*Journal of the Royal Statistical Society Series A: Statistics in Society*, 2023, **186**, 110–136
https://doi.org/10.1093/jrsssa/qnac003
Advance access publication 31 January 2023
**Original Article**

ROYAL STATISTICAL SOCIETY DATA | EVIDENCE | DECISIONS
Journal of the Statistics Society
**Series A**
Statistics in Society

# Reconciling reports: modelling employment earnings and measurement errors using linked survey and administrative data

**Stephen P. Jenkins**[1] **and Fernando Rios-Avila**[2]

[1]Department of Social Policy, London School of Economics and Political Science, London, UK
[2]The Levy Economics Institute of Bard College, Blithewood, Annandale-on-Hudson, NY, USA

*Address for correspondence:* Stephen P. Jenkins, Department of Social Policy, London School of Economics and Political Science, Houghton Street, London, WC2A 2AE, UK. Email: s.jenkins@lse.ac.uk

## Abstract

We develop and apply new statistical models for linked survey and administrative data on employment earnings, incorporating 4 types of measurement error. In addition, we allow error distributions to differ with individual characteristics, which improves model fit and allows us to investigate substantive hypotheses about factors associated with error bias and variance. Contributing the first UK evidence to a field dominated by findings about the USA, we show that measurement errors are pervasive, but the 4 types are quite different in nature. We also document substantial heterogeneity in each of the error distributions.

**Keywords:** measurement error, earnings, finite mixture models, linkage error, linked survey data and administrative data

## Introduction

Most studies of measurement error in household survey data on employment earnings have assumed that individually linked administrative data provide a benchmark earnings measure that is error-free. A few recent studies also allow for measurement errors in the linked administrative data. Our paper belongs to this second generation of research. Using individual-level data from the 2011/12 Family Resources Survey (FRS) linked with administrative data based on Pay As You Earn (PAYE) records ('P14' data) for the same individuals, we contribute the first UK evidence about measurement errors in employment earnings to a field dominated by findings about the USA. Our novel application uses new statistical models that are the first to incorporate 4 types of measurement error in models of earnings while also allowing error distributions to vary with observed individual characteristics.

We distinguish 4 types of measurement error in models of earnings based on linked datasets:

| Survey data | Administrative data |
|---|---|
| 1. Measurement error | 3. Linkage error |
| 2. Reference period error | 4. Measurement error |

The textbook Classical measurement error model provides a reference point; it focuses on survey measurement error alone (type 1). Let $s_i = \xi_i + \varepsilon_i$, where $s_i$ is log employment earnings for individual

**Received:** September 13, 2021. **Revised:** August 18, 2022. **Accepted:** October 28, 2022

$i$, and $\xi_i$ is true log earnings. Assuming error $\varepsilon_i$ has a mean of zero, positive variance, and zero covariance with true earnings, completes the Classical model. This simple specification has powerful implications. Survey responses provide unbiased measures of true earnings but overestimate true earnings inequality (the variance of log earnings is greater). If survey earnings are used as the dependent variable in a linear regression model (rather than true earnings), the slope coefficient for an explanatory variable is consistently estimated. But if survey earnings are used as the explanatory variable in a linear regression model, there is the well-known attenuation result: the slope coefficient underestimates the true coefficient (that derived were true earnings observed). Finally, if you also assume that log earnings in the linked administrative data source, $r_i$, are error-free—and hence equal to true log earnings—it is straightforward to quantify the overestimation of earnings inequality and the extent of attenuation and to examine how errors differ across survey respondents with different characteristics.

There are 2 problems with relying on the Classical model. First, the assumptions may not hold. Survey measurement errors may be biased (error $\varepsilon_i$ has a nonzero mean), and errors may be correlated with true earnings. With regard to the latter feature, most discussed has been the case of 'mean reverting' errors (a negative correlation), so that survey responses are overreports (relative to true earnings) below the mean but underreports above the mean. If the degree of mean reversion is sufficiently large, survey earnings inequality is less than true earnings inequality (Gottschalk & Huynh, 2010), not the reverse as for the Classical model. Also, with mean-reversion, the slope coefficient in a linear regression with survey earnings as the dependent variable is no longer consistent, and a slope coefficient in a linear regression with survey earnings as the independent variable need not be attenuated (Kapteyn & Ypma, 2007, 529–531).

The second problem is that there are other types of measurement error, and their existence further constrains the conclusions that can be drawn. When there are administrative data errors, formulae for biases in parameter estimates become more complicated, as Kapteyn & Ypma (2007) have shown. No longer does the difference between a linked survey and administrative report necessarily reflect a single error type, and the impact of the errors on inequality and regression estimates depends on the magnitude and direction of the various errors.

We consider 2 types of errors in survey data. *Survey measurement error* is as already discussed but extended to incorporate the possibility of mean reversion and bias. *Reference period error* arises when the reference period for respondents' survey earnings differs from the reference period for their earnings in the administrative data. There are 2 potential types of error in linked administrative data. There is *linkage error* if, when creating the linked data set, there is mismatch: the survey record for each of several survey respondents is incorrectly linked with some other person's administrative data record, thereby importing error into the linked data. Finally, there may be *administrative measurement error* in the administrative data, analogous to what arises in survey data albeit with different provenance. Both types of administrative data error mean that linked administrative data provide an imperfect benchmark for assessing errors in survey data.

The nature and importance of the different types of measurements may differ within a country and across countries (depending on the survey used and nature of the administrative data and the algorithm used for linkage), so it is important to add evidence about the UK to existing knowledge about other countries.

Table 1 lists earlier studies of measurement error in employment earnings that have used linked survey and administrative data. The top panel cites research that has assumed there are no errors in the administrative data reports (no measurement error and no linkage error), i.e., first-generation studies. Their focus on the USA is immediately apparent. The bottom panel lists the smaller number of studies that have accounted for administrative data errors of either kind (second-generation studies). We use '*L*' to indicate research allowing for linkage error and '*M*' to indicate research allowing for administrative data measurement error. Prior to this article, there has been only 1 study, Bollinger et al., (2018), that allows for both types of administrative data error. Kapteyn & Ypma (2007) is the pioneering second-generation study but incorporated linkage error only. (Our earlier research using UK data, Jenkins & Rios-Avila (2020), fitted the same models as Kapteyn & Ypma.) Here, we extend Kapteyn & Ypma's approach to incorporate measurement error in the administrative data, and additionally, our paper is the first to also allow error distributions to differ across individuals by modelling parameters as functions of covariates.

**Table 1.** First- and second-generation studies of measurement errors in employment earnings

| Country | Study | Survey data | Administrative data |
|---|---|---|---|
| (a) First-generation studies (assume no error in linked administrative data) | | | |
| USA | Bound and Krueger (1991) | CPS | SSA |
| USA | Bollinger (1998) | CPS | SSA |
| USA | Duncan and Hill (1985) | PSIDVS | Firm payroll data |
| USA | Bound et al. (1994) | PSIDVS | Firm payroll data |
| USA | Pischke (1995) | PSIDVS | Firm payroll data |
| USA | Bricker and Englehardt (2008) | HRS | SSA |
| USA | Pedace and Bates (2000) | SIPP | SSA |
| USA | Gottschalk and Huynh (2010) | SSA | SSA |
| USA | Kim and Tamborini (2014) | SSA | SSA |
| Austria | Angel et al. (2019) | SILC | Wage tax registers |
| Denmark | Kristensen and Westergaard-Nielsen (2007) | ECHP | Linked registers ('IDA') |
| Germany | Valet et al. (2019) | LINOS | IAB IEH |
| (b) Second-generation studies (allow for error in linked administrative data) | | | |
| Sweden | Kapteyn and Ypma (2007)$^{L}$ | SHARE | Linked registers ('LINDA') |
| UK | Jenkins and Rios-Avila (2020)$^{L}$ | FRS | P14 |
| USA | Abowd and Stinson (2013)$^{M}$ | SIPP | SSA |
| USA | Bollinger et al. 2018)$^{L,M}$ | CPS | SSA |
| Denmark | Bingley and Martinello (2017)$^{M}$ | SHARE | Linked registers |
| New Zealand | Hyslop and Townsend (2020)$^{M}$ | SoFIE | Wage tax registers ('IDI') |
| UK | The current paper$^{L,M}$ | FRS | P14 |

*Note*. CPS = Current Population Survey (Annual Social and Economic Supplement); ECHP = European Community Household Panel; FRS = Family Resources Survey; HRS = Health and Retirement Study; IAB IEH = Federal Institute for Employment Research Integrated Employment Histories; LINOS = Legitimation of Inequality over the Life Span survey; P14 = National Insurance administrative records (see main text); PSIDVS = Panel Study of Income Dynamics Validation Study; SHARE = Survey of Health, Ageing, and Retirement in Europe; SILC = Statistics on Income and Living Conditions; SIPP = Survey of Income and Program Participation; SoFIE = Survey of Family, Income and Employment; SSA = administrative records on employee earnings held by Social Security Administration. All the administrative data sources on earnings are based on employer reports as part of a national social insurance system.
$^{L}$Study allows for linkage error.
$^{M}$Study allows for measurement error in the administrative data. All studies listed, except for the current paper, assume that measurement error distributions are the same for all respondents.

There are some other features of previous research to note and contrast with the current paper. First, the administrative data sources for almost all studies, including ours, arise from government agency collation of information provided by employers about social insurance contributions and pay withheld from employees. [The exceptions are the Panel Study of Income Dynamics Validation Study (PSIDVS)–based papers; these use payroll data from a single company.] Hence, the sources of administrative data measurement error are likely to be similar across settings though may vary in importance.

Second, no study cited in Table 1 considers reference period error. Kapteyn & Ypma's (2007) model specification allows for 'contamination' error; we exploit this in the current paper. However, Kapteyn & Ypma discuss potential sources of contamination error in only 1 sentence (2007, 528) and do not address reference period issues as we do (nor does any other paper). This omission is appropriate in Kapteyn & Ypma's case because reference period issues are much more important in the UK context than in their Swedish application. More generally, reference period issues are irrelevant in countries where the survey collects data about annual earnings directly, and the year referred to is the reference period in the administrative data too. This is the case for the studies listed in Table 1, which are based on the Current Population Survey (CPS),

PSIDVS, Health and Retirement Study, Survey of Health, Ageing, and Retirement in Europe, Statistics on Income and Living Conditions, and European Community Household Panel. However, in Valet et al.'s (2019) study for Germany, Legitimation of Inequality over the Life Span survey respondents were asked their monthly gross earnings, but Federal Institute for Employment Research Integrated Employment Histories administrative earnings data are recorded as daily amounts. The authors do not take account of this when reconciling earnings reports. In the US SIPP-based studies, survey earnings refer to monthly earnings over a 4-month period; the annual earnings amounts used for comparisons with SSA annual earnings are derived by summing the monthly earnings reports for the relevant year. This process could conceivably lead to reference period error, but it is likely to be less serious than in the German or UK contexts. As we explain later, UK household surveys collect sub-annual earnings data (the month is the most common reporting period), whereas our P14 administrative data refer to annual earnings. We address this mismatch in our modelling.

Third, there is no previous research about earnings measurement errors using UK survey and administrative data (aside from our earlier work using the Kapteyn & Ypma (2007) model). Britton et al. (2019) mention several of the issues that motivate our contributions in the current paper, but they were unable to address them. For example, they mention the possibility of measurement error in administrative data on earnings (as well as in survey data) and also cite differences in earnings reference periods between their survey data (the UK Labour Force Survey) and administrative data (Student Loan Company records). Britton et al. (2019) did not have individually linked data and so were only able to compare marginal distributions of survey and administrative earnings. They could not calculate individual-specific differences between survey and administrative data reports, and yet such data are essential for assessing the nature of measurement error distributions. We have such data.

Our paper proceeds as follows. In the *Four types of measurement error* section, we discuss the 4 types of measurement error in detail with specific reference to the UK context. We review the sources of the various errors, explaining also how these relate to differences in error distributions across individuals which we model using covariates. The *FMMs of earnings incorporating 4 types of measurement error* section sets out the finite mixture models (FMMs) that we use to characterize true earnings and the 4 types of measurement errors and discusses how our models relate to those used in the earlier studies cited in Table 1. We explain how our models are identified and fitted by maximum likelihood, and how we derive post-estimation average predictive margins of model parameters for the sample as whole and for sets of respondents with different characteristics. The *The linked FRS-P14 dataset* section describes our linked survey and administrative data, the distributions of log earnings in each source, and the distribution of individual-level differences between survey and administrative data reports on earnings.

We present estimates of our statistical models for reconciling earnings reports in the *Model estimates* section. We find that measurement errors are pervasive but quite different in nature. We estimate the probability of measurement error in survey data to be around 94% and the probability of measurement error in linked administrative error around 37%. However, the standard deviation (SD) of survey data errors is markedly smaller than the administrative data error SD. In addition, there is a probability of linkage error of around 6%. Reference period error has a low prevalence (around 8%) but introduces substantial noise into survey earning reports. We also document substantial heterogeneity in error distributions. The *Summary and conclusions* section contains a summary and conclusions.

The Online Supplementary Material contains additional estimates and discussion, as we explain later. Jenkins & Rios-Avila (2021b) report further robustness checks and additional analysis. For example, we predict the survey and administrative data error types each observation is prone to, and we derive 'hybrid' earnings variables combining information from survey and administrative data that have greater reliability than earnings from each source separately. We also illustrate the nature of bias in estimates of regression model parameters that are introduced when error-ridden earnings measures are used as a dependent variable or an explanatory variable rather than 'true' earnings.

The readme.txt document accompanying this article describes how to download our Stata code (ado- and do-files) and discusses data access and replication. Replication materials are downloadable from https://bit.ly/3dCmbMi.

## Four types of measurement error

There are 4 possible types of measurement error in linked survey and administrative data. Any one of them may lead to differences between survey and administrative reports on earnings for a specific individual. All of them need to be considered when reconciling the earnings reports in linked survey and administrative data.

### Survey measurement error

Survey measurement error is typically assumed to follow the Classical model (discussed above), augmented to incorporate possible bias and mean-reverting errors, i.e., whether individuals systematically under- or over-report earnings. Errors arise for multiple reasons. Moore et al.'s survey (2000) points out that cognitive factors may be at least as important as motivated misreporting, citing examples such as respondent misunderstanding of the concept asked about, faulty retrieval by respondents because of faulty recall or low salience of some items, and various types of sensitivities to questions about money.

An example of misunderstanding of the earnings concept would be when FRS respondents mistakenly report earnings net of some deductions such as those related to salary sacrifice for pension contributions. FRS guidance for interviewers (Department for Work and Pensions, 2012) is intended to minimize such misunderstanding (and the net earnings questions come prior to the gross earnings ones), but it may arise nonetheless.

Question sensitivity is closely related to issues of social desirability bias. Bound et al. (2001) write in their survey that '[i]t is widely believed and well documented that … questions [about socially and personally sensitive topics] elicit patterns of underreporting (for socially undesirable behavior and attitudes) as well as overreporting (for socially desirable behaviors and attitudes)' (2001, 3746). This is the standard explanation for mean reversion in survey measurement error, and explicitly cited as such by Angel et al. (2019) for example. Our models allow for mean reversion in survey measurement error.

Moore et al.'s (2000, 342–345) review indicates that survey measurement error variances for employment earnings are significant but there is no clear evidence about bias (summarized by the mean of the measurement error distribution). Our models provide new evidence about both error bias and variance.

We hypothesize that the error variance may be greater among older workers (for cognition reasons) and among part-time workers compared to full-time workers (related to lower salience or less accurate recall for part-time work, often more variable). Bound & Krueger (1991) report that US women have smaller error variances than men (private, nonagricultural employees across the full age range), though Kapteyn & Ypma (2007) report no significant difference (for Swedish employees aged 50+). Bound & Krueger (1991) also report greater mean-reversion in earnings for men compared to women. In our models, we allow survey measurement error distribution parameters to differ by sex.

An additional issue in the UK context is that the FRS collects earnings information for 3 jobs at most (see below). If a respondent has earnings from a fourth or other employment, these would be missed by our FRS earnings variables but may be captured in the worker's P14 earnings record (which is compiled from all employments over the financial year). However, this factor is of negligible relevance in our data because the fraction of individuals with more than one job is tiny (around 4%).

Moore et al. (2000, 353–4) discuss US studies about how consultation of records by survey respondents may reduce differences between survey and administrative reports. Using our models, we quantify whether respondent consultation of payslips during the FRS interview is associated with a reduction in measurement error. (We do not know if payslips consulted by the respondent were also shown to the interviewer.)

Errors may also enter survey responses through interviewer key entry errors (e.g., mis-keying numbers) and subsequent data processing, though these factors are likely to be less important than cognitive ones and motivated misreporting given the computer-assisted collection of survey data nowadays.

In the light of this discussion, we use the following covariates to model heterogeneity in the survey measurement error distribution: respondent's sex, whether the respondent is aged 60+ or not,

whether the respondent works in a part-time or full-time job, and whether the respondent consulted a payslip. The *FMMs of earnings incorporating 4 types of measurement error* section explains how we specify distribution parameters as functions of covariates and the *The linked FRS-P14 dataset* section provides further rationales for these specific covariate definitions and summary statistics.

## Reference period error

Mismatch between the earnings reporting periods in the survey data and in the administrative data is an important issue for all major UK household surveys, including the FRS. Strictly speaking, this mismatch is not an 'error' per se, but a consequence of differences in the design of a country's data collection instruments. However, it is appropriate to refer to it as a form of error because it must be considered (along with the other 3 error types) when reconciling earnings reports in the 2 data sources. Put differently, we assume that annual earnings is the concept of primary interest, as in previous studies, and hence treat the reference period (annual) in the P14 data as correct but potentially incorrect in the FRS data.

We consider gross earnings in this paper, i.e., earnings prior to deductions for income tax and national insurance contributions, and other types of deduction. This is the earnings concept used by both the survey and administrative data sources.

In the FRS, gross earnings refer to jobs in progress at the interview date ('current' earnings). For each job in turn, up to a maximum of 3, the interviewer asks each employed respondent 'What was the gross wage/salary—i.e., the total, before any deductions?'. A follow-up question asks for the reference period to which that amount refers. See Department for Work and Pensions (2012) for the 2011/12 FRS User Guide with annotated questionnaire.

Around 70% of our sample (discussed below) report '1 calendar month' for their gross earnings reference period. The next most prevalent report is '1 week' (17%), then '4 weeks' (7%), '1 year/52 weeks/12 months' (4%), and '2 weeks' and 'other' (each 2%). Three other response options receive few responses. The FRS data producers convert the gross earnings responses for each job to weekly amounts pro rata—the originally-reported amounts are not released—which we converted to annual amounts (pounds per year). Earnings amounts are not top-coded.

FRS interviews are undertaken throughout the financial year and so respondents' earnings reference periods do not refer to specific calendar dates that are common to all. (Our data include responses from interviews undertaken in the 12 months between April 2011 and March 2012. There is an even spread over the year. For 2011/12 FRS documentation, see Department for Work and Pensions, 2013.) Hence, there are noncomparabilities between the annualized and genuinely annual measures that linked data analysis must address and we do this using a model-based approach. Kapteyn & Ypma (2007: 538) briefly mention reference period error as a source of 'contamination' and they model contamination in addition to mean-reverting survey measurement error. We bring this component to the fore, renaming it 'reference period' error.

The stabilities of an individual's employment and earnings are important factors for reference period error. If an employee stays in the same job(s) throughout the tax year or longer, receiving the same pay, no reconciliation of reports is required on reference period grounds: the survey's annualized current earnings measure equals the administrative data's annual measure by construction. However, the annualized FRS measure may be greater than the P14 annual earnings measure if the respondent experiences spells of unemployment or lower pay either before or after the reported reference period. That is, the interview response captures earnings in good times but misses the shortfall of earnings in bad times, but the P14 records both, in effect averaging them. Conversely, the annualized FRS earnings measure may be smaller than its P14 counterpart if the respondent has a higher-paid job outside the reported reference period whether through job change, promotion, or a cost-of-living increase, or if end-of-year bonuses are not reported in the earnings figure at the interview. Overall, it is unclear ex ante whether reference period error corresponds to under- or over-estimation of annual earnings on average. In our models, the former (latter) case corresponds to a negative (positive) mean for the reference period error distribution.

We expect higher-paying and full-time jobs to be more stable than lower-paying and part-time jobs (see, e.g., Golden, 2016), and shorter pay reference periods to be more prevalent among workers with lower-paying jobs. (In our FRS data, annualized earnings for workers reporting

earnings using an annual reference period are greater on average than those for workers reporting using a monthly reference period, and these are in turn greater than for those reporting using a weekly reference period.) Hence, we hypothesize that reference period error is negatively correlated with true earnings and our models incorporate this possibility—this is another innovation not incorporated in modelling to date. We also investigate how the distribution of reference period error varies with each of several measures of job stability.

To capture (in)stability-related heterogeneity in the reference period error distribution (in addition to this correlation), we use 3 indicators: whether the gross earnings reference period is reported as 'other', whether every earnings spell in the P14 data for a respondent span the 2011/12 tax year, and whether the respondent works full-time or part-time.

## Linkage error

Linkage error arises when a survey respondent is linked to the wrong individual in the administrative data, in which case the individual's linked administrative data earnings measure is a draw from the complete administrative earnings distribution. Kapteyn & Ypma (2007); Meijer et al. (2012) show that even a small linkage error rate has serious consequences for the reliability of administrative data compared to survey data: the former no longer provide a clean benchmark. Linkage error is also of substantive importance because when models incorporate it, survey measurement errors are no longer found to be mean-reverting (Jenkins & Rios-Avila 2020; Kapteyn & Ypma 2007). Observe that linkage error refers to matches that are incorrect; it does not refer to linkages that are unsuccessful, i.e., not achieved at all. All earnings measurement error studies to date analyse data for employees for whom a link is achieved.

The prevalence of linkage error is likely to depend on the algorithms and match keys used to undertake the linkage and, relatedly, the country context. In an ideal world, individuals would each have a unique personal identifier (e.g., social security or national insurance number) which is accurately recorded in survey and administrative data sources, and linkage algorithms would use this information directly. The Nordic countries are closest to this ideal. Kapteyn & Ypma (2007) report that every Swede has a unique social security number (SSN), also used in all administrative registers, and this is the linkage key in their study. And yet, Kapteyn and Ypma's pioneering research also estimated that, for their sample of Swedish individuals aged 50+, the probability of linkage error was around 5% (2007, Tables C1 and C2, Full model). They refer to its source being wrong or mistyped SSNs (p. 518) and point out that '[t]here is no … mechanism that would verify the correct linking of records in the construction of administrative analysis files' (p. 519).

Most US earnings measurement error studies have used unique Protected Identification Keys (PIKs) to undertake linkage between survey and Social Security Administration (SSA) earnings records. The SSA has a master file of PIKs derived from information originally provided to create an SSN and, for the survey(s), the Census Bureau creates PIKs by probabilistic record matching methods using information on name, address, birth date, and gender. Abowd & Stinson (2013) acknowledge the possibility of linkage error but, like all other US studies with one exception, do not model it. Bollinger et al. fit 1 model with linkage error, estimating its probability to be between 8% and 10% (2018, Appendix Table 2).

The UK situation is different again. The major household surveys, including the FRS, do not collect national insurance numbers because of doubts about the accuracy of the reports. (See Jenkins et al., 2008 about this issue. For related discussion of errors in linking of multiple administrative registers, see Harron et al., 2017.) In fact, linkage of individual record data from the major UK household surveys with administrative data sources is in its relative infancy. The current paper results from the Department for Work and Pensions' Secure Data Pilot initiative which included multiple projects investigating the use of administrative data along with survey data. The data linking undertaken for our project was a bespoke exercise. We return to this issue and more recent developments in the Conclusions.

When Department of Work and Pensions (DWP) statisticians created the linkage between 2011/12 P14 and FRS data, they linked individual records deterministically, using first name, last name, postcode, sex, and date of birth as the match keys. This algorithm is likely to lead to linkage error and so our models incorporate it.

An interesting issue is whether the linkage error probability is random, i.e., not varying with respondent characteristics, which is what Kapteyn & Ypma (2007) and all other earnings measurement error studies incorporating linkage error assume. A referee suggested that the DWP's matching algorithm may lead to systematic differences in linkage error probabilities across individuals, citing evidence about linkage of administrative data sources provided by, e.g., Bohensky's (2016) review.

We report estimates from models that allow differences between individuals of White and non-White ethnic group. Our hypothesis is that on average the latter group have more complex names than the former group and that this is associated with a greater probability of linkage error (first and last names are match keys). Preliminary modelling revealed no differences in linkage error probabilities by sex. We have no data on postcodes or exact dates of birth.

## Administrative measurement error

Administrative data on earnings may also contain measurement error per se in addition to linkage error (Abowd & Stinson, 2013; Bound & Krueger 1991). The most important reason is that administrative data on earnings are derived from employer reports to the relevant national social insurance authority. For the administrative data as for the survey data, there is an earnings 'reporter' but, in this case, the reporter is an employer rather than a survey respondent.

Our data refer to financial year 2011/12 when UK employers were mandated to provide end-of-year returns to the tax authorities (HMRC) about wages and salaries paid employees and the income taxes and national insurance contributions withheld as part of the PAYE system. Returns were made on P14 forms which could be returned on paper or electronically using approved payroll software. (We show a specimen P14 form in Online Supplementary Material, Appendix A3.) The earnings reported on P14 forms refer to gross earnings prior to deductions (the same earnings concept as in our survey data). UK P14 forms are thus similar to the W-2 forms returned by US employers to the SSA and used to compile administrative data on employee earnings. Other countries use similar processes.

Previous research suggests several potential sources of error. First, an employer's staff member may accidentally enter the wrong numbers on the paper form or mis-key entries into the payroll software that generates (in the UK context) the year-end P14 return to HMRC. Large mistakes may have been noticed and fixed in our P14 data file because HMRC had procedures by which employers can submit corrected returns, but smaller mistakes may have been overlooked or simply ignored. They may be over- or under-reports. Such mistakes may be more prevalent in small businesses without good payroll software, and more likely to occur for workers who are not on full-time or permanent contracts (supposing that records for these staff are of poorer quality). Public sector employers may be more accurate reporters than private sector employers on average because we expect the public sector to have better quality reporting software and this to be shared across the sector. We investigate these various hypotheses in our empirical application.

Second, there may be motivated misreporting by employers. The principal example cited by previous research is when an employer pays an employee informally, i.e., 'cash in hand' or 'under the table', perhaps to reduce liabilities for employer social insurance (national insurance) contributions. We conjecture that such practices are more prevalent among the same sorts of employers as we expect to have a higher prevalence of accidental mistakes.

Third, there is the issue of uncovered earnings—earnings reported in the survey but not in the administrative data because employers do not have to report them. Specifically, in our UK context, if employment earnings were below the lower earnings limit (LEL) for national insurance contributions, it was voluntary for employers to submit a P14 form for that employment. (In 2011/12, the LEL was £102 per week, i.e., just over £5,300 per year.)

Commenting on the second and third issues, the DWP informed us that 'some employers didn't submit records for people [for whom] there were no NI [national insurance] or tax liabilities—but these were largely people working for small employers who did not operate electronic payroll. We suspect these people would not appear in PAYE at all and would effectively be "cash in hand" employees. In reality we don't believe this was a massive issue and that most employment records were captured.' (DWP FRS Team, email 2020-02-25). The DWP statement is consistent with many employers submitting returns for below-LEL employees even if they did not have to.

Relatedly, observe that, if uncovered earnings were empirically important, we would expect to see marked changes in the probability density of P14 earnings around the LEL, but there are none: see our discussion of Figure 1 below. In addition, our statistical models can reveal whether uncovered earnings play a role. Specifically, if P14 earnings were systematically below FRS earnings at the bottom of the earnings distribution (apparent under-reporting below the mean), this would manifest itself in the form of mean-affirming errors in the administrative data. (The reasons for mean-reversion cited for survey errors, related to social desirability, do not apply to employer-reported administrative data.) We find little evidence of mean-affirmation, as we explain below. In sum, we conclude that uncovered earnings are not a major issue.

In the light of the discussion above, we use covariates to model heterogeneity in the P14 measurement error distribution that reflect potential differences in the quality of employers' P14 reporting: whether (the employee reports that) the employer provides a payslip or not; whether the survey respondent works part-time or full-time, and whether works in the private sector or public sector (main job).

## FMMs of earnings incorporating 4 types of measurement error

In this section, we propose FMMs that incorporate all 4 types of measurement error we discussed in the *Four types of measurement error* section. Our models extend Kapteyn & Ypma's (2007) most general ('Full') model. We account for reference period error by reinterpreting one feature of Kapteyn and Ypma's specification. Our most important modelling innovations are to incorporate the possibility of administrative data measurement error, and to allow every error distribution parameter to depend on covariates thereby introducing greater flexibility and thence better fit while also allowing us to investigate factors associated with higher or lower error bias and variance (the hypotheses set out in the *Four types of measurement error* section).

The intuition underlying the modelling strategy is that true annual earnings for an individual $i$, $\xi_i$, are unobserved but there are 2 observed earnings measures available: $s_i$ from the FRS and $r_i$ from the linked P14 data. Each measure is subject to error for the reasons discussed earlier, though not all individuals experience all types of error. For some, their FRS earnings measure is error-ridden and their linked P14 measure is not; for others it is vice versa; or both earnings measures are error-ridden. We can classify individuals into groups (latent classes) according to which types of error their earnings measures contain. Observed earnings are a combination ('mixture') of the distributions for the latent classes. We can identify the various error components by having the 2 observed earnings measures and by making assumptions about the nature of the measurement errors. We elaborate these remarks in the rest of this section. Our models are of log earnings (as in previous research) not earnings but, for brevity, we use the latter term.

### Nine types of survey and administrative data observation

We assume that the FRS earning distribution is a mixture of 3 types of observation, as summarized by eq. (1). In the first case (type $S1$), $s_i$ equals true earnings with probability $\pi_s$. In the second case (type $S2$), $s_i$ contains mean-reverting error with probability $(1–\pi_s)(1–\pi_\omega)$, with $\rho_s$ summarizing the correlation between error and true earnings. Third, there are observations subject to reference period error ($\omega_i$) in addition to survey measurement error (type $S3$), with probability $(1–\pi_s)\pi_\omega$.

$$s_i = \begin{cases} \xi_i \quad \text{with probability } \pi_s & \text{(type } S1\text{)} \\ \xi_i + \rho_s(\xi_i - \mu_\xi) + \eta_i \quad \text{with probability } (1-\pi_s)(1-\pi_\omega) & \text{(type } S2\text{)} \\ \xi_i + \rho_s(\xi_i - \mu_\xi) + \eta_i + \omega_i \quad \text{with probability } (1-\pi_s)\pi_\omega. & \text{(type } S3\text{)} \end{cases} \tag{1}$$

We assume that the linked P14 earnings distribution is a mixture of 3 types of observation, as set out in eq. (2). We distinguish between individuals for whom the linkage is correct (with probability $\pi_r$) and individuals who are incorrectly linked (probability $1–\pi_r$). Among the correctly linked observations, P14 earnings are either equal to true earnings, $\xi_i$, with probability $\pi_v$ (type $R1$), or are measured with error with probability $1–\pi_v$ (type $R2$). For each type $R2$ observation, the P14 measurement error may be correlated with true earnings with the correlation summarized by parameter $\rho_r$. There is mean reversion if $\rho_r < 0$ and mean affirmation (as mentioned in the previous section) if $\rho_r > 0$. In the third case (type $R3$), linkage error, the linked administrative data represent the P14

earnings not of the FRS respondent as intended but of someone else in the P14 dataset. The incorrectly-linked P14 earnings are $\zeta_i$. Their distribution is based on an unknown subset of observations in the population P14 database and need not have the same distribution as the distribution among our sample P14 observations.

$$r_i = \begin{cases} \xi_i & \text{with probability } \pi_r\pi_v & \text{(type } R1\text{)} \\ \xi_i + \rho_r(\xi_i - \mu_\xi) + v_i & \text{with probability } \pi_r(1-\pi_v) & \text{(type } R2\text{)} \\ \zeta_i & \text{with probability } (1-\pi_r) & \text{(type } R3\text{)} \end{cases} \quad (2)$$

In sum, there are 9 types of observation in the linked dataset corresponding to which of the 3 FRS and 3 P14 observation types are combined. For example, group 1 contains observations with the combination ($R1$, $S1$; error-free earnings in both data sources), group 2 contains observations with the combination ($R1$, $S2$), etc. Table 2 lists the 9 groups (observation types, i.e., latent classes) and their probabilities.

## Incorporating covariates

We allow distributions to vary with observed characteristics by expressing transformations of parameters as linear indices of characteristics, i.e.,

$$G(\gamma) = a_\gamma + \beta'_\gamma X_i. \quad (3)$$

For each parameter with generic label $\gamma$, $\alpha\gamma$ is a constant, $X_i$ is a vector of observed characteristics for individual $i$, and $\beta\gamma$ are the slopes associated with those characteristics. Transformations help estimation because they constrain parameters to stay within their theoretical bounds. Transformation function $G(.)$ is the identity function for means, the logarithmic function for SDs, the logistic function for probabilities, and Fisher's $z$ transformation for correlations. Mean-reverting errors in models with a heterogeneous mean earnings function refer to mean-reverting errors among individuals with the same observed characteristics.

To facilitate interpretation of parameter estimates, we back-transform them to their natural metrics, reporting Average Predicted Margins (APMs). That is, for each parameter $\gamma$, we predict the value of $\gamma$ for every observation using the fitted model, with values of all covariates (if included in the equation for $\gamma$) set at their sample values, and then report an estimation sample average of the derived $\gamma$ values, as well as the associated standard error.

**Table 2.** Groups (latent classes) in general mixture model of FRS and P14 earnings

| Group, $j$ | Description | Types | Probability, $\pi_j = \ldots$ |
|---|---|---|---|
| 1 | No error in P14 or in FRS earnings | $R1, S1$ | $\pi_r\,\pi_v\,\pi_s$ |
| 2 | No error in P14 earnings; error in FRS earnings | $R1, S2$ | $\pi_r\,\pi_v\,(1-\pi_s)(1-\pi_\omega)$ |
| 3 | No error in P14 earnings; error and reference period error in FRS earnings | $R1, S3$ | $\pi_r\,\pi_v\,(1-\pi_s)\pi_\omega$ |
| 4 | Error in P14 earnings; no error in FRS earnings | $R2, S1$ | $\pi_r\,(1-\pi_v)\pi_s$ |
| 5 | Error in P14 earnings; measurement error in FRS earnings | $R2, S2$ | $\pi_r\,(1-\pi_v)(1-\pi_s)(1-\pi_\omega)$ |
| 6 | Error in P14 earnings; measurement error and reference period error in FRS earnings | $R2, S3$ | $\pi_r\,(1-\pi_v)(1-\pi_s)\pi_\omega$ |
| 7 | Mismatched P14 earnings; no error in FRS earnings | $R3, S1$ | $(1-\pi_r)\pi_s$ |
| 8 | Mismatched P14 earnings; measurement error in FRS earnings | $R3, S2$ | $(1-\pi_r)(1-\pi_s)(1-\pi_\omega)$ |
| 9 | Mismatched P14 earnings; measurement error and reference period error in FRS earnings | $R3, S3$ | $(1-\pi_r)(1-\pi_s)\pi_\omega$ |

*Note*. $\pi_s$ = probability survey data are error-free; $\pi_\omega$ = probability of survey reference period error; $1-\pi_r$ = probability of linkage error; $1-\pi_v$ = probability administrative data contain measurement error. The KY+ and KY models have 6 latent classes ($j = 1, 2, 3$ and $7, 8, 9$). The General and Constrained General models have 9 latent classes ($j = 1, \ldots, 9$).

For each parameter, we report an average over all sample observations, labelled 'All' in our estimation tables below. To report how each parameter varies with different values of a specific covariate, we adapt this approach. For example, for a binary indicator variable *sex* taking the values 0 (male) and 1 (female), we calculate the APM of $\gamma$ for men by first setting all sample values of *sex* to 0 and then taking the average over the whole sample. (If other explanatory variables are included in the $\gamma$ equation, they are left at their sample values.) We calculate the APM of $\gamma$ for women analogously. This approach also allows us to test whether the difference between the APMs for men and women (or, more generally, any other binary contrast) is statistically significant. See Jenkins & Rios-Avila (2021a) for details.

## Distributional assumptions

We assume that true earnings ($\xi_i$), incorrectly-linked earnings ($\zeta_i$), and errors ($\upsilon_i$, $\eta_i$, $\omega_i$) are each independently normally distributed (conditional on observed characteristics) with the exception that true earnings and reference period errors ($\omega_i$) are bivariate normal. Thus, the distributions of the 'factors' may be written as:

$$\begin{pmatrix} \xi_i \\ \omega_i \end{pmatrix} = N\left( \begin{pmatrix} \mu_\xi \\ \mu_\omega \end{pmatrix}, \quad \begin{pmatrix} \sigma_\xi^2 & \rho_{\xi\omega}\sigma_\xi\sigma_\omega \\ \rho_{\xi\omega}\sigma_\xi\sigma_\omega & \sigma_\omega^2 \end{pmatrix} \right)$$
$$\zeta_i \sim N(\mu_\zeta, \sigma_\zeta^2),\ \eta_i \sim N(\mu_\eta, \sigma_\eta^2) \text{ and } \upsilon_i \sim N(\mu_\upsilon, \sigma_\upsilon^2), \tag{4}$$

where '$\mu$' and '$\sigma$' denote mean and SD, respectively, and $\rho_{\xi\omega}$ is the correlation between true earnings and reference period error that we cited in the *Four types of measurement error* section. $N(.)$ is the normal distribution. We do not restrict means to equal zero so that we can estimate whether errors introduce bias.

We assume normality to fit models by maximum likelihood and because it facilitates post-estimation derivations. The assumption is ubiquitous in this field, employed for example, by Abowd & Stinson (2013); Bollinger et al. (2018); Kapteyn & Ypma (2007). Moreover, we gain distributional flexibility by conditioning distributional parameters on characteristics. The normality assumptions do not constrain observed log earnings measures to be normally distributed since *s* and *r* are each a mixture of normal distributions and can vary with observed characteristics. The more substantive but untestable assumption is that true earnings are conditionally lognormally distributed.

When reporting estimates, we focus on 4 variants of our general model:

(a) the *General* model is the general model set out above;
(b) the *Constrained General* model is the General model with constraint $\rho_{\xi\omega} = 0$;
(c) the *KY+* model is Kapteyn & Ypma's (2007) Full model but also allowing $\rho_{\xi\omega} \neq 0$; and
(d) the *KY* model is Kapteyn & Ypma's (2007) Full model with constraint $\rho_{\xi\omega} = 0$.

The 2 KY models correspond to the case in which $(1-\pi_\upsilon) = 0$. Compared to the General models, the KY models have only 6 latent classes (in Table 1, $j = 1, 2, 3$ and 7, 8, 9). Comparisons of estimates of the 2 General models with estimates of the 2 KY models highlight the effects of neglecting measurement errors in the administrative data.

We have also fitted all the simpler cases of the KY model considered by Kapteyn & Ypma (2007) but do not report their estimates for brevity. The KY model always fitted better than the special cases of it.

Our models generalize those underpinning first generation measurement error studies (Table 1) because the latter do not incorporate linkage error or administrative data measurement error (or reference period error). Nor do they incorporate covariates. The Classical measurement error model falls into this group.

Among second generation studies, Bingley & Martinello's (2017) model for survey incomes is the same as that in our Constrained General model (and the KY models) except that there are no reference period errors. Their model of administrative data earnings is the same as our General model except that linkage error is ignored ($\pi_r = 1$) and there is no mean reversion ($\rho_r = 0$). No error distribution parameter varies with covariates in the Bingley & Martinello (2017) model.

Our General model shares some features of Bollinger et al.'s (2018) mixture models. Their models are designed principally to investigate connections between survey earnings response errors and earnings nonresponse per se. Bollinger et al. exploit the fact that the CPS has data from both an initial interview and an interview around 1 year later, and they compare survey measurement error patterns for those providing responses on both occasions and those providing a response only at the first interview. One variant of Bollinger et al.'s models incorporates mean-reverting survey measurement error and linkage error as ours do. However, Bollinger et al. do not allow for parameter heterogeneity other than in the mean of true earnings.

Hyslop & Townsend's (2020) model is not directly comparable to ours because their measurement error specifications are embedded within a model of earnings dynamics, i.e., true earnings are assumed to be the sum of a 'permanent' random walk component plus a transitory MA(1) component. (They have 8-year linked panel data.) Our models share some features with theirs because we both allow for survey and administrative data measurement error. However, Hyslop and Townsend ignore linkage error, they assume that there is no mean-reversion in the administrative data, and that error distribution moments are homogenous.

Abowd & Stinson's (2013) models are also not directly comparable with ours (or others) because they do not include any concept of 'true earnings'. Rather than fitting FMMs, Abowd and Stinson fit a bivariate linear mixed model in which there are measurement errors in both survey and administrative data, and mean observed survey and administrative earnings each vary with covariates. Because their pooled SIPP panel-SSA linked dataset provides more repeated measures than ours (across time and across jobs, not only across person), they can also fit multiple cross-data source correlations whereas we have a single individual random effect common to both earnings data sources. (Broadly speaking, our true earnings factor $\xi_i$ corresponds to Abowd and Stinson's 'common' earnings random effect, $c$.)

## Reliability

It is of considerable interest to know how consistently survey and administrative earnings measure true earnings, i.e., the reliabilities of the 2 observed measures. We report estimates of a commonly used psychometric measure of reliability, also employed by Meijer et al. (2012), i.e., the squared correlation between true earnings and an observed earnings measure:

$$\text{Reliability}(r) = \frac{[\text{cov}(\xi_i, r_i)]^2}{\text{var}(\xi_i)\text{var}(r_i)}, \text{Reliability}(s) = \frac{[\text{cov}(\xi_i, s_i)]^2}{\text{var}(\xi_i)\text{var}(s_i)}. \quad (5)$$

Reliability lies between 0 and 1. We derive reliability statistics using analytical expressions for the relevant variances and covariances that are implied by each of our models, aggregating appropriately across the latent classes (see Jenkins & Rios-Avila, 2021a). Because reliability statistics for $r$ and $s$ are model-contingent, they can be compared for a specific model but reliability estimates for $r$ (or $s$) should not be compared across models.

## Identification and estimation

Our mixture models are identified by the assumptions about the relationships between the 2 observed measures and true earnings and the nonnormal error structure arising from the mixture of normal distributions (Kapteyn & Ypma, 2007, 532). See also Yakowitz & Spragins (1968).

The additional structure our models have by comparison with standard FMMs plays an important additional role in identification. Specifically, the first latent class (group 1) contains observations for whom survey earnings equal administrative earnings and thence also true earnings. These observations are 'completely labelled' (Redner & Walker, 1984), i.e., membership of this class is known with certainty rather than latent. Parameters $\mu_\xi$ and $\sigma^2_\xi$ are identified by the completely labelled observations and having this baseline helps identification of the other parameters. Also, although the general model contains 9 latent class probabilities, these are each a function of only 3 probabilities (Table 1). Identification also relies on the assumption that latent class membership and true earnings are independent. The mean and SD of true earnings (conditional on covariates)

is the same for all latent classes: see the expressions for the latent class contributions to the sample likelihood shown in Online Supplementary Material, Appendix A1.

To fit our models, we must define when an observation's survey and administrative earnings measures are sufficiently close to count as 'equal'. Kapteyn & Ypma (2007) assumed that observations were completely labelled if survey and administrative earnings differed by less than SEK 1000 per year, which translates to a relatively large fraction of their sample, 14.8%. We are reluctant to use a completely labelled fraction that is so large because of the relevance of reference period error in the UK context. We assume that observations with $|r_i - s_i| < 0.005$ are completely labelled, which is 3.4% of our main sample. We have repeated analyses using a completely-labelled fraction more than twice as large, 7.7% (observations with $|r_i - s_i| < 0.010$), and we find that conclusions are robust. See also Jenkins & Rios-Avila (2020) and Jenkins & Rios-Avila (2021b, Appendices B–E).

We fit our models by maximum likelihood and report cluster-robust standard errors using the FRS household as the cluster. (The FRS attempts interviews with all individuals aged 16+ years in a sampled household; our estimation sample therefore contains some households in which more than 1 member provides linked data on earnings. See the notes to Table 3.) For details of the likelihood function and estimation method, see Online Supplementary Material, Appendix A1 and Jenkins & Rios-Avila (2021a).

We checked that our maximization algorithms were not converging to local maxima. Our principal strategy was to fit successively more complex models using starting values from less complex models. Our relatively parsimonious covariate specifications (see below) also helped avoid fitting problems. Convergence was straightforward in almost all cases and, in the few cases where problems arose, we addressed them by using multiple sets of initial values and checked that each led to the same maximum.

## The linked FRS-P14 dataset

Our analysis dataset is created by linking records for respondents to the FRS for financial/tax year 2011/12 to records for the same individuals in the P14 administrative data for the same year held by HMRC. We have already provided information about each of the 2 data sources and their

**Table 3.** Four types of measurement error, their sources, and covariates used in modelling

| Type of error and source(s) | Moments | Covariates from FRS-P14 Linked dataset |
|---|---|---|
| 1. *Survey measurement error (FRS data)* | $\mu_\eta$, $\sigma_\eta$, $\rho_s$ | |
| Respondent cognition, including misunderstanding, misremembering, and social desirability; motivated misreporting; Interviewer keying errors | | Male/female; whether aged 60+ years; part-time versus full-time job; whether respondent consulted payslip |
| 2. *Reference period error* | $\mu_\omega$, $\sigma_\omega$, $\rho_{\xi\omega}$ | |
| Instability of employment | | Whether FRS reference period is unusual; P14 employment spell(s) not spanning full year; part-time versus full-time job |
| 3. *Linkage error* | $\pi_r$ | |
| Matching algorithm | | Whether member of non-White ethnic group |
| 4. *Administrative measurement error (P14 data)* | $\mu_v$, $\sigma_v$, $\rho_r$ | |
| Employer data entry error, motivated misreporting, employer's administrative capacity, noncoverage of below-LEL earnings | | Whether employer provided payslip; part-time versus full-time job; private versus public sector job |

*Note*. LEL = lower earnings limit for employee National Insurance contributions (see main text). Moments for error distributions refer to the General model. In addition to the error distributions, the General model also describes true earnings (moments $\mu_\xi$, $\sigma_\xi$ with covariates sex, age (quadratic), educational qualifications, marital status, part/full-time job, and number of jobs), and earnings among mismatched respondents (moments $\mu_\zeta$, $\sigma_\zeta$, no covariates). See main text.

linkage. In this section, we provide additional information about the derivation of the linked data analysis sample and the definitions of the earnings variables underpinning the analysis, as well as preliminary description.

## Derivation of the linked-data analysis sample

In our 2011/12 FRS data, there are 13,851 employed respondents with at least 1 job, of whom 9,014 men and women (65%) gave their consent to data linkage. There are 6,432 men and women, 71% of the employees consenting to data linkage, for whom the DWP statisticians made a link between FRS and P14 records. This linkage success rate is smaller than the 80% reported by Lunn & McKay (2013) for consenting 2009/10 FRS respondents but not wholly comparable because their sample was different (not restricted to employees) and their administrative data source was the Work and Pensions Longitudinal Study (because their study was of accuracy of survey responses about receipt of cash benefits).

From this linked-data sample we dropped a small number of observations who declared themselves in the FRS to be 'self-employed' ($N = 23$) and then we also dropped observations for whom either FRS or P14 earnings were equal to zero ($N = 18$), giving us 6,391 employees (2,794 men; 3,599 women). Finally, we followed common practice further and dropped observations with imputed or otherwise edited values for gross earnings or reference period for any FRS job reported ($N = 420$). (In preliminary analysis, we refitted our models including imputed or otherwise edited observations, and estimates changed hardly at all.) The resulting estimation sample contains 5,971 individuals (2,595 men; 3,376 women). The age range is 16–84 years, with the vast majority (84%) aged 25–59 years, 6% aged 16–24 years, and 10% aged 60+ years.

We also have estimates for a subsample of 3,564 individuals aged 25–59 years in full-time work and not participating in any form of education. We do not report these estimates here for brevity and because results are similar to those for the main sample. See Jenkins & Rios-Avila (2021b, Appendices B–E).

## Representativeness of the estimation sample

The representativeness of our estimation samples is a potential issue: consent to data linkage and record linkage success among consenters may be selective processes. Following Bollinger et al. (2019, Appendix A4), we addressed this issue by constructing inverse-probability weights. We regressed the probability of the binary outcome 'consented to data linkage and successful linkage' on many individual characteristics using a probit model applied to the FRS sample of employed respondents, and derived weights equal to the inverse of the predicted probabilities. We then multiplied these weights by the FRS individual sample weight to create a new composite weight. See Online Supplementary Material, Appendix A for more about the derivation of the weights.

Reassuringly, the characteristics of our analysis sample are very similar to those of the full FRS sample (i.e., the sample that does not condition on consent or successful linkage). See Appendix, at the end of this paper, which shows weighted and unweighted means and SDs of earnings variables and means of other characteristics. Moreover, unweighted and composite-weighted estimates of our statistical models are generally very similar and so for brevity we report weighted estimates. (Unweighted estimates are in Online Supplementary Material, Appendix B.)

Overall, we conclude that consent and linkage biases are negligible in our linked data. This conclusion is similar to that of Sakshaug & Kreuter (2012) using German linked data. They report their 'results show that nonconsent biases are present for [a] few estimates, but are generally small relative to other sources of bias' (2012, 112).

Using weights in modelling reduces bias in parameter estimates (because the data are more representative of the target population) though has the potential downside of intrducing greater variability. Although we report weighted estimates, we cite some situations where the choice between weighted and weighted estimates may affect conclusions.

## The survey and administrative data earnings variables

We follow previous earnings measurement error research by analysing total gross earnings recorded in the survey and administrative data.

Our survey measure of earnings for each respondent $i$, $s_i$, is the logarithm of total gross earnings, i.e., the annualized sum of earnings across all jobs reported, where the annualization is undertaken as described earlier. Only 4% of our sample report earnings for more than 1 job (Appendix), and our preliminary modelling showed that measurement error distribution parameters did not differ between single-job and multiple-job holders. Hence, we do not examine multiple job holding further (except as a predictor of mean true earnings—see the *Model estimates* section).

Our administrative measure of earnings for each linked respondent $i$, $r_i$, is the logarithm of total gross annual earnings. For each individual in our linked P14 file, there is a row reporting earnings for each of the employments reported during the 2011/12 year, and we derive total earnings per individual by summing across the relevant rows. If a job starts (ends) within the 2011/12 year, the start (end) date is recorded; hence we can deduce whether a job spans the whole year or not, a measure of stability (see the *Four types of measurement error* section and below). No other information appears in the P14 data supplied to us.

## Distributions of earnings and earnings differences

We now describe the distributions of FRS and P14 earnings and the individual-level differences between them.

Figure 1 shows that the distributions of FRS earnings ($s$) and P14 earnings ($r$) are quite similar. Each has greater concentration around the mean than a normal distribution with the same mean and standard deviation and is slightly asymmetric. P14 earnings have a slightly lower mean than FRS earnings, 9.75 compared to 9.78, and greater SD, 0.85 compared to 0.82. The greater dispersion is inconsistent with a model in which P14 earnings represent the truth and FRS earnings contain only Classical measurement error (Kapteyn & Ypma 2007, 524).

Observe also that there are no marked changes in P14 earnings density around log earnings of 8.58, i.e., the value corresponding to the LEL of £5,304 per year in 2011/12. (Recall the discussion in the *Four types of measurement error* section about uncovered earnings.)

Figure 2 shows the distribution of differences between FRS and P14 earnings ($s_i$–$r_i$). There is a large spike at zero, with most differences tightly clustered around this value, which is the mean and median. There are close similarities with the corresponding graphs shown in earlier studies: see,

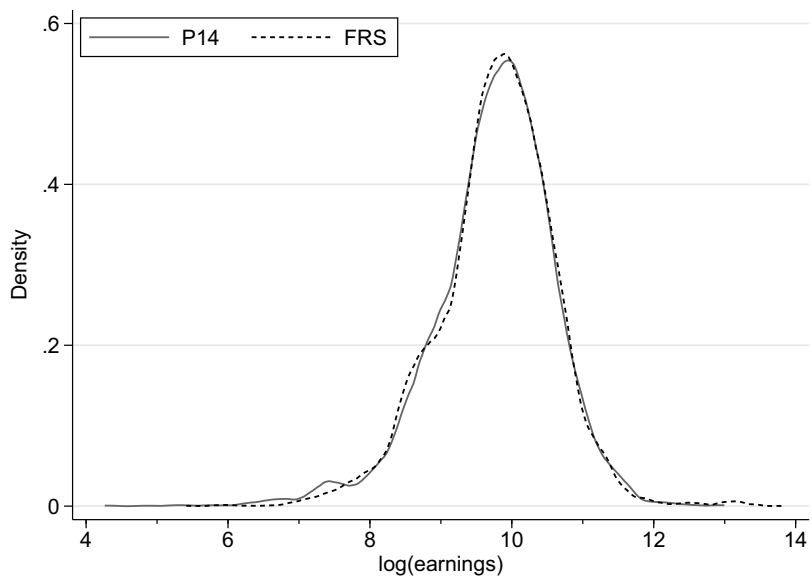


**Figure 1.** Distributions of Family Resources Survey (FRS) and P14 log(earnings).

*Note*: Kernel density estimates (Epanenikov kernel, 'optimal' bandwidth). Summary statistics for ($s$, $r$): mean (9.78, 9.75); $p5$ (8.39, 8.32); $p10$ (8.69, 8.70); $p50$ (9.84, 9.83); $p90$ (10.71, 10.70); $p95$ (10.98, 10.97); standard deviation (0.82, 0.85). Weighted estimates. Unweighted sample $N = 5{,}971$.

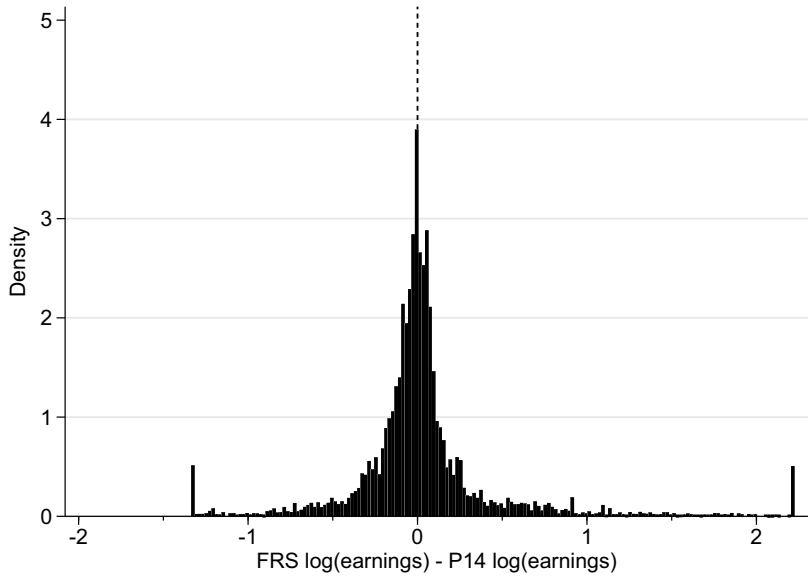


**Figure 2.** Distribution of difference between Family Resources Survey (FRS) and P14 earnings (*s–r*).

*Note:* Histogram with bin width = 0.02. Earnings differences are bottom-coded at *p*1 (−1.34) and top-coded at *p*99 (2.21) for purposes of presentation. Summary statistics for *s–r* (without bottom- or top-coding): mean, 0.032; standard deviation, 0.508; *p*5, −0.579; *p*10, −0.307; *p*50, −0.005; *p*90, 0.381; *p*95, 0.784. Weighted estimates. Unweighted sample $N = 5{,}971$.

e.g., Bound & Krueger (1991); Hyslop & Townsend (2020); Kapteyn & Ypma (2007); Kim & Tamborini (2014).

Figure 3 shows a scatterplot of FRS earnings against P14 earnings with a linear regression line superimposed. Were the Classical measurement error model to apply, the regression line would have a slope coefficient of 1; a slope of less than 1 is indicative of mean-reverting error. In Figure 3, the slope is 0.793 (SE 0.007) and significantly less than 1. Separate regressions by sex show the slope of the regression line is 0.693 (SE 0.013) for men and 0.830 (SE 0.010) for women, i.e., there is apparently greater mean-reversion in survey error for men than for women. These findings echo Bound & Krueger's (1991) for the USA. Also, separate regressions by age yield a slope coefficient of 0.742 (SE 0.031) for workers aged 60+ years and of 0.794 (SE 0.007) for workers aged less than 60 years.

Drawing conclusions about survey error mean reversion is contingent on the assumed model. We investigate mean reversion further, and differences in it by age and sex, using our non-Classical measurement error models.

## Covariates

For each of the 4 types of measurement error, the *Four types of measurement error* section discussed potential sources of the error and hence how one would expect heterogeneity in error distributions across the sample with systematic differences associated with survey respondent, job, and employer characteristics. Table 3 column 3 lists the covariates we use to capture this heterogeneity when fitting the models reported in the *Model estimates* section, separately for each of the 4 measurement error types (column 1) and for the moments in our General Model that summarize the respective error distributions (column 2). Our covariate specifications represent a balance between capturing important systematic variation on the one hand, and practical considerations on the other. The specifications are parsimonious to facilitate the fitting of our complex mixture models and, relatedly, almost all covariates are binary indicator variables. Also, we are limited to variables available in our dataset.

For true earnings, we suppose the mean ($\mu_\xi$) varies with sex, age, age-squared, educational qualifications, marital status, whether working part- or full-time, and number of jobs. We employ the same covariates in the SD equation ($\sigma_\xi$), except that we exclude marital status and number of jobs

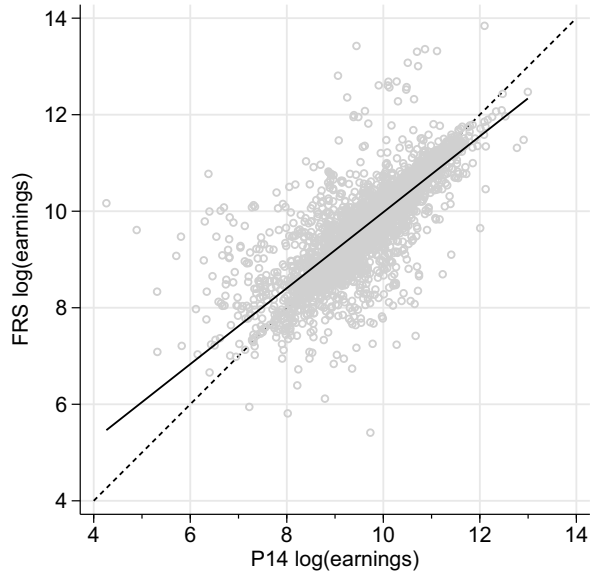


**Figure 3.** The relationship between Family Resources Survey (FRS) and P14 earnings.

*Note:* Scatterplot shows all ($r_i$, $s_i$) observations. A linear regression of FRS earnings on P14 earnings has slope coefficient 0.787 (*SE* 0.007), shown by the solid line. Weighted estimates. Unweighted sample $N = 5{,}971$.

as they were never significant in preliminary analyses. Earlier research has modelled mean earnings as a function of much the same characteristics, but our heteroskedastic error specification allows for greater flexibility in distributional shape while remaining feasible to fit.

We model the distribution of P14 earnings among incorrectly-linked observations without covariates. It makes little sense to relate the mean and SD of this distribution to the characteristics of the individuals to whom the linked earnings incorrectly refer. (The characteristics of the incorrectly-linked observations are unknown.)

We derived all covariates from the FRS because there are no covariates in the P14 data (with 1 exception discussed below). Like other researchers, we ignore potential measurement errors in covariates. The potentially most problematic variable is educational qualifications, for which around 9% of observations are missing. We assume these observations have educational qualifications below A-level standard. (In preliminary analysis, we allocated the missing observations to an additional separate educational qualifications category. The coefficients were similar to the below-A-level category and estimates of other model parameters were no different.)

Appendix, column (2), shows that our analysis sample is 54% female, and the average age is 41 years. One half have educational qualifications to A-level or higher (the minimum qualification for university entrance in the UK). Almost three-quarters are employed full-time in their main job and 70% live with a spouse. Only 4% report having more than 1 employment. Almost 30% work in the public sector in their main job. Around 60% consulted a current or recent payslip when reporting earnings. Employers did not supply payslips in 11% of cases. More than three-quarters (77%) are paid monthly; and 2% reported a nonstandard ('other') earnings reference period. For 61% of the sample, the earnings spells reported in the linked P14 data cover the full 2011/12 financial year, a measure of earnings stability.

## Model estimates

### Goodness of fit and reliability statistics

Table 4 summarizes goodness of fit statistics for the 2 General models and the 2 KY models. Every model fits substantially better—has substantially smaller Akaike information criterion (AIC) and Bayesian information criterion (BIC) values—than its counterpart without covariates (Online Supplementary Material, Appendix B). Of the 4 models, we prefer the Constrained

**Table 4.** Four models of log(earnings) with covariates: goodness of fit statistics and reliabilities

| | Models with administrative data measurement error | | Models without administrative data measurement error | |
|---|---|---|---|---|
| | General model ($\rho_{\xi\omega} \neq 0$) (1) | Constrained general model ($\rho_{\xi\omega} = 0$) (2) | KY+ model ($\rho_{\xi\omega} \neq 0$) (3) | KY model ($\rho_{\xi\omega} = 0$) (4) |
| Log pseudo-likelihood | −6458.8556 | −6459.0840 | −6759.6007 | −6762.8491 |
| AIC | 13027.7111 | 13026.1679 | 13603.2015 | 13607.6983 |
| BIC | 13395.9180 | 13387.6801 | 13884.3776 | 13882.1797 |
| Reliability($r$) | 0.7041 | 0.6999 | 0.6294 | 0.6186 |
| Reliability($s$) | 0.8222 | 0.8251 | 0.7846 | 0.7992 |

*Note*. Weighted estimates. Sample unweighted $N = 5{,}971$ individuals within 4,874 households. See the main text for the specifications of the 4 models and definition of reliability. Estimation based on a completely-labelled fraction of 3.4% (observations with $|r_i - s_i| < 0.005$: see main text). AIC and BIC are the Akaike and Bayesian information criteria, respectively.

General one (as we do for the corresponding models without covariates). It has the smallest AIC and BIC values, and the log pseudo-likelihood is almost identical to that for the General model. The only difference in specification between these 2 models is that the General model allows a nonzero correlation between reference period error and true earnings, but we find $\hat{\rho}_{\xi\omega} = -0.058$ (SE 0.075), i.e., of the expected sign but not significantly different from zero. Goodness of fit for the 2 General models is substantially better than for the 2 KY models. The important lesson is that models of earnings and measurement error should allow for measurement errors in the administrative data.

Table 4's bottom rows show reliability statistics. Regardless of model and for each reliability statistic, the conclusion is clear: the FRS data are more reliable than the linked P14 data. For example, according to the Constrained General model, Reliability is 0.83 for FRS earnings but only 0.70 for linked P14 earnings. That is, the errors in the linked P14 data (measurement error and linkage error) are more consequential that the errors in the FRS data (measurement error and reference period error). Our estimates of reliabilities for the KY model also favour the survey data, which is what Meijer et al. (2012) found for Kapteyn & Ypma's (2007) models (with only linkage error) and Swedish data. Overall, the full set of estimates points to the major source of unreliability in the administrative data being linkage error.

## Parameter estimates

We focus on the Constrained General model henceforth, with estimates shown in Tables 5 and 6. The tables show parameter estimates in the form of Average Predictive Margins (APMs), as described in the FMM*s of earnings incorporating 4 types of measurement error* section, derived both for the full sample ('all') and for the levels of the various covariates, where relevant. Table 5 shows APMs for the error probabilities (left-hand side) and latent class probabilities (right-hand side). Table 6 reports the other parameter estimates: the top left-hand set refers to APMs for means ($\mu$) and their SEs and the top right-hand set to APMs for SDs ($\sigma$). The bottom panel of the table contains APMs for mean-reversion parameters ($\rho_s$, $\rho_r$). We discuss the estimates for true and incorrectly linked earnings first and then the estimates for the error distributions.

*True earnings*: The estimated mean of true earnings is 9.81 and SD 0.49 ('all' estimates), by comparison with 9.78 and 0.82 for observed FRS earnings and 9.75 and 0.85 for P14 earnings. The observed measures slightly underestimate mean true earnings and substantially overestimate the true SD. The latter result is inconsistent with mean-reversion in survey earnings, as we confirm below. Differences across individuals are as we expect. Men have not only greater mean earnings than women but also more dispersed earnings at each age. Both gender differences are statistically significant, as indicated by the '+' (see the notes to Table 5). The estimates imply average earnings

**Table 5.** Constrained general model estimates: probabilities

| Error probabilities | APM | (SE) | Latent class probabilities | | | | |
|---|---|---|---|---|---|---|---|
| | | | | APM (White) | (SE) | APM (non-White) | (SE) |
| $\pi_s$ | 0.0550[b] | (0.0156) | $\pi_1$ | 0.0329[b] | (0.0026) | 0.0314[b] | (0.0027) |
| $\pi_\omega$ | 0.0773[b] | (0.0160) | $\pi_2$ | 0.5210[b] | (0.1537) | 0.4981[a] | (0.1607) |
| $\pi_r$ (All) | 0.9434[b] | (0.0276) | $\pi_3$ | 0.0437[b] | (0.0079) | 0.0417[b] | (0.0085) |
| $\pi_r$ (White) | 0.9476[b] | (0.0252) | $\pi_4$ | 0.0193 | (0.0127) | 0.0184 | (0.0117) |
| $\pi_r$ (non-White) | 0.9060[b] | (0.0550) | $\pi_5$ | 0.3052[a] | (0.1087) | 0.2918[a] | (0.0968) |
| $\pi_\upsilon$ | 0.6306[b] | (0.1515) | $\pi_6$ | 0.0256 | (0.0139) | 0.0245 | (0.0127) |
| | | | $\pi_7$ | 0.0029 | (0.0022) | 0.0052 | (0.0043) |
| | | | $\pi_8$ | 0.0457[c] | (0.0207) | 0.0820 | (0.0459) |
| | | | $\pi_9$ | 0.0038 | (0.0024) | 0.0069 | (0.0050) |

*Note*. As for Table 4. The values of $\hat{\pi}_1$ (completely labelled fractions) have SEs attached because our software calculated them from $\hat{\pi}_r\hat{\pi}_\upsilon\hat{\pi}_s$ (see Table 1). *P*-value for test of $\pi_r$ (White) = $\pi_r$ (non-White) is 0.19. [a]$P < 0.001$. [b]$P < 0.01$. [c]$P < 0.05$.

levels of £26,870 per year for men and £20,746 for women (in 2011/12 prices). [Given lognormality, expected true earnings equals $\exp(\hat{\mu}_\xi + \hat{\sigma}^2_\xi/2)$.] Individuals with at least university entry-level educational qualifications (1+ A-level exam grade passes or higher) earn more than less-qualified individuals, and their SD is also greater. Hence predicted average earnings for the more educated group are substantially greater: £28,863 per year compared with only £18,901 per year. Individuals working full-time earn more than double than those working part-time on average: £28,611 per year compared with only £13,220 per year. True earnings are greater for married individuals than single people, and for those with more than 1 job. Average true earnings and their dispersion increase with age up to the mid-forties but then both flatten off: predicted average earnings are £17,445 per year at age 25, rising to £27,343 per year at age 45 and £27,199 at age 55.

*Incorrectly linked earnings*: Mean $\hat{\mu}_\zeta$, 8.85, is smaller than the sample mean of P14 earnings (9.75), but $\hat{\sigma}_\zeta = 1.23$ is larger than the SD of sample P14 earnings (0.85). However, it is not as large as the ratio reported by Kapteyn & Ypma (2007), a difference we attribute to their sample's narrower age range (50+ years) than ours.

*Error and latent class probabilities*: The probability of linkage error, $1–\hat{\pi}_r$, is 6% for the sample as a whole, which helps explain the reliability statistics reported earlier. Bollinger et al. (2018, Appendix Table 2) using US data report a slighter larger probability, 10% for men and 8% for women. Kapteyn & Ypma (2007, Table C2) report a probability of 4% using Swedish data (but their model does not allow for administrative data measurement error). Bee & Rothbaum (2019, 19) hypothesize that '[a]llowing for mis-reporting in administrative data would likely reduce [Kapteyn and Ypma's] estimates of mis-linkage'. This is what we find: according to our KY model estimates (not shown), $1–\hat{\pi}_r = 10\%$ rather than 6%.

There is no clear evidence in favour of our hypothesis that linkage error probabilities are larger for non-White respondents than White respondents. Although the point estimate for the former group is 9% and 5% for the latter, the difference is not statistically significant ($P = 0.19$). The imprecision may reflect small numbers of non-White respondents or it might also reflect additional variability introduced by weighting. When we fit the constrained general model to unweighted data, the linkage error probabilities are much the same (11% and 5%), but the difference is statistically significant ($P = 0.03$).

FRS measurement error is very prevalent, with a probability $1–\hat{\pi}_s$ of 94%. The probability of P14 measurement error, $1–\hat{\pi}_\upsilon$, is only about four-tenths as large, 37%, but this is nontrivial in magnitude, nonetheless. [The probability is larger than what is reported by Kapteyn & Ypma (2007), around 85%, but the difference may reflect their assumption of a completely labelled fraction of 15% by contrast with our 3.4%.] The probability of reference period error is smaller still: $\hat{\pi}_\omega =$ 8%. This is consistent with most of our sample having stable jobs with little change in pay over the year.

**Table 6.** Constrained general model estimates: distributional parameters

| | | APM | (SE) | | APM | (SE) |
|---|---|---|---|---|---|---|
| ***All*** | $\mu_\xi$ | ***9.8101****** | ***(0.0126)*** | $\sigma_\xi$ | ***0.4949****** | ***(0.0083)*** |
| Male | | 9.9370***$^{+++}$ | (0.0177)$^{+++}$ | | 0.5235***$^{++}$ | (0.0137)$^{++}$ |
| Female | | 9.7035*** | (0.0133) | | 0.4732*** | (0.0091) |
| Education: less than A-level | | 9.6264***$^{+++}$ | (0.0129)$^{+++}$ | | 0.4411***$^{+++}$ | (0.0114)$^{+++}$ |
| Education: A-level or more | | 9.9943*** | (0.0181) | | 0.5520*** | (0.0111) |
| Full-time employee | | 10.0361***$^{+++}$ | (0.0123)$^{+++}$ | | 0.4509***$^{+++}$ | (0.0094)$^{+++}$ |
| Part-time employee | | 9.1762*** | (0.0252) | | 0.6266*** | (0.0177) |
| Married, cohabiting | | 9.8388***$^{+++}$ | (0.0140)$^{+++}$ | | | |
| Single, divorced, separated, widowed | | 9.7429*** | (0.0175) | | | |
| Has 1 job | | 9.8055***$^{+}$ | (0.0127)$^{+}$ | | | |
| Has 2+ jobs | | 9.9247*** | (0.0492) | | | |
| Age = 25 years | | 9.5604*** | (0.0208) | | 0.4128*** | (0.0125) |
| Age = 35 years | | 9.8258*** | (0.0144) | | 0.4821*** | (0.0096) |
| Age = 45 years | | 9.9507*** | (0.0141) | | 0.5310*** | (0.0111) |
| Age = 55 years | | 9.9352*** | (0.0153) | | 0.5515*** | (0.0134) |
| ***All*** | $\mu_\zeta$ | ***8.8460****** | ***(0.2527)*** | $\sigma_\zeta$ | ***1.2310****** | ***(0.1241)*** |
| ***All*** | $\mu_\eta$ | ***−0.0215****$^{++}$ | ***(0.0071)*** | $\sigma_\eta$ | ***0.1127****** | ***(0.0186)*** |
| Payslip(s) not consulted | | −0.0429***$^{+++}$ | (0.0082)$^{+++}$ | | 0.1434***$^{+++}$ | (0.0203)$^{+++}$ |
| Payslip(s) consulted (all jobs) | | −0.0063 | (0.0081) | | 0.0903*** | (0.0181) |
| Male | | −0.0277**$^{++}$ | (0.0086) | | 0.1316***$^{++}$ | (0.0226)$^{++}$ |
| Female | | −0.0163* | (0.0068) | | 0.0983*** | (0.0171) |
| Aged < 60 years | | −0.0148*$^{+++}$ | (0.0072)$^{+++}$ | | 0.1042***$^{+++}$ | (0.0194)$^{+++}$ |
| Aged 60+ years | | −0.1044*** | (0.0177) | | 0.2075*** | (0.0237) |
| Full-time employee | | −0.0186* | (0.0076) | | 0.1007*** | (0.0131) |
| Part-time employee | | −0.0298*** | (0.0090) | | 0.1465*** | (0.0369) |
| ***All*** | $\mu_\omega$ | ***0.0114*** | ***(0.1068)*** | $\sigma_\omega$ | ***1.0823****** | ***(0.1189)*** |
| Reference period: not 'other' | | 0.0243*$^{+++}$ | (0.1082)$^{+++}$ | | 1.0952***$^{+++}$ | (0.1229) |
| Reference period: other | | −0.5727*** | (0.1146) | | 0.4596 | (0.5086) |
| Employment spells do not span year | | −0.2833*$^{+}$ | (0.1212)$^{+}$ | | 0.9042***$^{+++}$ | (0.1280) |
| Employment spells all span year | | 0.2034 | (0.1703) | | 1.1959***$^{+++}$ | (0.1491) |
| Full-time employee | | 0.1726 | (0.1302)$^{++}$ | | 1.1203***$^{+++}$ | (0.1252) |
| Part-time employee | | −0.4411**$^{++}$ | (0.1679) | | 0.9708***$^{+++}$ | (0.2943) |
| ***All*** | $\mu_\upsilon$ | ***−0.0325*** | ***(0.0637)*** | $\sigma_\upsilon$ | ***0.2984****** | ***(0.1120)*** |
| Payslip provided by employer | | −0.0055 | (0.0497) | | 0.2685*$^{+++}$ | (0.1142)$^{+++}$ |
| Payslip not provided by employer | | −0.2411 | (0.1872) | | 0.5157***$^{+++}$ | (0.1251) |
| Full-time employee | | −0.0242 | (0.0671) | | 0.2310 | (0.1179)$^{+++}$ |
| Part-time employee | | −0.0559 | (0.0674) | | 0.4839***$^{+++}$ | (0.1022) |
| Private sector employee | | −0.0118 | (0.0659)$^{++}$ | | 0.3063**$^{++}$ | (0.1057) |
| Public sector employee | | −0.0872 | (0.0613) | | 0.2778*$^{+++}$ | (0.1339) |

(*continued*)

**Table 6.** Continued

| | | APM | (SE) | | APM | (SE) |
|---|---|---|---|---|---|---|
| ***All*** | $\rho_s$ | ***−0.0076*** | ***(0.0159)*** | $\rho_r$ | ***0.0902*** | ***(0.0719)*** |
| Male | | −0.0050 | (0.0217) | | | |
| Female | | −0.0098 | (0.0130) | | | |
| Aged < 60 years | | −0.0018 | (0.0169) | | | |
| Aged 60+ years | | −0.0792* | (0.0395) | | | |
| Payslip provided by employer | | | | | 0.0553 | (0.0352) |
| Payslip not provided by employer | | | | | 0.3619 | (0.4496) |
| Full-time employee | | | | | 0.0958 | (0.0799) |
| Part-time employee | | | | | 0.0746 | (0.0743) |
| Private sector employee | | | | | 0.1522 | (0.0915)$^{+}$ |
| Public sector employee | | | | | −0.0720 | (0.0569) |

*Note*. As for Table 4. APM = Average Predicted Margin. Cluster-robust standard errors in parentheses (cluster is household). Statistical significance indicators for tests of APM = 0: $^{*}P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$. Statistical significance indicators for tests of pairwise APM binary contrasts = 0: $^{+}P < 0.05$, $^{++}P < 0.01$, $^{+++}P < 0.001$. 'All' estimates refer to APMs (and associated SEs) calculated using the full analysis sample. Weighted estimates (see main text).

The error probabilities determine the probabilities of latent class membership (as summarized in Table 2). Table 5 shows that 4 classes account for almost 90% of the observations. The most prevalent class is type 2 ($R1$, $S2$), i.e., observations with error-free P14 earnings and measurement error-ridden FRS earnings: $\hat{\pi}_2 \approx 50\%$. (The estimate is 0.52 for White respondents and 0.50 for non-White respondents but the difference is not statistically significant.) Second most prevalent is type 5 ($R2$, $S2$), i.e., observations for whom both P14 earnings and FRS earnings contain measurement error: $\hat{\pi}_5 \approx 30\%$. Third, types 3 ($R1$, $S3$) and 8 ($R3$, $S2$), with reference period error and linkage error respectively, have probabilities of 4%–5% each. The probability of having error-free earnings (type 1; $R1$, $S1$) is around 3%, by construction, and the remaining 4 classes have a collective membership probability of at most around 4%.

*Error distributions: overall patterns*: We begin by reporting headline estimates for the sample overall ('all') and then turn to discuss how estimates differ by characteristics. Table 6 shows that survey, administrative, and reference period errors introduce little bias into observed earnings. Means $\hat{\mu}_\eta = -0.02$, $\hat{\mu}_\upsilon = -0.03$, and $\hat{\mu}_\omega = 0.01$, but only the first of these differs significantly from zero. In contrast, all 3 error types introduce significant variability of different magnitudes. The survey measurement error SD $\hat{\sigma}_\eta = 0.11$, but the administrative measurement error SD is more than twice as large: $\hat{\sigma}_\upsilon = 0.29$. The reference period error SD is substantially larger still: $\hat{\sigma}_\omega = 1.08$.

There is little evidence of mean-reverting or mean-affirming errors. We cannot reject hypotheses that $\hat{\rho}_s$ and $\hat{\rho}_r$ each equal zero ($\hat{\rho}_r = 0.09$ but is imprecisely estimated). Thus, our research confirms the finding of other second-generation studies that there is no mean reversion in survey measurement errors once one controls for administrative data error. As a result, earnings inequality estimates derived from FRS data over-estimate true earnings inequality. Dispersion introduced into the observed measure by measurement error (and reference period error) is not offset by mean-reversion. Ours is the first study to show that there is a similar result for administrative data. See Jenkins & Rios-Avila (2021b, *Summary and conclusions* section) for further discussion, including inequality estimates based on summary indices other than the SD of log earnings.

*Survey measurement error heterogeneity*: Table 6 shows that there are differences in survey measurement error mean and SD, represented by the differences in binary contrasts for the covariates listed. (Recall that Table 3 summarizes the rationale for including each of the variables.) Looking at means and taking the overall estimate of −0.02 as a benchmark, we see that there is statistically significant downward bias of greater magnitude for some groups, notably respondents who did not consult a payslip ($\hat{\mu}_\eta = -0.04$) and those aged 60+ years ($\hat{\mu}_\eta = -0.10$). The error SD is larger for respondents not consulting a payslip (0.14) compared to the SD for those who do (0.09).

There are also statistically significant differences in error SDs between men (0.13) and women (0.10), respondents aged less than 60 (0.10) and aged 60+ (0.21). Our estimates confirm that within-interview validation of responses can substantially reduce survey measurement error dispersion (and bias) and hence provide strong support for the FRS's long-standing procedure of seeking such validations. Our findings suggest that efforts by data collectors to raise the prevalence of record consultation further would improve survey data quality. In addition, the estimates are consistent with our hypothesis about older people experiencing greater cognition issues and with research for the USA finding that women are more accurate survey reporters than men. Our conjecture that there are differences according to whether the respondent was a full-time or part-time employee is not supported in the sense that, although there is greater bias and variability for part-timers than full-timers, the contrasts are not statistically significant.

Our finding that survey errors are not mean-reverting overall also applies to key subgroups. Table 6, bottom panel, shows estimates of $\hat{\rho}_s \approx 0$ for men and women and between respondents aged less than 60 and aged 60+. Again, although we find mean-reversion if we ignore administrative data measurement error and linkage error (recall the discussion of Figure 1), it disappears once these issues are accounted for.

*Reference period error heterogeneity*: Our expectation that individuals with unstable employment(s) would have greater differences between FRS (annualized) and P14 (annual) earnings than individuals with stable jobs is borne out by our estimates using several measures of instability. For example, first, the small minority who report unusual ('other') survey earnings reference periods have $\hat{\mu}_\omega = -0.57$ compared to $\hat{\mu}_\omega = 0.02$ for those with more standard periods. (In preliminary analyses, we found no clear differences in reference period error distributions when we differentiated between the full range of reported reference period options.) There are also substantial differences in reference period error according to whether all earnings spells recorded in the linked P14 data for the respondent span the 2011/12 financial year. Mean $\hat{\mu}_\omega$ is more negative for individuals with less stable jobs compared to those with stable jobs, −0.20 compared to 0.28. Similarly, part-time employees have a more negative mean ($\hat{\mu}_\omega = -0.44$) than do full-time workers ($\hat{\mu}_\omega = -0.17$). Although we find that greater earnings instability is associated with substantial under-estimation of average true annual earnings, we draw no conclusions about differences in error SDs: although some cross-group differences in $\hat{\sigma}_\omega$ are apparent, none of the contrasts is statistically significant. (This is also apparent in the unweighted estimates, so the issue is not imprecision arising from variability in the weights.)

*Administrative data measurement error heterogeneity*: The pattern of estimates here bears some similarities with those for survey data measurement error in the sense that cross-group differences are more apparent in subgroup error SDs than in error means. For example, for most subgroups, $\hat{\mu}_\upsilon \approx 0$. For workers whose employers do not provide payslips (a signal of inaccurate employer reporting), $\hat{\mu}_\upsilon = -0.24$, i.e., consistent with our expectations, but the contrast with those in jobs with payslips provided is not statistically significant.

The error SD for workers whose employers do not provide payslips is roughly twice the size of the SD for those in jobs with payslips ($\hat{\sigma}_\upsilon = 0.52$ compared with $\hat{\sigma}_\upsilon = 0.27$; a statistically significant contrast). Similarly, the error SD for part-time workers is around twice the size of the SD for full-timers ($\hat{\sigma}_\upsilon = 0.23$ compared with $\hat{\sigma}_\upsilon = 0.53$; a statistically significant contrast). These differences are consistent with the hypotheses we stated in the *Four types of measurement error* section. We also posited that there would be differences between private and public sector employees but we do not find this.

We found administrative data measurement errors to be neither significantly mean-reverting nor mean-affirming overall. In the *Four types of measurement error* section, we noted that if noncoverage of below-LEL earnings was poor in the P14 data, this may reveal itself in the form of mean-affirming errors among some subgroups, i.e., $\hat{\rho}_r > 0$. We find no clear-cut evidence for this. For example, we cited employees in jobs without payslips provided as a potentially relevant subgroup and for them, $\hat{\rho}_r = 0.36$. This estimate is positive (and large) but also very imprecisely estimated. For private sector employees, we find $\hat{\rho}_r = 0.15$, but this estimate does not differ statistically from zero.

## Summary and conclusions

Much research has argued that survey data on earnings suffer from measurement error; relatively little research has considered errors in linked administrative data on earnings, whether arising

from linkage error or measurement errors per se. We have modelled all 3 types of error, and we have also addressed reference period error, because UK household surveys provide 'annualized' earnings measures rather than genuinely annual measures as in the administrative data. In sum, our general statistical model is the first to incorporate 4 types of error when modelling employment earnings and their accuracy. Another of our innovations is to demonstrate how models with error distributions dependent on covariates can be employed to provide better fitting models than models without covariates (as in previous research) and to investigate differences in error distributions across individuals. And, our research provides the first empirical evidence for the UK to put alongside that for other countries (mostly the USA).

We find that the probability of measurement error in 2011/12 FRS earnings reports error is around 94%, whereas the probability of P14 measurement error is 37%, i.e., much smaller but nontrivial in magnitude. Moreover, the variance of P14 errors is larger than the variance of FRS errors. For the FRS, too, measurement error is not the only problem: there is also reference period error. This has a probability of around 8% and adds substantial noise to the FRS annualized earnings measure. Administrative data measurement error adds noise over and above that introduced by linkage error (which has a small but consequential probability of around 6%). On balance, however, the FRS data have greater reliability—are more strongly correlated with true earnings—than are the linked P14 data.

Our estimates highlight factors associated with poorer data quality. For example, survey measurement error variance is greater for workers not consulting a payslip to validate oral responses and for older workers. P14 measurement error variance is greater for respondents whose employers do not provide payslips and among employees working part-time. On the one hand, there is a positive take-away: restricting survey samples to full-time earners of 'standard' working age (as labour economists commonly do) reduces the variance of measurement error, and annualized current earnings measures are less noisy measures of annual earnings. (This is confirmed by comparisons with our estimates for the subsample of full-time workers aged 25–59: see Jenkins & Rios-Avila, 2021b, Appendices B–E.) Our findings point to initiatives that may improve data quality, e.g., encouraging greater use of payslips and other records in survey interviews, and improving employers' payroll systems especially for employees in part-time and less stable jobs.

We have examined earnings and errors among individuals with positive employment earnings (as in almost all previous earnings measurement error studies). An additional source of error might arise if there are employed survey respondents who incorrectly say they have no current job (and hence are routed away from the earnings questions) but correctly have positive annual earnings recorded in the P14 data. To address this issue requires linked data differently constructed from ours (which refer to individuals who report having at least 1 job) as well as more complex statistical models.

Our statistical models refer to the situation when there is a single cross-section of linked earnings data available. A few previous first- and second-generation studies have had access to linked data for panels of individuals, including Abowd & Stinson (2013); Bound & Krueger (1991); Gottschalk & Huynh (2010); Hyslop & Townsend (2020). Longitudinal linked data open up possibilities to estimate the correlation of errors within persons over time. (Previous research has estimated the first-order autocorrelation in errors, typically finding positive values, which implies that over- or underreporting is persistent.) Extending our mixture modelling approach to linked panel data raises interesting questions and challenges. For example, should researchers allow latent classes and membership probabilities to vary over time to allow for changing reporting behaviours? There are also related data and identification issues. Bound & Krueger (1991, 15) point out that researchers need long panels to identify an autoregressive process in the measurement errors separately from a person fixed effect or other time series process.

Our research has utilized administrative data created via employers reporting to the tax authorities as part of the UK's national insurance system, i.e., data created similarly to those used in earnings measurement error studies for other countries. Future work for the UK might also consider other linked data sources such as the earnings data provided by, e.g., the Annual Survey of Hours and Employment (a survey of employers). Models such as ours could be used to assess the nature of employer reporting error in this source alongside the other 3 relevant types of error. It would also be useful to use linked data methods to assess the quality of earnings data in other UK household surveys such as the Labour Force Survey and Understanding Society.

Readers should remember that our research is about errors in employment earnings, not total labour earnings (i.e., including self-employment or business income) as studied by Britton et al. (2019). Nor have we examined the accuracy of survey reporting about other income components such as welfare benefits. For these other types of income, the measurement error processes and nature of the relevant administrative data sources are likely to differ. For example, Brewer et al. (2017) argue that there is systematic under-reporting of welfare benefit income in the UK. It would be valuable to investigate this issue further using data linking FRS responses with the individual-level data on benefits held by the DWP, taking forward the work of Lunn & McKay (2013); McKay (2012). For related US research, see inter alia Celhay et al. (2021) who investigate measurement error in survey reports in 3 types of US welfare benefit income using linked administrative data for New York state.

There are other topics to address in further research as well. For example, we have characterized the distributions of earnings and errors using mixture models with conditionally normal errors. It is straightforward to use more flexible functional forms to characterize marginal distributions of earnings and errors (e.g., Generalized Beta of the Second Kind). However, for measurement error studies, we need multivariate counterparts to these distributions to specify cross-factor correlations and the likelihood function, and we are not aware of any.

It is important for future research to update our analysis to a later year to assess whether the nature of measurement error has changed. The UK has changed the technology used to administer the withholding system for income tax and national insurance. Starting April 2012, HMRC began to phase in a system of Real Time Information (RTI) and, since April 2014, employers must communicate to HMRC information about tax and other deductions under PAYE every time an employee is paid. Year-end P14 forms no longer exist, and HMRC's administrative data on employee earnings are now based on RTI (Office for National Statistics, 2019). Moreover, taking advantage of legislative changes, the FRS no longer asks respondents for consent to data linkage; instead, they are informed prior to interview that responses will be linked to administrative data for statistical and research purposes. In addition, data linkage incorporates probabilistic matching. As a result, linkage rates are expected to be close to 90% (Burke & Matejic, 2018), i.e., substantially greater than for our 2011/12 data.

The FRS-related changes are likely to reduce issues related to selectivity of consent, but it is unclear how the probability of linkage error and the intrinsic quality of administrative data earnings measures have changed. (Remember that linkage error is an issue for apparently successful linkages, and administrative data errors remain possible because employer reporting remains in RTI.) Also, reference period issues continue for linked data analysis, albeit in different form, because of the 'calendarization' methodology used to compile the RTI data (Office for National Statistics, 2019) which produces 'daily-ised' earnings estimates. In sum, the modelling approach taken in this paper will continue to be useful when linked survey-RTI data become available to researchers. It is also applicable to linked datasets for other countries.

## Acknowledgements

Thanks to the FRS team at the UK Department of Work and Pensions (DWP) for facilitating this project as part of their Secure Data Pilot initiative and for helpfully responding to queries. For helpful comments and suggestions, thanks also to Joint Editor Jouni Kuha and 2 anonymous referees, Dean Hyslop, Erik Meijer, Steve Pischke, and seminar participants at Autonoma Madrid, Leeds, Melbourne Institute, Motu Research, Paris School of Economics, and the 2022 RES Conference.

## Data availability

Data availability statement is at the end of the Introduction.

*Conflicts of interest:* None declared.

## Supplementary material

Online Supplementary material is available at *Journal of the Royal Statistical Society*, *Series A* (http://mtp.oxfordjournals.org/).

## References


Abowd J., & Stinson M. (2013). Estimating measurement error in annual job earnings: A comparison of survey and administrative data. *Review of Economics and Statistics*, *95*(5), 1451–1467. https://doi.org/10.1162/REST_a_00352

Angel S., Disslbacher F., & Humer S. (2019). What did you really earn last year? Explaining measurement error in survey income data. *Journal of the Royal Statistical Society, Series A*, *182*(4), 1411–1437. https://doi.org/10.1111/rssa.12463

Bee A., & Rothbaum J. (2019). *The administrative income statistics (AIS) project: Research on the use of administrative records to improve income and resource estimates, SEHSD Working Paper Number 2019-36*. U.S. Census Bureau.

Bingley P., & Martinello A. (2017). Measurement error in income and schooling and the bias of linear estimators. *Journal of Labor Economics*, *35*(4), 1117–1148. https://doi.org/10.1086/692539

Bohensky M. (2016). Bias in data linkage studies. In K. Harron, C. Dibben & H. Goldstein (Eds.), *Methodological developments in data linkage, ch. 4* (pp. 63–82). Wiley.

Bollinger C. R. (1998). Measurement error in the current population survey. *Journal of Labor Economics*, *16*(3), 576–594. https://doi.org/10.1086/209899

Bollinger C. R., Hirsch B. T., Hokayem C. M., & Ziliak J. P. (2018). *The good, the bad and the ugly: measurement error, non-response and administrative mismatch in the CPS. Working Paper, Gatton College of Business*. University of Kentucky.

Bollinger C. R., Hirsch B. T., Hokayem C. M., & Ziliak J. P. (2019). Trouble in the tails? What we know about earnings nonresponse 30 years after Lillard, Smith, and Welch. *Journal of Political Economy*, *127*(5), 2143–2185. https://doi.org/10.1086/701807

Bound J., Brown C., Duncan G. J., & Rodgers W. L. (1994). Evidence on the validity of cross-sectional and longitudinal labor market data. *Journal of Labor Economics*, *12*(3), 345–368. https://doi.org/10.1086/298348

Bound J., Brown C., & Mathiowetz N. (2001). Measurement error in survey data. In J. J. Heckman & E. Leamer (Eds.), *Handbook of econometrics, ch. 59*. (Vol. *5*), (pp. 3705–3843). North-Holland.

Bound J., & Kreuger A. (1991). The extent of measurement error in longitudinal data: Do two wrongs make a right? *Journal of Labor Economics*, *19*(1), 1–24. https://doi.org/10.1086/298256

Brewer M., Etheridge B., & O'Dea C. (2017). Why are households that report the lowest incomes so well-off? *Economic Journal*, *127*(605), F24–F49. https://doi.org/10.1111/ecoj.12334

Bricker J., & Engelhardt G. V. (2008). Measurement error in earnings data in the health and retirement study. *Journal of Economic and Social Measurement*, *33*(1), 39–61. https://doi.org/10.3233/JEM-2008-0297

Britton J., Shephard N., & Vignoles A. (2019). A comparison of sample survey measures of earnings of English graduates with administrative data. *Journal of the Royal Statistical Society, Series A*, *182*(3), 719–754. https://doi.org/10.1111/rssa.12382

Burke D. and Matejic P. (2018) *Family Resources Survey and related series – update and developments*. Family Finance Surveys User Conference presentation. https://dam.ukdataservice.ac.uk/media/621289/burkematejic.pdf

Celhay P., Meyer B. D., & Mittag N. (2021). *Errors in reporting and imputation of government benefits and their implications*. IZA Discussion Paper 14396. IZA.

Department for Work and Pensions. (2012). *Family resources survey user guide: household schedule, benefit unit schedule 2011-2012*. https://doc.ukdataservice.ac.uk/doc/7368/mrdoc/pdf/frs_2011_12_user_guide.pdf

Department for Work and Pensions. (2013). *Family resources survey United Kingdom, 2011/12*. Department for Work and Pensions.

Duncan G., & Hill D. (1985). An investigation of the extent and consequences of measurement error in labor economic survey data. *Journal of Labor Economics*, *3*(4), 508–522. https://doi.org/10.1086/298067

Golden L. (2016). *Still falling short on hours and pay. Part-time work becoming new normal*. Employment Policy Institute.

Gottschalk P., & Huynh M. (2010). Are earnings inequality and mobility overstated? The impact of nonclassical measurement error. *The Review of Economics and Statistics*, *92*(2), 302–315. https://doi.org/10.1162/rest.2010.11232

Harron K., Dibben C., Boyd J., Hjern A., Azimaee M., Barreto M. L., & Goldstein H. (2017). Challenges in administrative data linkage for research. *Big Data & Society*, *4*(2). https://doi.org/10.1177/2053951717745678

Hyslop D. R., & Townsend W. (2020). Earnings dynamics and measurement error in matched survey and administrative data. *Journal of Business and Economic Statistics*, *38*(2), 457–469. https://doi.org/10.1080/07350015.2018.1514308

Jenkins S. P., Lynn P., Jäckle A., & Sala E. (2008). The feasibility of linking household survey and administrative record data: New evidence for Britain. *International Journal of Social Research Methodology*, *11*(1), 29–43. https://doi.org/10.1080/13645570701401602

Jenkins S. P., & Rios-Avila F. (2020). Modelling errors in survey and administrative data on labour earnings: sensitivity to the fraction assumed to have error-free earnings. *Economics Letters*, *192*, 109253. https://doi.org/10.1016/j.econlet.2020.109253

Jenkins S. P., & Rios-Avila F. (2021a). *Finite mixture models for linked survey and administrative data: estimation and post-estimation*. IZA Discussion Paper 14404. IZA. Forthcoming in The Stata Journal.

Jenkins S. P., & Rios-Avila F. (2021b). *Reconciling reports: modelling employment earnings and measurement errors using linked survey and administrative data*. IZA Discussion Paper 14405. IZA.

Kapteyn A., & Ypma J. Y. (2007). Measurement error and misclassification: a comparison of survey and administrative data. *Journal of Labor Economics*, *25*(3), 513–551. https://doi.org/10.1086/513298

Kim C., & Tamborini C. R. (2014). Response error in earnings: An analysis of the survey of income and program participation matched with administrative data. *Sociological Methods and Research*, *43*(1), 39–72. https://doi.org/10.1177/0049124112460371

Kristensen N., & Westergaard-Nielsen N. (2007). A large-scale validation study of measurement errors in longitudinal survey data. *Journal of Economic and Social Measurement*, *32*(2-3), 65–92. https://doi.org/10.3233/JEM-2007-0283

Lunn S., & McKay S. (2013). Data linking the family resources survey with social security benefits. In M. Jäntti, V.-M. Törmälehto & E. Marlier (Eds.), *The use of registers in the context of EU–SILC: Challenges and opportunities, ch. 12* (pp. 161–174). Publications Office of the European Union.

McKay S. (2012). *Evaluating approaches to Family Resources Survey data linking. Working Paper 110*. Department for Work and Pensions.

Meijer E., Rohwedder S., & Wansbeek T. (2012). Measurement error in earnings data: using a mixture model approach to combine survey and register data. *Journal of Business and Economic Statistics*, *30*(2), 191–201. https://doi.org/10.1198/jbes.2011.08166

Moore J. C., Stinson L. L., & Welniak E. J. Jr. (2000). Income measurement error in surveys: A review. *Journal of Official Statistics*, *16*(4), 331–361.

Office for National Statistics. (2019). *New methods for monthly earnings and employment estimates from Pay As You Earn Real Time Information (PAYE RTI) data: December 2019*. Office for National Statistics.

Pedace R., & Bates N. (2000). Using administrative records to assess earnings reporting error in the survey of income and program participation. *Journal of Economic and Social Measurement*, *26*(3-4), 173–192. https://doi.org/10.3233/JEM-2000-0180

Pischke J. S. (1995). Measurement error and earnings dynamics: Some estimates for the PSID Validation Study. *Journal of Business and Economic Statistics*, *13*, 305–314.

Redner R. A., & Walker H. F. (1984). Mixture densities, maximum likelihood and the EM algorithm. *SIAM Review*, *26*(2), 195–239. https://doi.org/10.1137/1026034

Sakshaug J. W., & Kreuter F. (2012). Assessing the magnitude of non-consent biases in linked survey and administrative data. *Survey Research Methods*, *6*(2), 113–122.

Valet P., Adriaans J., & Liebig S. (2019). Comparing survey data and administrative records on gross earnings: nonreporting, misreporting, interviewer presence and earnings inequality. *Quality and Quantity*, *53*(1), 471–491. https://doi.org/10.1007/s11135-018-0764-z

Yakowitz S. J., & Spragins J. D. (1968). On the identifiability of finite mixtures. *Annals of Mathematical Statistics*, *39*(1), 209–214. https://doi.org/10.1214/aoms/1177698520

**Appendix Table A1.** Means: analysis sample versus FRS sample

| Characteristics | All employees in analysis sample | | All employees in FRS | |
|---|---|---|---|---|
| | Unweighted (1) | Weighted (IPW weights) (2) | Unweighted (3) | Weighted (FRS weights) (4) |
| Log(annualized employment earnings), *s* | 9.77 (0.81) | 9.78 (0.82) | 9.80 (0.81) | 9.81 (0.82) |
| Log(annual employment earnings), *r* | 9.75 (0.84) | 9.75 (0.85) | n/a | n/a |
| Male | 0.43 | 0.46 | 0.44 | 0.46 |
| Female | 0.57 | 0.54 | 0.56 | 0.54 |
| Age (years) | 44 | 41 | 42 | 41 |
| Aged < 60 years | 0.90 | 0.92 | 0.91 | 0.93 |
| Aged 60+ years | 0.10 | 0.08 | 0.09 | 0.07 |
| Education: less than A-level | 0.53 | 0.50 | 0.49 | 0.49 |
| Education: A-level or more | 0.47 | 0.50 | 0.51 | 0.51 |
| Employed full-time | 0.71 | 0.74 | 0.72 | 0.74 |
| Employed part-time | 0.29 | 0.26 | 0.28 | 0.26 |
| Married, cohabiting | 0.70 | 0.70 | 0.70 | 0.70 |
| Single, divorced, separated, widowed | 0.30 | 0.30 | 0.30 | 0.30 |
| Has 1 job | 0.96 | 0.96 | 0.96 | 0.96 |
| Has 2+ jobs | 0.04 | 0.04 | 0.04 | 0.04 |
| Payslip(s) consulted (all jobs) | 0.65 | 0.59 | 0.58 | 0.60 |
| Payslip not provided by employer (any job) | 0.10 | 0.11 | 0.11 | 0.11 |
| Pay reference period: | | | | |
| 1 week | 0.17 | 0.16 | 0.17 | 0.16 |
| 4 weeks or 1 calendar month | 0.77 | 0.77 | 0.77 | 0.77 |
| 1 year, 12 months, or 52 weeks | 0.04 | 0.05 | 0.05 | 0.05 |
| Other | 0.02 | 0.02 | 0.02 | 0.02 |
| Job spells in P14 data do not span 2011/12 | 0.39 | 0.39 | n/a | n/a |
| Job spells in P14 data all span 2011/12 | 0.61 | 0.61 | n/a | n/a |
| Private sector employee | 0.69 | 0.72 | 0.69 | 0.70 |
| Public sector employee | 0.31 | 0.28 | 0.31 | 0.30 |
| Non-White ethnic group | 0.07 | 0.10 | 0.08 | 0.10 |

*Notes*. Columns 1 and 2 refer to the main analysis sample of 'all individuals' (unweighted $N = 5{,}971$), as described in the main text. The weights used to derive column 2 are the composite IPW weights described in the main text and Online Supplementary Material (and also used to derive the estimates reported in the main text). Columns 3 and 4 refer to a sample of all FRS respondents reporting at least 1 employment (i.e., not conditioning on consent or successful linkage). For comparisons with the analysis sample, the calculations in columns 3 and 4 exclude respondents with a job who have imputed wages or declare themselves to be self-employed (unweighted $N = 12{,}518$). The weights used to derive column 4 are the standard FRS individual respondent survey weights ('gross3'). n/a: P14 information is not available for nonconsenting nonlinked respondents. For *s* and *r*, the numbers in parentheses are the respective SDs.