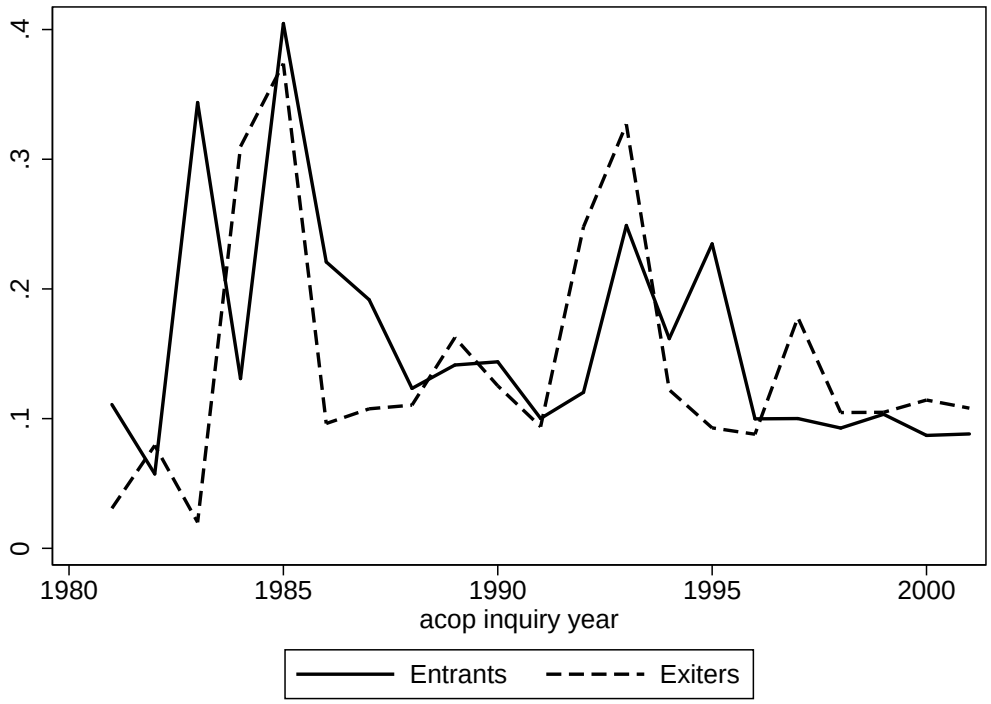
The rest of this paper is organised as follows: before examining the continuers' bias more carefully first theoretically in section 3 and then empirically in section 4, I will carefully discuss other data problems that might

have affected the productivity decomposition series in section 2. section **??** concludes.

2 The ARD from 1980 to 2000

For a more elaborate introduction to the Annual Respondents Database (ARD) consult Criscuolo et al. (2003). Here I focus on those aspects and

Figure 3: Share of exitors and entrants
(Annual calculations)

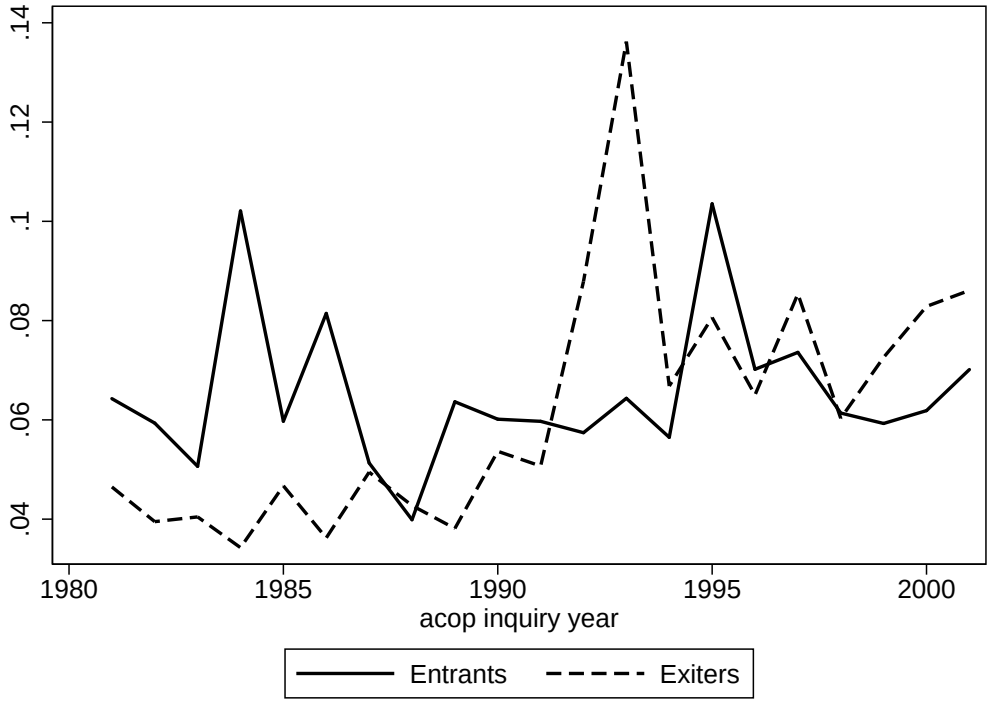


Source: Author's calculations based on ARD.
Notes: "Exiters" = plants that exist in $t-1$ but not in t as fraction of all plants. "Entrants" is defined accordingly.

problems of the data which are particularly relevant for productivity decomposition calculations. Figure 3 shows entering and exiting plants as share of all plants[7] over the 1980 to 2000 period. While the shares are generally fluc-

[7] i.e. continuing, entering and exiting

Figure 4: Share of exitors and entrants
(Calculations on sample used for decomposition calculations)



Source: Author's calculations based on ARD.
Notes: The definition of the series is as in figure 1

tuating somewhere between 10 and 20 percent there are distinct periods at the beginning of both the 80s and 90s with much higher rates. The concern is that these are not so much driven by genuine entry and exit but rather the particularities of how we measure these figures here. Figure 3 is computed from information on the register underlying the ARD. For our purposes the register consists of long lists of plant reference numbers which we observe in various years. If in year t_0 a certain number i, say, is observed but not in a later year t_1 – with $t_1 > t_0$ – the we count an exit for year t_1. Similarly, if a number is observed in year t_1 but not in year t_0 then this counted as an entry event. Continuers have to be observed in both years. Changes in the way the ONS organises its register can therefore lead to spurious exit and entry counts. The following register modifications over the sample period might have introduced errors in measured entry and exit rates.

- In 1983/84 the ONS undertook a major register overhaul. Before that the register suffered from unaccounted death of plants; i.e. the ONS had no structured way of registering the death of plants which meant that many plants which went out of business continued to be kept in the register. This explains the rather low exit rates for the 1980 to 83 period and the spikes in exit rates in 1983 and 84. The figures in those two years simply include all unaccounted for deaths from the earlier periods.

- From 1993 to 94 the ONS completely changed its register coding. What happened was that the earlier Annual Census of Prodduction(ACOP) register was replaced by the Interdepartmental Business Register(IDBR).[8] This meant that each plant register ID changed nominally to a different number. Unfortunately only incomplete records of the changeover exist and a lot of research time has been spent in recent years[9] to create a complete lookup table between the two reference keys. Figure 3 suggests that this effort has been fairly successful. The problem with this kind of register modification is that some plants might erro-

[8] Interdepartemental as in various government departments dealing with businesses use the same register

[9] in particular by Richard Harris and his colleagues

neously counted as exiters if we fail to find the new IDBR reference key for them. They actually would at the same time – also erroneously – counted as entrants. Thus widespread measurement error of this kind should show up as a spike in both exit and entry rates. 1994 – which would be the relevant year for the coding change – does not show such a simultaneous spike.

- Over time the register population changes simply because the ONS successively covered more sectors. The biggest change of this kind concerns the service sectors which have been added after 1997. However, with the inclusion of recycling as a separate sector since 1995, even within manufacturing there have been changes of this kind.

Besides changes in the register the ARD was subject to changes in sampling rules. The major change here has been the inclusion of plants with less than 20 employees after 1994. In earlier years they have not been sampled. Besides this major change the ONS is continously changing the sampling rules for various size bands of plants. Table 4 in the appendix summarises this. We see for example that from 1998 onwards only firms with more than 250 employees were included with probability 1 into the sample (last row of table 4). In earlier years the threshold was either 50 or 100 employees.

I take the following steps in order to ensure that the effects of these issues on my decomposition calculations are minimised. Firstly, spurious belated statistical deaths because of register overhaul are primarily a problem for very small rarely sampled units. As a consequence this issue should solve itself for the sample of plants which can be used for decomposition calculations as it is a subset of the sampled plants.[10]
Secondly, to avoid any spurious results from the inclusion of recycling I dropped the sector entirely together with the "not elsewhere classified" sector (sic92=36).
Thirdly, to make pre and post 1995 values comparable I give plants with less than 20 employees after 1994 a weight of zero in all decomposition calcula-

[10]Of course this does not imply that the decomposition calculations would not be different if we would have the chance to include them in our calculations. All we can do about this, however, is find appropriate sampling weights, an issue to which I come back below

tions.

To examine how these changes affect exit and entry rates consider figure 4 which recalculates entry and exit rates for the sample of plants that enter decomposition calculations. Figure 4 looks considerably more plausible than figure 3. The one thing that remains a bit suspicious is the spike of the exit rate in 1993. The fact that in that year the UK was suffering from a major recession and that there is not contemporaneous spike in the entry rate supports the idea that it reflects actual economic behaviour.

Besides that the figure shows that entry and – even more so – exit rates have increased over the sample period. This is in line with the findings from figure 1. However, as suggested in the introduction, part of this increased importance of exit and entry could be driven by changes in the sampling rules of the ONS reported in table 4. How to correct for that will be the focus of the reminder of this paper.

3 The continuer bias

To pin things down more clearly let's introduce some algebra. To focus I base my argument on one particular decomposition method, namely the one proposed by Foster et al. (1998)(FHK). Their decomposition applies to aggregated productivity measures y_t which we can write as

$$y_t = \sum_i \theta_{it} y_{it} \tag{1}$$

where y_{it} is the relevant plant level productivity measure and θ_{it} a weight based on plant level factor inputs,[11] L_{it} say. The natural example of such a productivity measure is value added per employee where θ_{it} becomes the employment share of plant i. FHK show that we can write the change in aggregate productivity between a period 1 and a base period 0, $\Delta y = y_1 - y_0$, as

$$\Delta y = \sum_C [\theta_{i0}\Delta y_i + \Delta\theta_i(y_{i0} - y_0) + \Delta\theta_i\Delta y_i] + \sum_N \theta_{i1}(y_{i1} - y_1) + \sum_E \theta_{i0}(y_{i0} - y_0) \tag{2}$$

[11] ...and possibly sampling weights, but let's ignore that for the time being.

where C, N and E are the sets of continuing, entering and exiting plants. The contribution of continuing plants consists of several sub-components: a component measuring the changes of productivity at continuing plants at constant market shares, a component derived from changes in market share and a residual interaction component. The performance of exiting and entering plants is expressed relative to the aggregate productivity in the base year. How FHK arrive at 2 is reproduced in the appendix for completeness. To make my argument it is helpful, however, to look at a less sophisticated way of writing Δy:

$$\Delta y = \sum_{C} \left[\theta_{i1} y_{i1} - \theta_{i0} y_{i0}\right] + \sum_{N} \theta_{i1} y_{i1} - \sum_{E} \theta_{i0} y_{i0} \tag{3}$$

If there is no random sampling then expressions such as 2 and 4 can be viewed as non probabilistic characterisations of the complete population of plants at various points in time. With random sampling what we do is calculate estimates of components of 4 by performing the same calculations on the sample only:

$$\tilde{\Delta y} = \sum_{i \in C \cap S^{dec}} \left[\tilde{\theta}_{i1} y_{i1} - \tilde{\theta}_{i0} y_{i0}\right] + \sum_{i \in N \cap S^{dec}} \tilde{\theta}_{i1} y_{i1} - \sum_{i \in X \cap S^{dec}} \tilde{\theta}_{i0} y_{i0} \tag{4}$$

where S^{dec} is the set of plants that can be used for the decomposition calculation; i.e. the continuing plants that are sampled in $t = 0$ and $t = 1$, the exiting plants sampled in $t = 0$ and the entering plants in $t = 1$. $\tilde{\theta}_{it}$ are the factor shares computed from the sample rather than the population.

A sufficient condition to make representative statements about the population from a sub sample S^{dec} is that the probability of finding a particular plant (or group of plants) in S^{dec} corresponds to its importance in the population:

$$P\{i \in Pop | i \in S^{dec}\} = P\{i \in Pop\} \tag{5}$$

Because by Bayes Rule

$$P\{i \in Pop | S^{dec}\} = \frac{P\{i \in S^{dec} | i \in Pop\}}{P\{i \in S^{dec}\}} P\{i \in Pop\} \tag{6}$$

where $P\{i \in S^{dec}\} = \sum_i P\{i \in S^{dec} | \imath \in Pop\} P\{i \in Pop\}$ is the marginal probability of a plant being sampled. Thus to have a representative sample

we must require that

$$P_i = \frac{P\{i \in S^{dec} | i \in Pop\}}{P\{i \in S^{dec}\}} = 1 \tag{7}$$

for all units i which is true if all i's have the same probability of being sampled:

$$P_i = P_j \text{ for all } i, j \tag{8}$$

If this is not satisfied the expected value of statistics such as 4 is in general not equal to the population value but becomes

$$E\{\tilde{\Delta y}\} = \sum_C \left[E\{\tilde{\theta}_{i1}\}y_{i1} - E\{\tilde{\theta}_{i0}\}y_{i0}\right] + \sum_N E\{\tilde{\theta}_{i1}\}y_{i1} - \sum_X E\{\tilde{\theta}_{i0}\}y_{i0} \tag{9}$$

with $E\{\tilde{\theta}_{it}\} = \frac{P_i L_{it}}{\sum_i L_{it}} \neq \theta_{it}$.
Note however that 5 – while sufficient – is actually stronger than needed. If the deviation of $E\{\tilde{\theta}_{it}\}$ from θ_{it} is neither correlated with y_{it} nor with the status of plants[12] then $\tilde{\theta}_{it}$ would still be consistent. More formally, if

$$E\{\tilde{\theta}_{it}\} = \theta_{it} + \varepsilon_{it} \tag{10}$$

where

$$\begin{gathered} E\{\varepsilon_{it}|y_{it}\} = E\{\varepsilon_{it}\} = 0 \text{ and} \\ E\{\varepsilon_{it}|i \in C\} = E\{\varepsilon_{it}|i \in E\} = E\{\varepsilon_{it}|i \in N\} = E\{\varepsilon_{it}\} = 0 \end{gathered} \tag{11}$$

then $E\{\tilde{\Delta y}\} = \Delta y$. However it is exactly the second part of condition 11 which is violated because of random sampling.

Let's introduce a little example to get a better grasp. Assume that the statistics agency which runs the plant survey draws entirely randomly a fraction of ρ plants each year from the population to survey. Then for continuing plants we have that, denoting the sampled plants in each year with S_t:

$$\begin{aligned} P\{i \in S^{dec} | i \in C\} &= P\{i \in S_0\} \cdot P\{i \in S_1 | i \in S_0\} \\ &= \rho^2 \end{aligned} \tag{12}$$

[12] i.e. the sorting of plants into exiters, entrants and continuers

whereas for exiting and entering plants it is

$$P\{i \in S^{dec} | i \in X\} = P\{i \in S_0\} = \rho \tag{13}$$

and

$$P\{i \in S^{dec} | i \in N\} = P\{i \in S_1\} = \rho \tag{14}$$

so that 11 does not hold. More in particular, because $P_{it}(C) < P_{it}(N)$, $P_{it}(X)$ we have that

$$E\{\varepsilon_{it} | i \in C\} < E\{\varepsilon_{it}\} \tag{15}$$

so that continuing plants will be under-represented in calculations such as 4. Luckily this description of the problem already entails its solution: To correct for the bias we need to multiply the contribution of continuing plants by the inverse of their lower relative sampling probability, i.e. we introduce the following weights

$$w_i = \begin{cases} \frac{P\{i \in S^{dec} | i \in N\}}{P\{i \in S^{dec} | i \in N\}} = \frac{1}{\rho} & \text{for } i \in C \\ \\ 1 \text{ otherwise} \end{cases} \tag{16}$$

Consequently an unbiased decomposition formula is

$$\hat{\Delta y} = \sum_{i \in C \cap S^{dec}} \left[\hat{\theta}_{i1} y_{i1} - \hat{\theta}_{i0} y_{i0} \right] + \sum_{i \in N \cap S^{dec}} \hat{\theta}_{i1} y_{i1} - \sum_{i \in X \cap S^{dec}} \hat{\theta}_{i0} y_{i0} \tag{17}$$

with

$$\hat{\theta}_{it} = \frac{w_i L_{it}}{\sum_{i \in S^{dec}} w_i L_{it}} \tag{18}$$

Is 17 feasible? In most practical case yes, because ρ is either known or can be estimated as the share of plants that are sampled each year, which requires availability of the underlying survey register of course.
Are the sample weights in 16 realistic? Hardly, because sampling procedures are rarely entirely uniform. The next sub-section discusses how to obtain weights in these more general situations.

3.1 Weights with non-random sampling

Sampling reality is more complex than suggested in the last section along various dimensions. These include

- Stratified sampling: Depending on their characteristics – often their size – plants are more or less likely to be sampled.
- Changes in the sample rules between period 0 and 1.
- Sampling is not random across time; in particular a plant that is sampled in period 0 might have a lower probability of being sampled again in period 1.

This latter effect would considerably reinforce the bias discussed above.

3.2 Estimating weights from the register

A procedure to account for all 3 effects simultaneously – when the survey register is available – is to

1. Stratify the population not only according to the strata used for sampling by the statistical agency but also according to whether a plant is continuing, exiting or entering.
2. For each stratum calculate the share of plants which are part of S^{dec}. This means for the continuing plants in particular that they need to be sampled in $t = 0$ and $t = 1$.
3. Use the inverse of these shares as weights w_i. Note that strictly speaking we have to multiply each weight by the probability of a numeraire stratum to be in the sample. However if the population includes a stratum which is sampled with probability one this can implicitly serve as numeraire stratum.

3.3 Using a priori information on sampling weights

The estimation of weights from register data often creates implausibly high values; e.g. values of 1000 or more[13] . The reasons for this are not entirely clear but it could happen for example if the number of eventually sampled

[13]This problem occurs even if we are not concerned with correcting for the continuer's bias

plants is exceptionally low for reasons such as non response. Also if the register information is not very accurate and huge numbers[14] of terminated businesses are not deleted from the register this could happen. Typically this happens in very small sector-size band-region cells. Therefore, if the outlier values are left in the datasets it means that very unimportant small sectors could suddenly start to dominate all results. Dropping outliers on the other hand is an unsatisfactory procedure as well. Where should the line be drawn as to what constitutes an outlier? To avoid the issue altogether researchers increasingly rely on weights which are derived from sampling probabilities as reported by the statistical offices[15] In this section I discuss a way to account for the correction bias which relies on a combination of using published sampling probabilities with estimates derived from the sample; i.e. not the register. The procedure is as follows

- For exiters and entrants use weights based on inverse sampling probabilities as reported in table 4[16]; i.e.

$$w_i = \begin{cases} \frac{1}{P\{i \in S_0\}} \text{ if } i \in X \\ \frac{1}{P\{i \in S_1\}} \text{ if } i \in N \end{cases} \tag{19}$$

- For continuers the challenge is to find an estimate of the re-sampling probability $P\{i \in S_1 | i \in S_0\}$. I obtain such an estimator by simply looking at the fraction of plants sampled in period 0 that are also sampled in period 1

$$\hat{P}\{i \in S_1 | i \in S_0\} = \frac{\sum_i \mathbb{I}\{i \in S_0 \cap i \in S_1\}}{\sum_i \mathbb{I}\{i \in S_0\}} \tag{20}$$

  where $\mathbb{I}\{\cdot\}$ is an indicator function equal to one if the condition in braces is true. Weights are then obtained by inverting the product of the

[14] relative to the sampled numbers

[15] This is not without problems either because the information on sampling procedures is sometimes fairly incomplete. For example while table 4 reports typical sampling probabilities across the years there is an issue as to how these varied exactly across sectors and how re-sampling is handled; i.e. conditional on having been sampled in a given year the sampling probability for a plant is presumably lower than the values indicated in table 4.

[16] For entrants we have to use the weights relevant in period 1, for exiters the weights from period 0.

reported sampling weights in period 0 with the estimated conditional re-sampling probability; i.e.

$$w_i = \frac{1}{\hat{P}\{i \in S_1 | i \in S_0\} P\{i \in S_0\}} \tag{21}$$

In the empirical section I will report two types of weights calculated in this fashion

- Method 0 computes the estimates of the re-sampling probability for each year-size band cell separately.
- Method 1 assumes a uniform re-sampling probability across size bands in each year.

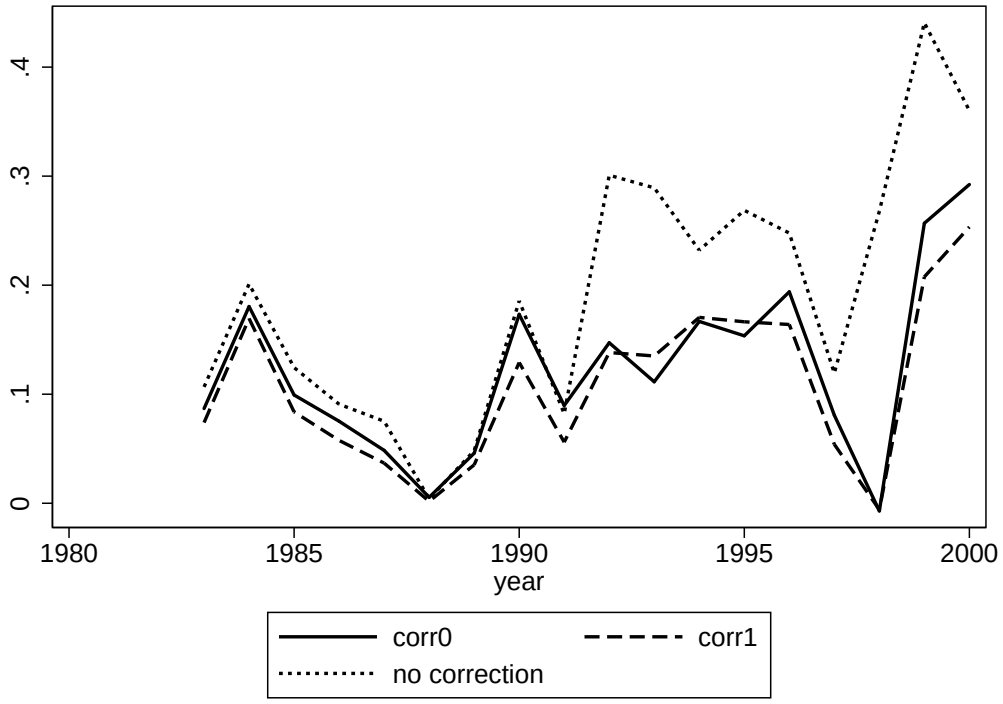
3.4 Further discussion

Other studies have used sampling weights before. However, when the spurious under-sampling of continuers is taken into account simply weighting observations by their inverse sampling probabilities might make matters worse and not doing any weighting at all might well turn out to be a kind of second best solution. Why is that? Suppose the following assumptions are correct: the smaller firms which usually get higher sampling weights are concentrated more in the group of exiters and entrants than in the group of continuing plants. Hence by using sampling weights we implicitly raise the contribution of entry and exit. However, because of the continuers bias, the group of exiters and entrants is already over-represented in the usable sample. Or put differently: if we just used the sample without weights the continuers bias and the bias because of undersampling of smaller plants might just cancel because they work in opposite directions. A little example makes this point clear. Assume that we have two groups of plants: large and small plants. Large plants are sampled with probability 0.5, small ones with probability 0.25 in every year. Further suppose that all large plants never exit whereas all small plants always exit and all entry occurs in the group of small plants as well. Now, in order to compute an unbiased estimate of the productivity growth decomposition we would have to use a weighting factor of $4 = \frac{1}{0.5^2}$ for

large continuing plants and an weighting factor of equally $4 = \frac{1}{0.25}$ for small plants; i.e. the relative weight is equal to 1 which means we no not need to weight at all. Thus not weighting at all is equivalent to the first best weights and using "naive weights" implies a relative weight of 2 between small and large plants which is wrong.

4 Empirical implementation

Figure 5: Correcting for the continuer bias



Source: Author's calculations based on ARD.
Notes: Median net entry share across 2 digit sectors. The "no correction" series is the same as in figure 1. "corr0" and "corr1" implement the correction methods described in section 3.3

The way to correct for the continuer's bias is thus the computation of modified sampling weights as described earlier. Table 1 reports descriptive statistics on the weights I obtained for the 1980 to 2000 period using the two

Table 1: Inverse sampling weights with continuer bias correction
(Descriptive statistics for manufacturing)

	method 0			method 1		
year	mean	10th perc.	90th perc.	mean	10th perc.	90th perc.
1984	3.205	1.000	7.262	2.600	1.000	4.948
1985	5.134	1.206	11.980	3.327	1.492	5.967
1986	5.910	1.261	13.960	3.611	1.620	6.481
1987	3.751	1.490	7.432	2.622	1.857	3.713
1988	6.574	1.248	15.129	3.783	1.637	6.549
1989	3.860	1.000	8.187	3.052	1.000	5.448
1990	6.469	1.272	15.174	3.744	1.633	6.532
1991	5.681	1.202	12.760	3.518	1.510	6.040
1992	2.875	1.000	6.319	2.170	1.000	3.237
1993	5.693	1.000	15.447	3.293	1.000	5.937
1994	6.828	1.160	18.843	3.503	1.579	6.317
1995	8.783	1.392	22.133	4.252	1.941	7.762
1996	7.193	1.251	16.927	3.989	1.757	7.029
1997	7.830	1.366	19.283	4.182	1.887	7.550
1998	12.148	1.333	30.907	5.517	1.333	10.372
1999	10.899	1.333	27.134	5.322	1.333	9.828
2000	8.859	1.333	21.154	5.217	1.333	9.609

Source: Author's calculation based on ARD.

Notes: Mean, 10th and 90th percentile of the inverse sampling weights from random sampling incorporating the correction method for the continuer's bias described in section 3.1. Method 0 estimates re-sampling probabilities for each size band-year cell separately. Method 1 assumes uniform re-sampling weights across size bands.

Table 2: Regressions of net entry contribution

	no correct	correct0	correct1	diff0	diff1
Post90	0.169	0.075	0.076	0.021	0.041
	(0.029)***	(0.022)***	(0.022)***	(0.007)***	(0.009)***
constant	0.093	0.082	0.064	0.011	0.023
	(0.022)***	(0.016)***	(0.017)***	(0.005)**	(0.006)***
obs	341	341	341	341	341

Notes: Regressions are for 2 digit sectors and for all 3 intervals from 1980 to 2000. Post90 is a dummy equal to one for years after 1990.

Table 3: Regressions of net entry contribution
(Looking at import penetration and ICT)

	no correct	correct0	correct1
IMP	0.337	0.084	-0.022
	(0.306)	(0.158)	(0.096)
ICT	2.755	1.497	1.370
	(1.535)*	(0.766)**	(0.708)*
constant	0.046	0.214	-0.026
	(0.143)	(0.113)*	(0.063)
obs	323	323	323

Notes: IMP is import penetration computed as the share of domestic sales in a sector; i.e. $IMP = \frac{Imports}{DomesticOutput-Exports+Imports}$. Data source for the sectoral import, export and output data is the OECD STAN database. ICT is the share of ICT capital in total capital stock. Data source is NIESR sectoral productivity dataset available at http://www.niesr.ac.uk/research/nisec.htm

methods described in section 3.3. The table shows the mean, the 10th and the 90th percentile of the weights for the whole manufacturing sector over the 1980 to 2000 period. Columns 1 to 3 show these statistics for the weight calculation method 0 and columns 4 to 6 for method 1[17]. Overall the values seem to be within reasonable bounds. A couple of things are worth pointing out. Firstly, the weights obtained with method 1 are generally lower. This is driven by the fact that there a couple of smaller cells with lower re-sampling ratio which only shows if we have a finer differentiation into size bands in method 0. Secondly, over time and in particular in the 1990s the weights increased which is in line with the increase in random sampling discussed earlier.

Consider next figure 5 which shows the median net entry contribution across 2 digit sectors from figure 1 along with a re-computation of the same series using the corrected sampling weights according to methods 0 and 1, respectively. While for all years the resulting series suggest lower net entry contribution shares the difference is particularly marked for the 1990s. The difference between the series derived from method 0 and 1 is not uniform but does not seem to be substantial in any year.

Somewhat more statistical evidence on the impact of the correction is provided in table 2. It shows median regressions across 2 digit sectors and for all three year intervalls from 1980 to 2000 of the net entry contribution to productivity growth on a dummy which is equal to one for the years after 1990. In column 1 I use the uncorrected net entry series. I obtain a significant coefficient for the 1990s dummy suggesting that the median net entry share increased by about 16 percentage points. Columns 2 and 3 repeat this regression using net entry series corrected for the continuer's bias with method 0 and method 1, respectively. While the 1990s coefficients remain significant they drop to less than half to about 7.5 percent. Also note that the constant which represents the pre 1990s net entry share reduces by 1 and 3 percentage points, respectively, compared to the uncorrected series. Overall the corrections suggest that in the 1990s the net entry share was between 14 and 16 rather than 26 percent. This is a substantial difference. To make that point even clearer the last two columns of table 2 report a regression

[17] see section 3.3

of the difference between the un-corrected and corrected series on the 1990s dummy. Significant and positive values for both the 1990s dummy and the constant suggest that the impact of the correction is statistically relevant in the 1980s but even more in the 1990s.

How does the continuer's bias correction interact with the impact of ICT and globalisation? This is the topic of table 4 which reports regressions of the net entry share on the ICT intensity and the import penetration of a sector.[18] All regressions also include 2 digit sector dummies. Again column 1 reports regressions for the un-corrected, columns 2 and 3 for the corrected series. In all three columns the coefficients for ICT intensity are significantly positive whereas the coefficients on import penetration are not. This is in line with the findings of Criscuolo et al. (2004). Compared to the results for the un-corrected series in column 1 the point estimates for the ICT coefficient are substantially lower in columns 1 and 2, roughly 1.4 as opposed to 2.7. Interestingly, however, both regressions using the corrected series indicate a more significant relationship between the two variables. This suggests that the correction accounted for some noise in the data and brings out genuine structural relationships more clearly.

5 Conclusion

This paper draws attention to a systematic bias in productivity growth decomposition calculations that occurs when these calculations are done using plant or firm level panel data that is subject to random survey sampling. I discuss various ways to correct for the bias. Using data from the UK Annual Respondents Database I show that the impact of the correction is dramatic and statistically significant. In particular in the 1990s – a period when a larger fraction of plants in the population became subject to random sampling – I end up with 10 percentage points lower estimates of the net entry contribution to productivity growth. This could potentially be a blow for some recent studies which tried to link the dramatic increase in net entry contribution to structural changes such as the emergence of ICT technology

[18] For details on the variables see the table notes and Criscuolo et al. (2004).

and increase globalisation. It turns out however that the contrary is true: The positive and significant relationship between ICT and net entry share becomes even more significant if the the corrected net entry share series are used.

References

Bartelsman, E. J. and Dhrymes, P. J. (1998). Productivity dynamics: Us manufacturing plants 1972-1986. *Journal of Productivity Analysis*, 9:5–34.

Criscuolo, C., Haskel, J., and Martin, R. (2003). Building the evidence base for productivity policy using business data linking. *Economic Trends*, 600:39–51.

Criscuolo, C., Haskel, J., and Martin, R. (2004). Import competition, productivity and restructuring in uk manufacturing. *Oxford Review of Economic Policy forthcoming*.

Foster, L., Haltiwanger, J., and Krizan, C. (1998). Aggregate productivity growth: lessons from microeconomic evidence. *NBER Working Paper*, (6803).

A ARD Sampling rules

Table 4: Sampling in ARD source data, 1970-2000

Census year	Employment size band	Sampling fraction	Comments
1970-1971	< 25	0 (exempt)	In some industries, < 11
	25 or more	All	In some industries 11 was lower limit.
1972-1977	< 20	0 (exempt)	
	20 or more	All	
1978-1979	< 20	0 (exempt)*	All industries
	20-49	0.5	In 68 industries
	50 or more	All	In 68 industries
	20 or more	All	In all other industries
1980-1983	< 20	0 (exempt)	All industries
	20-49	0.25	In most industries
	50-99	0.5	In most industries
	100 or more	All	All industries
1984	< 20	0 (exempt)	All industries
	20-49	0.5	England only
	50 or more	All	20 or more outside England
1985-1988	< 20	0 (exempt)	All industries
	20-49	0.25	In most industries
	50-99	0.5	In most industries
	100 or more	All	All industries
1989	< 20	0 (exempt)	All industries
	20-49	0.5	England only
	50 or more	All	20 or more outside England
1990-1994	< 20	0 (exempt)	All industries
	20-49	0.25**	In most industries
	50-99	0.5	In most industries
	100 or more	All	All industries
1995-1997	< 10	0.2	
	10-49	0.25	
	50-99	0.5	
	100-199	0.75	
	200 or more	All	50% of industries, others with smaller thresholds
1998-2000	< 10	0.25	100% rotation
	10-49	0.5	50% rotation
	50-249	0.75	25% rotation
	250 or more	All	

Source: Oulton (1997) and Author's updates Note: For 1997 and earlier years these are sampling frames for ACOP. From 1998 onwards they refer to ABI. * In 1978 a small sample of establishments employing less than 20 was also drawn. ** 0.2 in 1993.

Table 4 summarises the changing survey sampling rules in the surveys providing data to the ARD. Column 3 reports the sampling probabilities for various size bands in various years. A major change in the 1990s has been

the succesive increase of the threshold for random sampling from plants with less than 100 employees to plants with less than 250 employees.

B The FHK decomposition

FHK start from the observation that many productivity measures can be written as follows:

$$y_t = \sum_i \theta_{it} y_{it} \tag{22}$$

where y_t represents the aggregate productivity measure y_{it} the same measure at the plant level and θ_{it} a weighting factor which in the case of labour productivity corresponds to the labour share of plant i and in the case of TFP is a more complex index of all production factors. The difference in aggregate productivity between a base year 0 and an end year 1 can then be written as

$$\begin{aligned} y_1 - y_0 \;=\; & \textstyle\sum_{i\in C} \theta_{i1} y_{i1} - \theta_{i0} y_{i0} && \} \Delta C \\ & + \textstyle\sum_{i\in E} \theta_{i1} y_{i1} && \} \Delta E \\ & - \textstyle\sum_{i\in N} \theta_{i0} y_{i0} && \} \Delta N \end{aligned} \tag{23}$$

where C, E and N are the sets of continuing, exiting and entering plants, respectively. Consider first the elements of the first sum over continuing plants. We can write

$$\begin{aligned} \theta_{i1} y_{i1} - \theta_{i0} y_{i0} &= \theta_{i1}(y_{i1} - y_{i0}) - \theta_{i0} y_{i1} + \theta_{i1} y_{i1} \\ &= \theta_{i1}(y_{i1} - y_{i0}) + (\theta_{i1} - \theta_{i0}) y_{i1} - (\theta_{i1} - \theta_{i0}) y_{i0} + (\theta_{i1} - \theta_{i0}) y_{i0} \\ &= \theta_{i0} \Delta y_i + \Delta \theta_i y_{i0} + \Delta \theta_i \Delta y_i \end{aligned} \tag{24}$$

Next note that

$$\left(\sum_E \theta_{i0} + \sum_C \theta_{i0} \right) - \left(\sum_N \theta_{i1} + \sum_C \theta_{i1} \right) = 0 \tag{25}$$

because the weights have to add up to one in each period. Consequently we can add the following expression to the RHS of 23:

$$\sum_E \theta_{i0} y0 - \sum_C (\theta_{i0} - \theta_{i0}) y_0 - \sum_N \theta_{i1} y_0 \tag{26}$$

In combination with the result in 24 we can thus write

$$\Delta y = \sum_{C} [\theta_{i0}\Delta y_i + \Delta\theta_i(y_{i0} - y_0) + \Delta\theta_i\Delta y_i] + \sum_{N} \theta_{i1}(y_{i1} - y_1) + \sum_{E} \theta_{i0}(y_{i0} - y_0) \tag{27}$$

CENTRE FOR ECONOMIC PERFORMANCE
Recent Discussion Papers

652	Jörn-Steffen Pischke	Labor Market Institutions, Wages and Investment
651	Anthony J. Venables	Evaluating Urban Transport Improvements: Cost Benefit Analysis in the Presence of Agglomeration and Income Taxation
650	John Van Reenen	Is There a Market for Work Group Servers? Evaluating Market Level Demand Elasticities Using Micro and Macro Models
649	Rachel Griffith Stephen Redding Helen Simpson	Foreign Ownership and Productivity: New Evidence from the Service Sector and the R&D Lab
648	Fredrik Andersson Simon Burgess Julia I. Lane	Cities, Matching and the Productivity Gains of Agglomeration
647	Richard B. Freeman Douglas Kruse Joseph Blasi	Monitoring Colleagues at Work: Profit-Sharing, Employee Ownership, Broad-Based Stock Options and Workplace Performance in the United States
646	Alberto Bayo-Moriones Jose E. Galdon-Sanchez Maia Güell	Is Seniority-Based Pay Used as a Motivation Device? Evidence from Plant Level Data
645	Stephen Machin Olivier Marie	Crime and Benefit Sanctions
644	Richard B. Freeman	Are European Labor Markets As Awful As All That?
643	Andrew B. Bernard Stephen Redding Peter K. Schott	Comparative Advantage and Heterogeneous Firms
642	Patricia Rice Anthony J. Venables	Spatial Determinants of Productivity: Analysis for the Regions of Great Britain
641	Kwok Tong Soo	Zipf's Law for Cities: A Cross Country Investigation

640	Alan Manning	We Can Work it Out: the Impact of Technological Change on the Demand for Low Skill Workers
639	Bianca De Paoli	Monetary Policy and Welfare in a Small Open Economy
638	Kimberly Ann Elliott Richard B. Freeman	White Hats or Don Quixotes? Human Rights Vigilantes in the Global Economy
637	Barbara Petrongolo	Gender Segregation in Employment Contracts
636	Ann Bartel Richard B. Freeman Casey Ichniowski Morris Kleiner	Can a Work Organization Have an Attitude Problem? The Impact of Workplaces on Employee Attitudes and Economic Outcomes
635	Paul Gregg Rosanna Scutella Jonathan Wadsworth	Reconciling Workless Measures at the Individual and Household Level: Theory and Evidence from the United States, Britain, Germany, Spain and Australia
634	Stephen Nickell	Employment and Taxes
633	Fabiano Schivardi Roberto Torrini	Threshold Effects and Firm Size: the Case of Firing Costs
632	Paul Gregg Jonathan Wadsworth	Two Sides to Every Story: Measuring the Polarisation of Work
631	Jeremy Grant Thomas Kirchmaier	Corporate Ownership Structure and Performance in Europe
630	Neal Knight-Turvey Andrew Neal Michael A. West Jeremy Dawson	The Impact of an Innovative Human Resource Function on Firm Performance: the Moderating Role of Financing Strategy
629	Nicholas Oulton	A Statistical Framework for the Analysis of Productivity and Sustainable Development
628	Vasileios Gkionakis	Short Job Tenures and Firing Taxes in the Search Theory of Unemployment

The Centre for Economic Performance Publications Unit
Tel 020 7955 7673 Fax 020 7955 7595 Email info@cep.lse.ac.uk
Web site http://cep.lse.ac.uk