# Who works for whom and the UK gender pay gap

## Online Appendix

Sarah Louise Jewell Giovanni Razzu Carl Singleton

*Corresponding author: c.a.singleton@reading.ac.uk: Department of Economics, University of Reading, UK.*

Published at: *British Journal of Industrial Relations.* (2019; In Press)
https://doi.org/10.1111/bjir.12497

## Appendix A. Further description of the data and sample construction

This section provides additional details regarding the datasets used and how we have constructed sub-samples and derived variables thereof. The relevant documentation and variable descriptions attached to these datasets are publicly available from the UK Data Service. The ONS has also published various documents concerning the data quality and consistency of the ASHE.

We focus on methodological details in the period 2002-16. Throughout these years, the ASHE is intended to be a true random sample of all employees in employment, irrespective of employment status, occupation, size of employer etc. Given the legal obligation of employers to respond using payrolls, the ASHE has always had a consistently high response rate of around 55%. The response rate for employee jobs is around 60%, implying that for those firms who do respond they tend to provide complete responses for all their employees in the sampling frame. There is no cumulative attrition from the panel. Any individual in the sample not included in the ASHE, in any year, for whatever reason, remains in the sampling frame for the following year. Conditional on a hundred percent response to the survey, the ASHE would be a truly random one percent sample of employees: all with a National Insurance number which has a numerical part ending in 14. However, there are three major sources of under-sampling, one due to firms not responding, and the other two both occurring if individuals do not have a current tax record. This can happen for some individuals who have recently moved job, or for those who earn very little (mostly working part-time), and who are therefore not paying income tax or National Insurance when their employers are looked up by the ONS. For either of these reasons, a worker would not have (yet) been assigned to an administrative employer reference, known in the UK as a PAYE number (Pay As You Earn). Therefore, the statistical authority would not be able to find an employer address to send the survey questionnaire to. From 2004 the ASHE aimed to increasingly sample some of these employees previously under-represented. It added supplementary data for individuals without a PAYE reference, and attempted to represent the employees whose jobs would have changed between the determination of the sampling frame each January and the reference period in April. Nevertheless, the ASHE datasets remain representative of the employee payments and hours worked in the UK. We view the dataset as providing on average an approximate one percent sample of the employees within all UK firms, as a repeated snapshot every April.

From 2005, a new survey questionnaire was introduced for the ASHE, which was intended to reduce the latitude for employers' own interpretations of what was being asked of them. From 2007 there were further notable methodology changes. Before, occupations were classified as follows: if the respondent stated an employee's job had not changed in the past year, then the previous year's occupational classification was applied - otherwise, it was manually coded. Afterwards an automatic coding, text recognition, tool was used. "The effect of using ACTR was to code more jobs into higher paying occupations. The jobs that tended to be recoded into these higher paying occupations generally had lower levels of pay than the jobs already coded to those occupations. Conversely, they tended to have higher levels of pay than the other jobs in the occupations that they were recoded out of. The impact of this was to lower the average pay of both the occupation group that they had moved from and that they had moved to." In 2007 and 2008 the target sample size of the ASHE was reduced by 20 percent, with reductions targeted at those industries exhibiting the least variation over time in earnings patterns. However, the 1% of population sampling frame was reinstated from 2009 onward.

We use the ASHE annual cross-sections for each year from 2002 to 2016 and construct a panel as follows: in case of multiple jobs per individual, we exclude non-main jobs. In case of missing main job

markers, we impute these based on the job with the highest working hours. In a next step, we link employees across consecutive years based on their unique personal identifiers. We can also impute missing enterprise reference numbers (entrefs) backwards, since the ASHE contains a variable which indicates whether an employee is holding the same job as in the last reference period. After linking two consecutive years in this way, we use local unit identifiers to impute missing entrefs across individuals within the same year (the ONS states that the local unit identifiers are not consistent across years, but rather they are created to identify establishments within years). We continue to update missing entrefs in this way back to and including 2002. While for the years 2003-16 we are only imputing values for a couple of missing entrefs per year, in 2002 we add a large proportion. We could also impute entrefs in this way for 1998-2001, but the sample then becomes increasingly unrepresentative of the UK employee population.

We only keep observations for individuals aged 25-64 in each period, and which have not been marked as having incurred a loss of pay in the reference period through absence, employment starting in the period, or short-time working, and which are marked as being at an adult rate of pay (i.e. dropping trainees and apprenticeships). This is practically the same filter applied by the ONS in their annually published results on UK "Patterns of Pay" using the ASHE. We drop observations with missing records for basic hours, gross weekly earnings, or hourly wage rates. Basic hours are intended by the survey to be a record for an employee in a normal week, excluding overtime and meal breaks. Gross weekly pay is the main recorded value in the survey, and from this overtime records are subtracted. However, all other payments received in the period are included within this gross pay, including incentive-related pay and any premiums for weekend or night work. Hourly wage rates are then derived by the ONS through dividing by basic hours worked. We drop observations with over a hundred or less than one basic hour worked, as these could reflect measurement error and the mistaken inclusion of overtime in the usual hours of work. Full-time is defined as working over thirty basic hours in a week. But there are a tiny number of discrepancies in some years, we believe relating to teaching contracts, where the definition applied by the ONS differs. We however recode these such that for all observations the thirty hours threshold applies. To further address some potential for measurement error, we drop observations whose derived hourly rate of pay, excluding overtime, is less than eighty percent of the applicable National Minimum Wage (NMW) each April, allowing for the different age-dependent rates of the NMW over time (which, in this application, is always the highest adult rate, given that we restrict our attention to individuals at least 25 years old). We set this threshold lower than the NMW to avoid dropping observations where employers have rounded pay figures about the NMW, where the degree of rounding could vary with the actual value of the NMW, a behaviour on the part of employers which has been hypothesised by the ONS.

To create a tenure variable, we use the recorded employment start date of individuals. The ASHE contains information on when an employee started working for an enterprise from 2002 onwards. We drop a tiny number of unrealistic entry dates, where the start date lies in the future, or where it implies an employee started working aged fifteen or younger. There are some inconsistencies across years in these records. First, an employee can be employed by the same enterprise for three consecutive years, holding the same job, but the start dates recorded in the first and third years, though identical, can vary from the second. In this case we update the "one-off" deviation with the value of the previous year. Second, if we observe an employee in a chain of consecutive years in the same firm, holding the same job, but the start dates differ for some years, then we impute the earliest date available. Finally, we use the employment start date to impute a tiny number of missing entrefs for employees backwards to 2002 again. This allows us to not have to observe employees in a chain of consecutive years to make imputations. Again, we then use within-year local unit identifiers to update longitudinal entrefs within a year, for a handful of employees with missing entrefs.

The ASHE contains the number of employees in an enterprise as listed in the administrative Inter-Departmental Business Register (IDBR). A small minority of employees in the same enterprise and year have missing or varying values for this variable. We impute the same value for all employees within year and enterprise as the modal value for the firm. For 2002-10 occupations are classified using the four-digit SOC2000, and for 2011-16 using the SOC2010. We use both classifications in our analysis, rather than cross-walking. When we use 2-digit occupations as control variables, the base or excluded category is 41 – Administrative Occupations. When we use 3-digit occupations in our robustness checks, the excluded category is 211 – Science Professionals. To derive a firm's time -invariant industry

classification, we first convert ONS Standard Industrial Classification (SIC) 2007 to 2003, using files made available by the UK Data Service. This conversion uses the 2008 Annual Respondents Dataset, where both classifications were applied, and where any 2007 code mapping to multiple 2003 codes was decided using whichever of the two bore a greater share of economic output. We then take the modal SIC2003 section (one-digit) classification of the firm in the sample. We then group industry sectors as follows: Manufacturing/Construction/Engineering, or just "Manufacturing," is given by SIC2003 sections C-F; Retail/Wholesale/Services or just "Non-financial (sales) services", is given by G-H; 'Financial services" is given by J-K; and Primary/public/other services, or just "Other", is given by A-B, I and L-Q. The Manufacturing sector is the excluded category in the regression models. We assign each firm to the public or private sector using their modal value in the sample of the ASHE variable "idbrsta", which records the legal status of the enterprise according to the IDBR. We assign "Private" to be private companies, sole proprietors and partnerships, with everything else being Public, including central government and local authorities. We derive an individual's birth cohort by taking their modal value of the dataset year minus their age within the sample, and in the regression models the excluded category is the earliest cohort. The excluded year effect is for 2002.

The household survey-based pay statistics in Figure 1 and A1 are derived using the longitudinal British Household Panel Survey (2002-08) and its successor the Understanding Society Survey (2009-16). They use the Great Britain and Northern Ireland samples and waves, but without any other boost samples. Years refer to tax years (April-March). Some individuals (less than 1%) were interviewed twice in these periods, in which case we use their first earnings response. The hourly wage is estimated from responses by employees aged 25-64 for monthly earnings in their main job. It is derived by taking usual monthly pay, converting this to a weekly figure (multiplying by 3/13), and then dividing by the sum of usual normal and usual overtime weekly hours. Only observations with usual weekly hours between 1 and 100 hours were used. Hourly wages below 0.8 of the applicable National Minimum Wage rate were dropped. Any individuals with missing values for pay, hours, age, sex or interview date were excluded from the statistics.

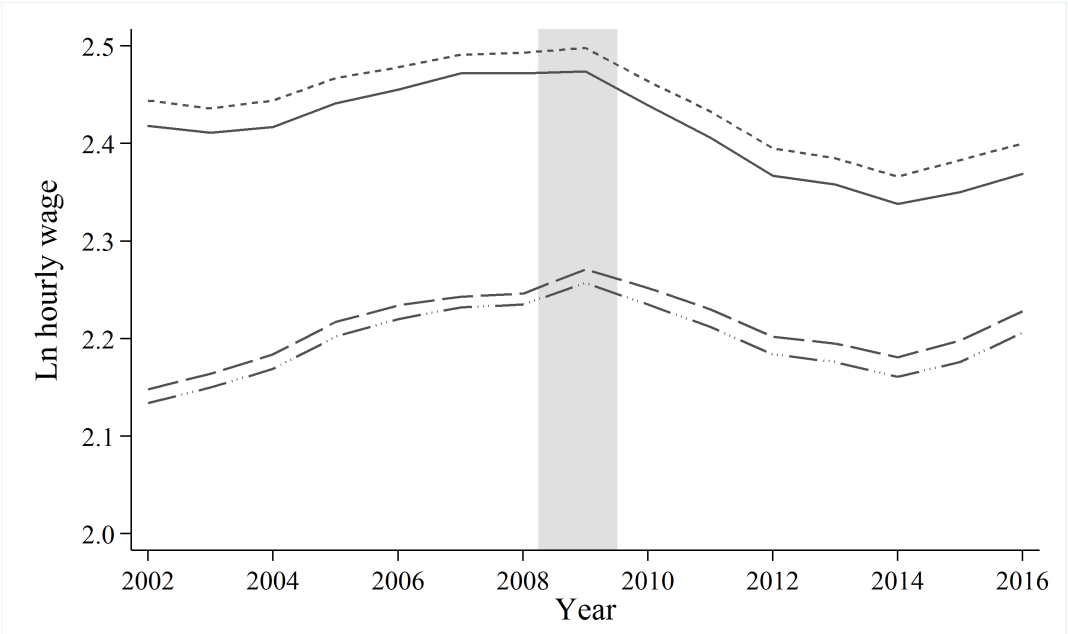
TABLE A1: Distribution of the number of different jobs held by workers in the Analysis sample (%), 2002-2016

| | Worker-year weighted | | | Worker weighted | | |
|---|---|---|---|---|---|---|
| Number of jobs | Male (1) | Female (2) | Total (3) | Male (4) | Female (5) | Total (6) |
| 1 | 40.93 | 40.95 | 40.94 | 48.17 | 48.20 | 48.18 |
| 2 | 32.61 | 32.60 | 32.60 | 30.52 | 30.56 | 30.54 |
| 3 | 16.90 | 17.01 | 16.96 | 14.13 | 14.16 | 14.14 |
| 4 | 6.75 | 6.72 | 6.73 | 5.20 | 5.14 | 5.17 |
| 5 | 2.12 | 2.06 | 2.09 | 1.52 | 1.49 | 1.51 |
| 6 | 0.54 | 0.54 | 0.54 | 0.37 | 0.38 | 0.37 |
| 7+ | 0.15 | 0.12 | 0.13 | 0.10 | 0.08 | 0.09 |
| *N* / *P* | 824,806 | 883,326 | 1,708,132 | 124,501 | 131,808 | 256,304 |

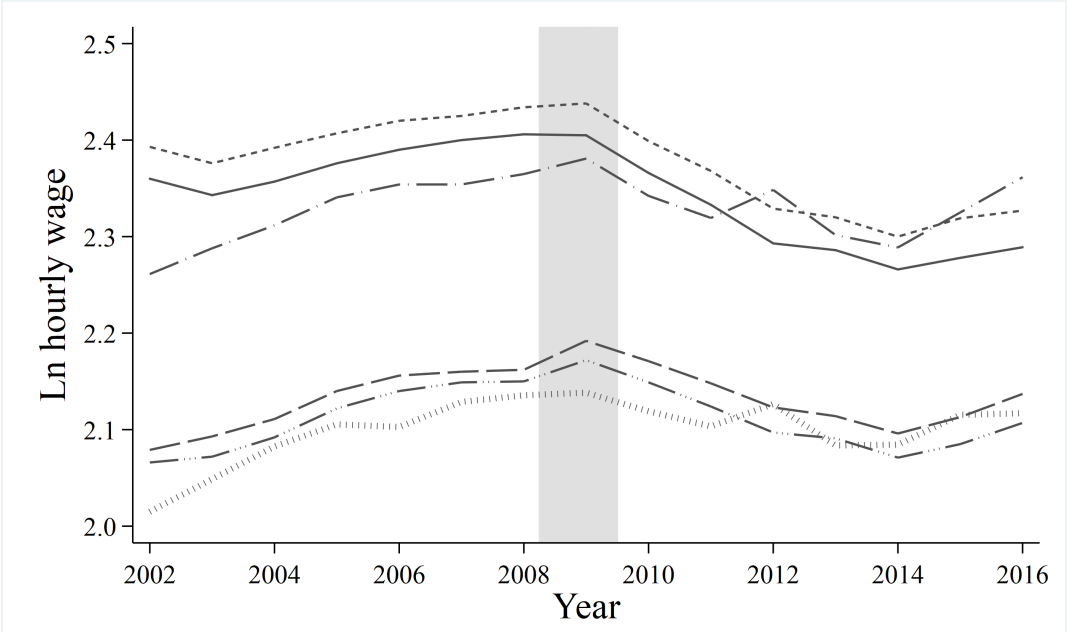
Notes.- author calculations using the ASHE 2002-16, all employees age 25-64.

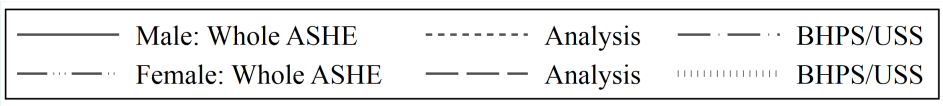
FIGURE A1: Mean and median log real hourly pay of UK men and women, 2002-16

A. Mean




B. Median

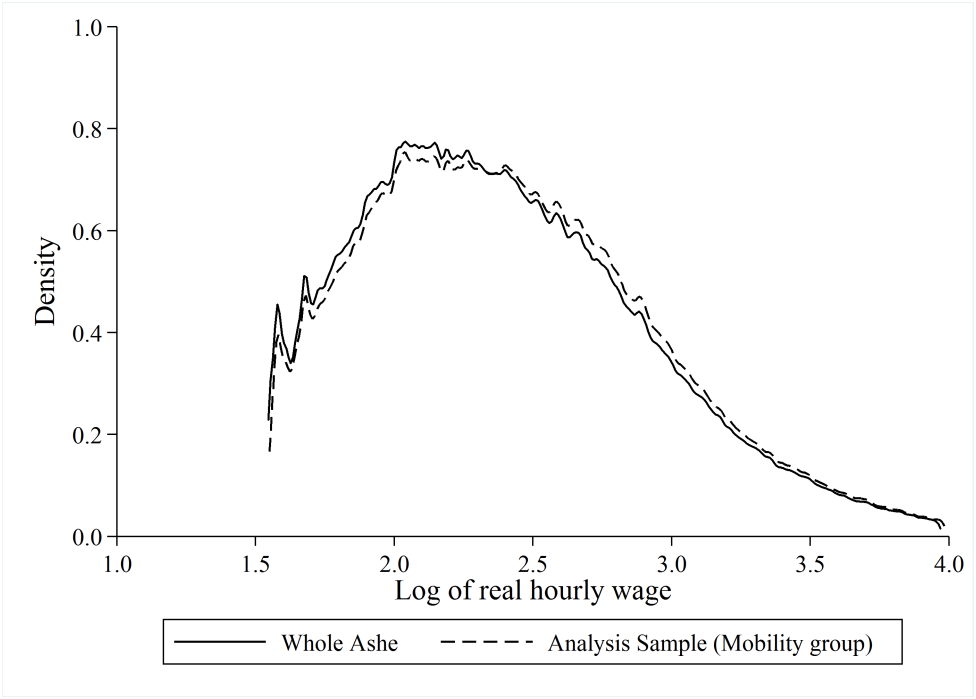


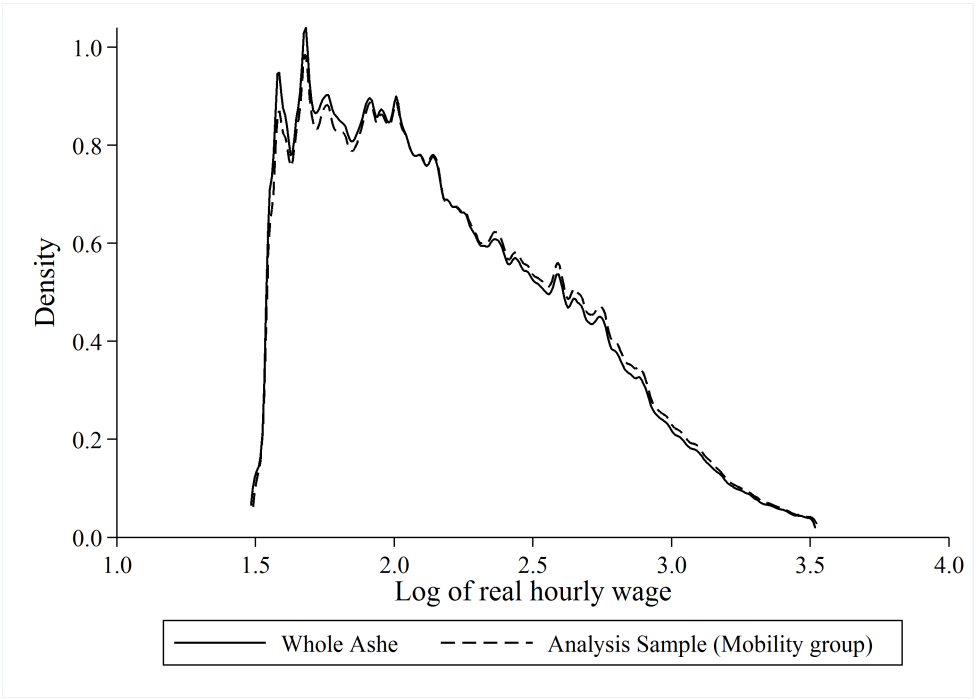

Notes.- see Figure 1.

FIGURE A2: Distributions of male and female real hourly pay: comparison of Whole UK-representative ASHE sample with the Analysis sub-sample used for the main results

A. Male




B. Female



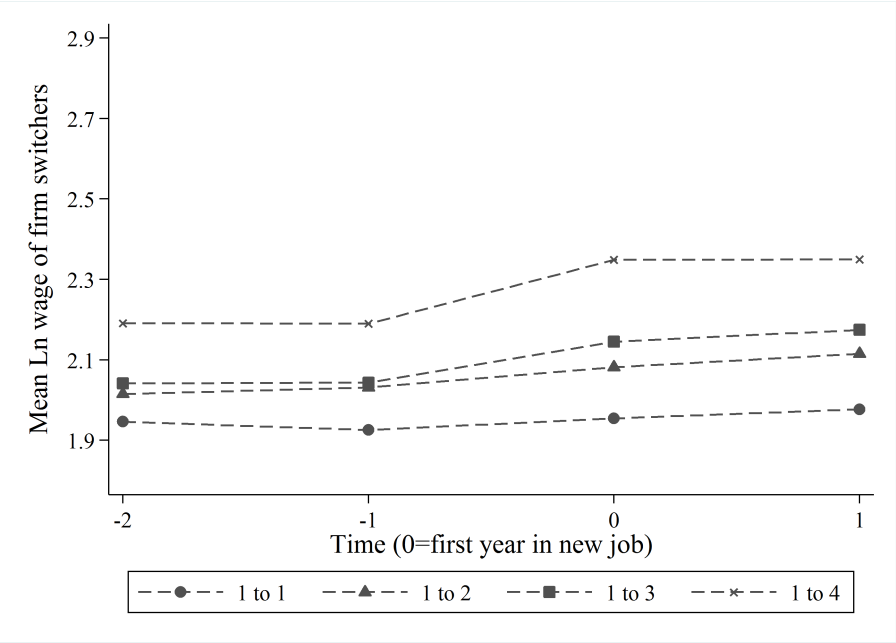

Notes.- see Figures 1 & 2

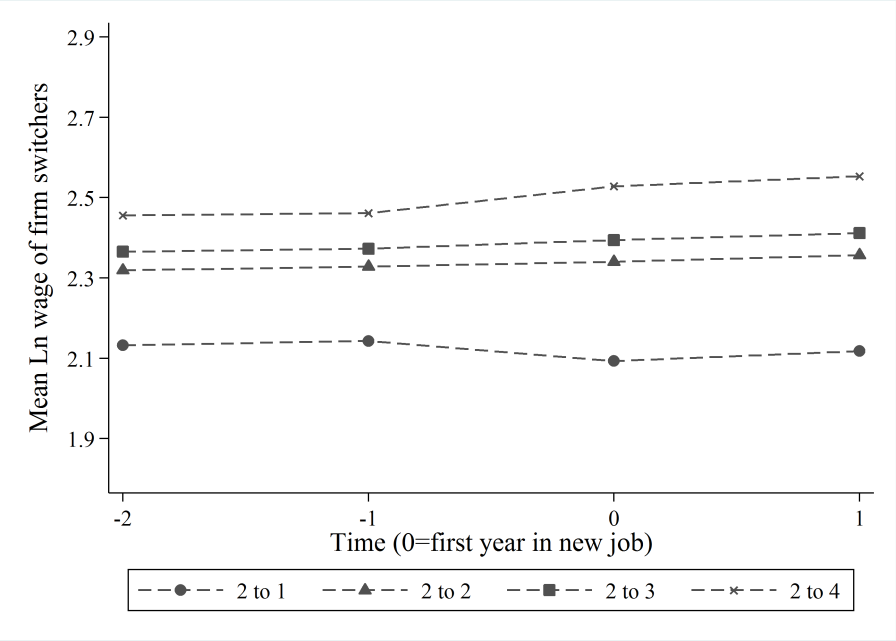
## Appendix B. Robustness of the regression model

A key assumption of the AKM-type models is that the mobility of workers is exogenous to the unobserved time-varying heterogeneity of the worker-firm match: for example, match-specific shocks to wages and productivity, which could induce workers to switch between firms. Card et al. (2013, 2016) and Card et al. (2018) have demonstrated how to carry out a simple and transparent test of this assumption, which we apply to our Analysis sample from the ASHE. This test takes the form of an event study of how wages change when employees switch between firms. If the assumptions of the AKM-type Full model are correct, as described in the main text, that firms pay proportional wage premiums to all of their employees, then we would expect to observe that employees who switch from firms where their co-workers are relatively low-paid (in the economy) would then experience wage increases after moving to firms where their new co-workers were relatively high-paid, and vice versa. The model also predicts that the wage gains and losses for employees moving in the different directions between any two firms should be symmetric. Furthermore, wages before and after switching should be relatively stable, i.e. firm switching is not driven by a deterioration or an expected future growth in match-specific quality.

We construct this event study using the Analysis sample of employees as follows. First, we assign to each employee, in every year, the mean wage among all of his co-workers in the same firm and year (excluding himself). Then, year-by-year, we assign each employee to the quartile of how relatively well or poorly paid on average those co-workers were, compared with all the other employees and their respective co-workers in the economy. Then we select a sample of employee-firm switching events from within the Analysis sample. For a switch to be included in the event study, an employee must be observed working in the same firms in the two years before and after their switch took place. We refer to such employees as 'switchers'. Throughout the whole sample period, this gives us a sample of 21,455 switching events. We then define 16 sub-types of event, defined by the co-worker wage quartiles of the switchers before and after they switched firms. For each of these event types, and the employee-employer relationships represented by each, we then compute the mean log real hourly wages of the switchers, conditional on the number of years before or since switching. These statistics are displayed in Figure B1 and Table B1. For example, in the first panel of the figure below, we plot how, on average, real hourly wages evolved for switchers who were originally in the lowest quartile for average co-worker real hourly wages, before they then switched to a different firm. As predicted by the model, there is a step-change increase in individual employee wages for those who moved to a firm where their new co-workers were then relatively high-paid. Whereas, those who moved between firms where their co-workers were similarly low-paid experienced no substantial wage increases. Similarly, in the last panel below, we can see that among those employees who switched away from firms with co-workers who were relatively well-paid, individuals then on average experienced larger real hourly wage decreases if they switched to firms with relatively low-paid co-workers. In other words, Figure B1 shows that employees do experience step-change wage increases (decreases) when they switch to firms where their new co-workers are more (less) high-paid than their old co-workers. Furthermore, the magnitude of these observed average employee wage changes upon firm switching display some symmetry, as is also predicted by the Full model. Also, note that none of the series in any panel of Figure B1 cross, nor are there noticeably different trends before and after in the average employee wages across the different types of switching (high to low, high to high, low to low etc.).

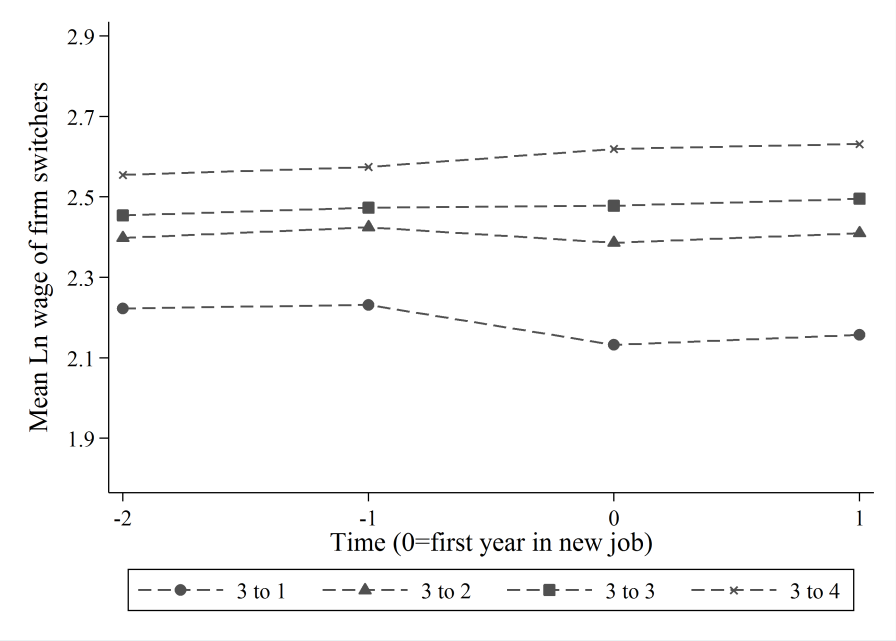
FIGURE B1: An event-study of average employee log real hourly wages before and after switching firms, depending on the quartile of co-worker average wages in the old and new firms

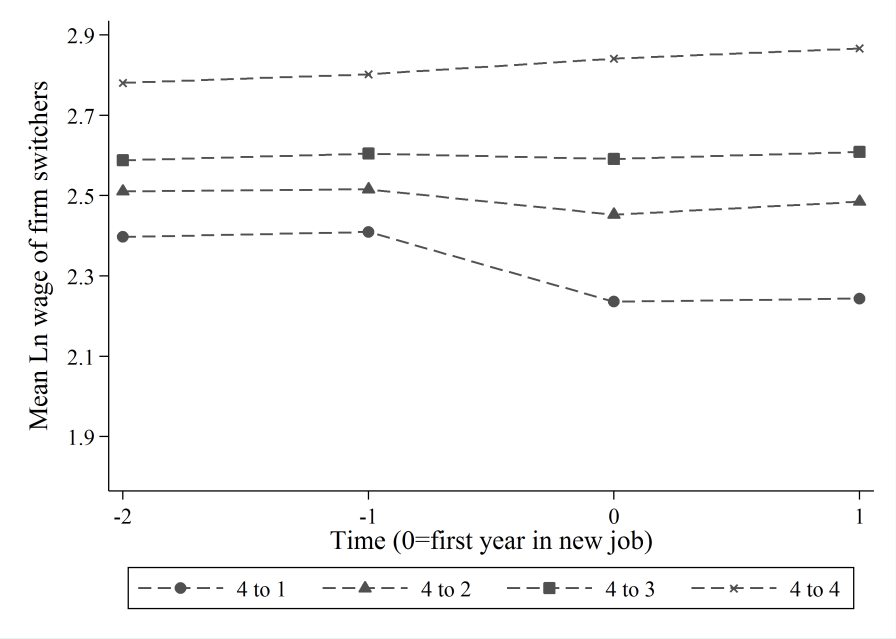



(continued on the next page)

(continued from the previous page)




Notes.- "X to Y" indicates the quartile of co-worker wages for employees in their old firm (X), from which they switch to their new firm (Y). Each event series uses firm switches throughout the ASHE 2002-2016 Analysis sample period. See Table B1 for a summary of this data.

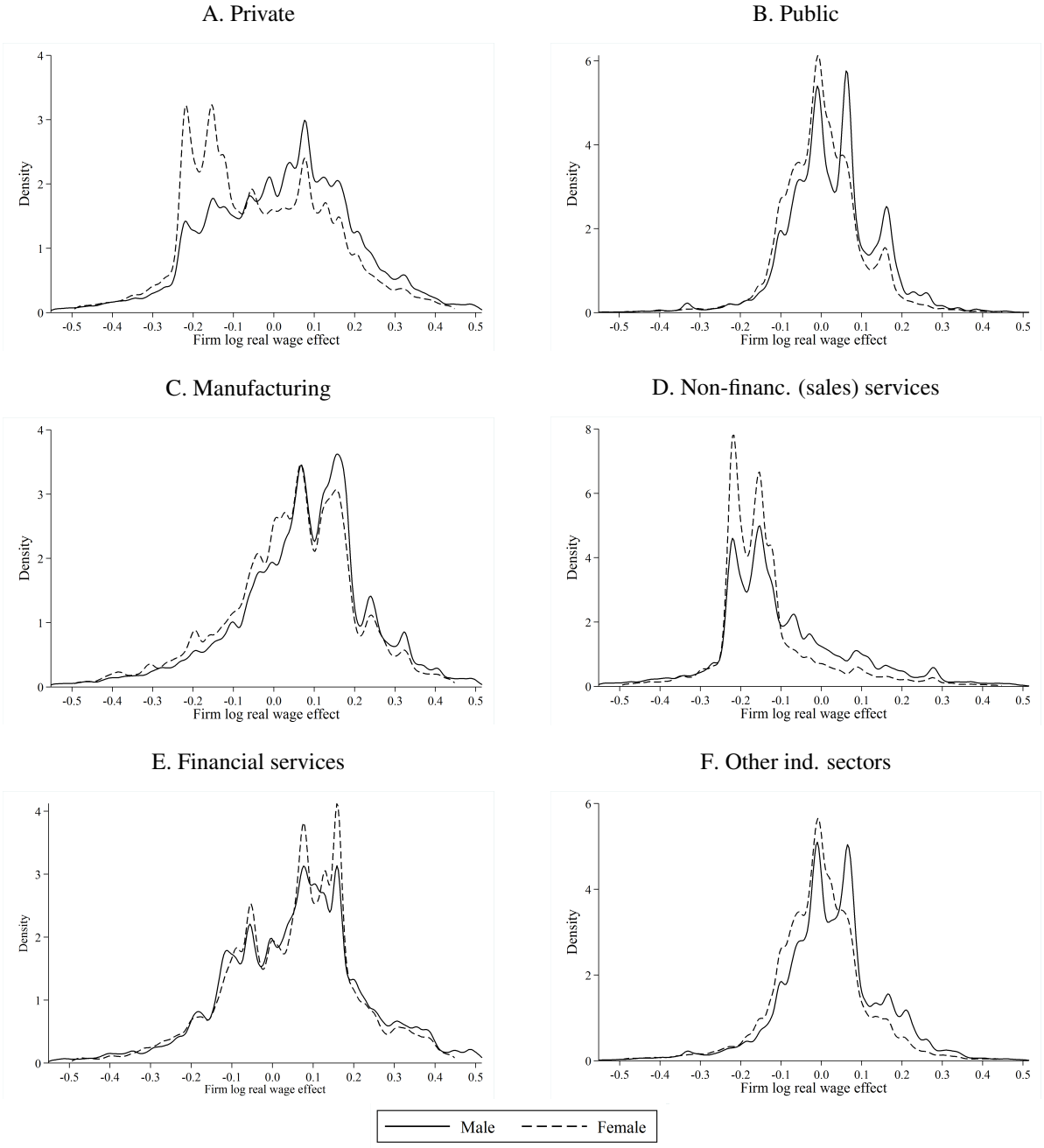
TABLE B1: Summary of event-study: Mean log wages of employees two years before switching firms and two years after, 2002-2016

| Quartile to Quartile | Switches (1) | % of switches (2) | Years since switching -2 (3) | -1 (4) | 0 (5) | 1 (6) | 3-year wage change (7) |
|---|---|---|---|---|---|---|---|
| 1 to 1 | 3,498 | 16.30 | 1.95 | 1.93 | 1.95 | 1.98 | 0.03 |
| 1 to 2 | 1,073 | 5.00 | 2.01 | 2.03 | 2.08 | 2.12 | 0.10 |
| 1 to 3 | 705 | 3.29 | 2.04 | 2.04 | 2.15 | 2.17 | 0.13 |
| 1 to 4 | 458 | 2.13 | 2.19 | 2.19 | 2.35 | 2.35 | 0.16 |
| 2 to 1 | 931 | 4.34 | 2.13 | 2.14 | 2.09 | 2.12 | -0.01 |
| 2 to 2 | 1,582 | 7.37 | 2.32 | 2.33 | 2.34 | 2.36 | 0.04 |
| 2 to 3 | 1,660 | 7.74 | 2.37 | 2.37 | 2.39 | 2.41 | 0.05 |
| 2 to 4 | 768 | 3.5 | 2.46 | 2.46 | 2.53 | 2.55 | 0.10 |
| 3 to 1 | 583 | 2.72 | 2.22 | 2.23 | 2.13 | 2.16 | -0.07 |
| 3 to 2 | 1,018 | 4.74 | 2.40 | 2.42 | 2.39 | 2.41 | 0.01 |
| 3 to 3 | 2,573 | 11.99 | 2.45 | 2.47 | 2.48 | 2.50 | 0.04 |
| 3 to 4 | 1,194 | 5.57 | 2.55 | 2.57 | 2.62 | 2.63 | 0.08 |
| 4 to 1 | 385 | 1.79 | 2.40 | 2.41 | 2.24 | 2.24 | -0.15 |
| 4 to 2 | 571 | 2.66 | 2.51 | 2.52 | 2.45 | 2.48 | -0.03 |
| 4 to 3 | 1,249 | 5.82 | 2.59 | 2.60 | 2.59 | 2.61 | 0.02 |
| 4 to 4 | 3,207 | 14.95 | 2.78 | 2.80 | 2.84 | 2.87 | 0.09 |

Notes.- author calculations using the ASHE 2002-16, all employees age 25-64. £2002. See Figure B1.

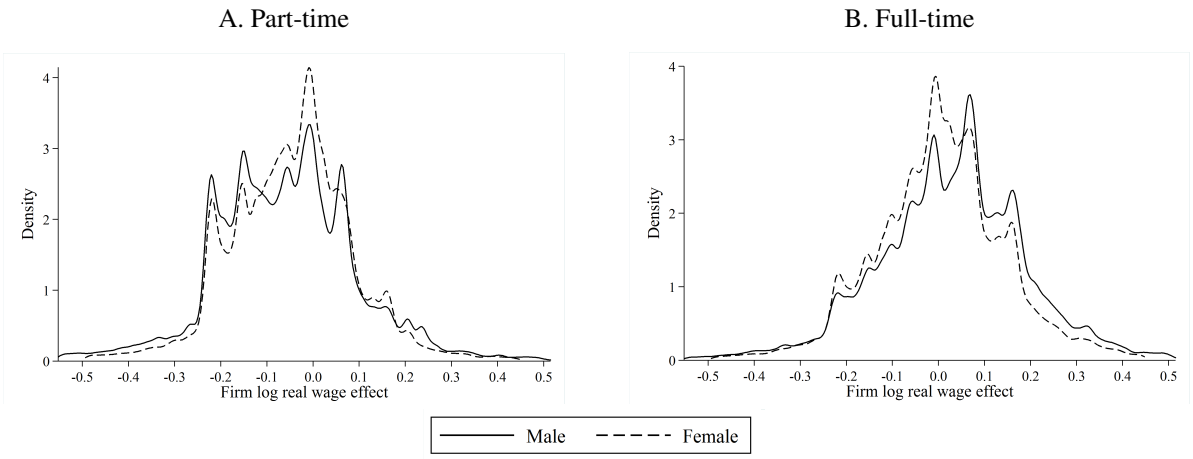
## Appendix C. Additional figures

FIGURE C1: Distribution of estimated firm-specific wage effects ($\widehat{\phi}_{J(it)}$): employees in the private and public sectors, and working within groups of industry sectors



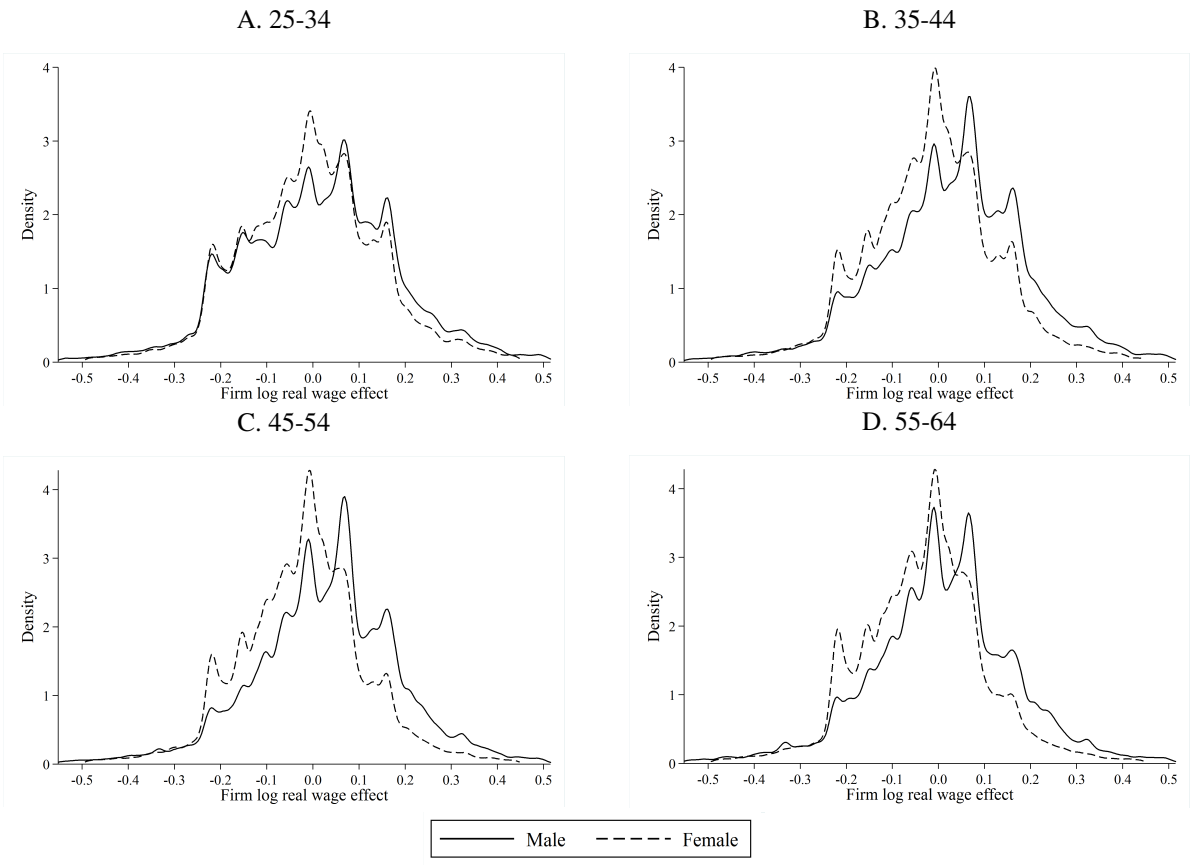

Notes.- see Figure 3. Both male and female kernel densities were estimated with a bandwidth of one log point. The top and bottom one percent of the overall set of estimated firm-specific effects are not displayed in any of the sub-figures. See Table 3 and Appendix A for descriptions of the industry groupings. In (a), the sample size of employee-years is 580,000 male and 430,000 female; in (b) 240,000 and 460,000; in (c) 190,000 and 60,000; in (d) 130,000 and 150,000; in (e) 170,000 and 150,000; in (f) 330,000 and 530,000;

FIGURE C2: Distribution of estimated firm-specific wage effects ($\widehat{\phi}_{J(it)}$): employees in part- and full-time jobs




Notes.- see Figure 3 and C1. Both male and female kernel densities were estimated with a bandwidth of one log point. The top and bottom one percent of the overall set of estimated firm-specific effects are not displayed in any of the sub-figures. In (a), the sample size of employee-years is 70,000 male and 360,000 female; in (b) 760,000 and 520,000.

FIGURE C3: Distribution of estimated firm-specific wage effects ($\widehat{\phi}_{J(it)}$): age groups




Notes.- see Figure 3 and C1. Both male and female kernel densities were estimated with a bandwidth of one log point. The top and bottom one percent of the overall set of estimated firm-specific effects are not displayed in any of the sub-figures. In (a), the sample size of employee-years is 220,000 male and 230,000 female; in (b) 250,000 and 260,000; in (c) 230,000 and 260,000; in (d) 130,000 and 130,000.

## Appendix D. Additional tables

TABLE D1: Summary of estimated Full model with two-way fixed effects & decomposition of raw gender pay gap: gender-specific covariate effects

| | Male (1) | Female (2) | Total (3) |
|---|---|---|---|
| St. dev. of log wages - $std_{it}(w_{it})$ | 0.55 | 0.49 | 0.53 |
| $N$: worker-years | 824,806 | 888,326 | 1,708,132 |
| $P$: workers | 131,903 | 124,501 | 256,404 |
| $F$: firms | | | 86,779 |
| St. dev. worker effects - $std_{it}(\widehat{\alpha}_i)$ | 0.46 | 0.37 | 0.42 |
| St. dev. firm effects - $std_{it}(\widehat{\phi}_{J(it)})$ | 0.20 | 0.17 | 0.18 |
| St. dev. observables - $std_{it}(\mathbf{x}_{it}'\widehat{\boldsymbol{\beta}})$ | 0.53 | 0.46 | 0.51 |
| Correlation - $corr_{it}(\widehat{\alpha}_i,\ \widehat{\phi}_{J(it)})$ | -0.022 | -0.010 | 0.004 |
| Adjusted $R^2$ | | | 0.904 |
| RMSE | | | 0.164 |
| *Variance shares* (X/$var_{it}(w_{it})$: | | | |
| Worker effects - $var_{it}(\widehat{\alpha}_i)$ | 0.69 | 0.57 | 0.64 |
| Firm effects - $var_{it}(\widehat{\phi}_{J(it)})$ | 0.13 | 0.12 | 0.12 |
| Covariance - $2covar_{it}(\widehat{\alpha}_i,\ \widehat{\phi}_{J(it)})$ | -0.01 | -0.01 | 0.00 |
| Residuals - $var_{it}(\hat{\varepsilon}_{it})$ | 0.07 | 0.09 | 0.02 |
| Other | 0.12 | 0.22 | 0.22 |
| *Raw gender wage gap decomp. (shares):* | | | |
| Raw gap - $E_{it}[w_{it}\mid i \in M] - E_{it}[w_{it}\mid i \in F]$ | | | 0.223 |
| Worker -$E_{it}[\widehat{\alpha}_i\mid i \in M] - E_{it}[\widehat{\alpha}_i\mid i \in F]$ | | | 0.174 |
| | | | (0.78) |
| Firm - $E_{it}\left[\widehat{\phi}_{J(it)}\mid i \in M\right] - E_{it}\left[\widehat{\phi}_{J(it)}\mid i \in F\right]$ | | | 0.036 |
| | | | (0.16) |
| Occupations | | | 0.013 |
| | | | (0.06) |
| Observable / other | | | 0.00 |
| | | | (0.00) |

Notes.- author calculations using the ASHE 2002-16, all employees age 25-64. £2002. Pay excludes overtime. Gap is male minus female. Estimated Full model includes gender-specific covariates in $\mathbf{x}_{it}$ for year fixed effects, squared and cubed terms for employee age, a cubic polynomial for employee tenure, a cubic polynomial for firm size (n. of employees) and a dummy variable for whether a worker was employed full-time.

TABLE D2: Summary of estimated Full model with two-way fixed effects & decomposition of raw gender pay gap: weekly earnings

| | Male (1) | Female (2) | Total (3) |
|---|---|---|---|
| St. dev. of log earnings - $std_{it}(w_{it})$ | 0.63 | 0.76 | 0.74 |
| $N$: worker-years | 824,806 | 888,326 | 1,708,132 |
| $P$: workers | 131,903 | 124,501 | 256,404 |
| $F$: firms | | | 86,779 |
| St. dev. worker effects - $std_{it}(\widehat{\alpha}_i)$ | 0.47 | 0.48 | 0.49 |
| St. dev. firm effects - $std_{it}(\widehat{\phi}_{J(it)})$ | 0.26 | 0.27 | 0.27 |
| St. dev. observables - $std_{it}(\mathbf{x}_{it}'\widehat{\boldsymbol{\beta}})$ | 0.60 | 0.72 | 0.71 |
| Correlation - $corr_{it}(\widehat{\alpha}_i,\ \widehat{\phi}_{J(it)})$ | -0.196 | -0.187 | -0.152 |
| Adjusted $R^2$ | | | 0.902 |
| RMSE | | | 0.232 |
| *Variance shares* (X/$var_{it}(w_{it})$: | | | |
| Worker effects - $var_{it}(\widehat{\alpha}_i)$ | 0.56 | 0.40 | 0.44 |
| Firm effects - $var_{it}(\widehat{\phi}_{J(it)})$ | 0.17 | 0.13 | 0.13 |
| Covariance - $2covar_{it}(\widehat{\alpha}_i,\ \widehat{\phi}_{J(it)})$ | -0.12 | -0.08 | -0.07 |
| Residuals - $var_{it}(\hat{\varepsilon}_{it})$ | 0.08 | 0.09 | 0.04 |
| Other | 0.31 | 0.46 | 0.46 |
| *Raw gender wage gap decomp. (shares):* | | | |
| Raw gap - $E_{it}[w_{it}\mid i \in M] - E_{it}[w_{it}\mid i \in F]$ | | | 0.498 |
| Worker -$E_{it}[\widehat{\alpha}_i\mid i \in M] - E_{it}[\widehat{\alpha}_i\mid i \in F]$ | | | 0.245 |
| | | | (0.49) |
| Firm - $E_{it}\left[\widehat{\phi}_{J(it)}\mid i \in M\right] - E_{it}\left[\widehat{\phi}_{J(it)}\mid i \in F\right]$ | | | 0.067 |
| | | | (0.14) |
| Occupations | | | 0.010 |
| | | | (0.02) |
| Other | | | 0.176 |
| | | | (0.35) |

Notes.- see Table 2.