A comparison of sample survey measures of earnings of English graduates with administrative data

Jack Britton (IFS), Neil Shephard (Harvard) and Anna Vignoles (Cambridge)

Royal Statistical Society Discussion Meeting
October 2018

Introduction

- Administrative data is increasingly being used in economic analyses
- Can help understand problems in survey data (Bound et al., Abowd and Stinson, Koijen et al.)
  - Comprehensive coverage, longitudinal, potentially clear measures
  - Cannot assume administrative data is always “the truth”
- Power of administrative tax records is clear (e.g. Chetty et al.)

Introduction

- Aim to build a new database (the "Golden Sample") of the earnings of graduates
  - Significant policy interest in graduate earnings, particularly post tuition fee increases
  - Limited data on how graduate earnings vary by subject or institution
  - Labour Force Survey has graduate earnings, but limited to variation by subject (Walker and Zhu)
- Need to compare and reconcile literature using LFS and administrative data (which will be used by DfE and others going forward)

Literature

- Literature on problems measuring income in surveys (Bound et al. Abowd and Stinson, Koijen et al).
  - Difficult measuring income of low, hourly and intermittently paid (Skinner et al.)
  - LFS used weekly pay divided by usual hours. Now uses hourly pay but missing rate is high.
  - Mixed evidence on whether pay and measurement error are positively correlated (Bound et al. Rodgers)
- Literature on variation in graduate earnings (Naylor et al. Walker and Zhu), inequalities in earnings (Cunha and Heckman) and gender wage gap (Machin and Puhani, Chevalier)

Disclaimers

- HM Revenue & Customs (HMRC) agrees that the figures and descriptions of results in the attached document may be published. This does not imply HMRC's acceptance of the validity of the methods used to obtain these figures, or of any analysis of the results. Copyright of the statistical results may not be assigned. This work contains statistical data from HMRC which is Crown Copyright. The research datasets used may not exactly reproduce HMRC aggregates. The use of HMRC statistical data in this work does not imply the endorsement of HMRC in relation to the interpretation or analysis of the information.

- The Student Loans Company (SLC) agrees that the figures and descriptions of results in the attached document may be published. This does not imply SLC's acceptance of the validity of the methods used to obtain these figures, or of any analysis of the results. Copyright of the statistical results may not be assigned. This work contains statistical data from SLC which is protected by Copyright, the ownership of which is retained by SLC. The research datasets used may not exactly reproduce SLC aggregates. The use of SLC statistical data in this work does not imply the endorsement of SLC in relation to the interpretation or analysis of the information.

Data

- Hard linked data sets using NIs
- Student Loan Company (SLC)
  - UK domiciled students who borrow (85-90% of those eligible)
  - Includes drop outs (10%)
- 10% random sample of Pay As You Earn (PAYE) - 90% of income tax
- Self Assessment (SA) - regarded as definitive
  - % SA increases with age
- 263k sample, cohorts 1998-2011 earnings 2008/9-2012/13

Data

	Overall	Men	Women
1998	14,487	6,927	7,560
1999	22,621	10,590	12,031
2000	23,506	10,853	12,653
2001	23,924	11,025	12,899
2002	23,891	11,060	12,831
2003	23,972	11,024	12,948
2004	23,577	10,767	12,810
2005	25,103	11,439	13,664
2006	25,383	11,340	14,043
2007	25,352	11,292	14,060
2008	20,847	8,990	11,857
2009	6,510	3,029	3,481
2010	2,993	1,334	1,659
2011	851	360	491
All	263k	120k	143k

Table: Number of Golden sample (10% sample of loan database) borrowers and tax data in 2011-12. PAYE (Pay As You Earn) and SA (self-assessment) denotes databases. Either denotes being in either PAYE or SA or both. Cohort denotes the first year the borrower received a loan from the SLC.

Data

- Focus on earned income (employment, profits from partnerships and self employment)
- No report is equated to zero earnings
- May miss very low earners
- Miss most of those who move abroad

Low/no earners

		% No tax form			% Earnings = £0 (or no form)			% Earnings $<$ £8,000 (includes 0s & missings)		
Median age	Cohort	All	Male	Female	All	Male	Female	All	Male	Female
31	1998	13.0	12.6	13.3	15.6	15.2	16.0	27.3	26.7	27.9
30	1999	11.7	11.4	11.9	14.4	14.4	14.5	26.2	25.7	26.7
29	2000	11.4	11.2	11.5	14.2	14.1	14.2	26.1	25.7	26.5
28	2001	10.1	9.9	10.3	13.0	12.7	13.2	25.0	24.5	25.5
27	2002	9.6	9.9	9.3	12.5	12.8	12.2	25.3	25.5	25.0
26	2003	9.0	8.9	9.0	12.0	11.8	12.2	25.8	25.4	26.1
25	2004	8.0	8.3	7.7	10.9	11.5	10.5	25.9	26.8	25.2
24	2005	7.5	7.4	7.5	10.8	11.0	10.6	29.1	30.3	28.2
23	2006	7.5	7.8	7.2	11.0	11.6	10.5	34.3	36.3	32.6
22	2007	7.0	7.8	6.3	10.5	11.6	9.6	43.2	45.1	41.8
21	2008	8.4	9.1	7.8	11.6	12.4	11.0	61.6	63.2	60.4
21	2009	10.9	11.6	10.4	15.8	17.2	14.5	61.1	64.6	58.0
20	2010	11.0	12.0	10.2	16.1	17.5	15.0	67.9	72.0	64.6
18	2011	10.1	13.1	7.9	14.9	18.3	12.4	90.6	90.6	90.6

Table: Golden Sample for 2011-12. % of individuals with no filed income form & the % with no or low earnings. Columns are cumulative so the share with earnings $<$ £8,000 includes those with earnings $=$ £0 and those with no form. Median age does not decrease by one each year in the GS because of small sample sizes and variation in the ages of HE leavers (since individuals only enter our dataset once they have left HE).

Data

- Starkly high proportion with low earnings
- HMRC (2014) estimate under-reporting: 25% of SA and 1.5% of PAYE
- 10% of sample fully or partially self employed

Data

- Also have information on non borrowers (most of whom will not be graduates)
- Even starker % of low earners
- Need to reweight earnings to allow for graduates who do not borrow

Non graduate "Silver Sample"

		% No tax form			% Earnings = £0 (or no form)			% Earnings < £8,000 (includes zeros & missings)		
Median age	Cohort	All	Male	Female	All	Male	Female	All	Male	Female
31	1998	22.1	21.5	23.0	27.3	26.7	27.9	46.3	43.3	49.9
30	1999	22.6	21.3	24.2	27.7	26.6	29.0	47.5	43.8	51.9
29	2000	23.5	21.8	25.5	28.5	27.0	30.4	48.8	45.2	53.2
28	2001	24.3	22.4	26.5	29.1	27.6	31.0	49.7	46.1	54.0
27	2002	24.8	23.1	26.8	29.7	28.3	31.4	51.2	47.9	55.1
26	2003	25.0	23.2	27.2	29.9	28.2	31.9	51.9	48.5	55.8
25	2004	24.9	22.7	27.5	30.1	28.1	32.5	52.9	49.8	56.6
24	2005	24.2	21.8	27.0	29.3	27.3	31.7	53.8	51.2	56.9
23	2006	23.7	21.4	26.4	29.0	26.9	31.4	55.8	53.4	58.6
22	2007	22.8	20.3	25.6	28.2	25.9	30.9	58.6	55.7	61.9
21	2008	21.6	19.4	24.1	27.8	25.4	30.5	61.6	59.0	64.5
21	2009	20.4	19.5	21.3	26.4	25.6	27.3	64.2	62.0	66.7
20	2010	18.4	17.1	19.9	24.4	23.1	25.8	68.8	66.0	71.8

Table: Silver Sample database for 2011-12. % with no filed income tax form and % with no and low earnings. Median age does not decrease by one each year in the SS because the age distribution is matched exactly to the GS (see footer to Table 2).

Data

- Labour Force Survey April 2011 - March 2012
- Matched sample on cohort (age based), gender, graduate status
- England domiciled currently (not at time of enrollment in HE)
- Response rate to earnings question 70% - includes proxy earnings
- Pool cohorts to increase sample size

Key differences between administrative data and LFS

- Majority of potential biases are downward for GS and upward for LFS
  - GS is legally required measure of earnings on which tax is levied
  - LFS annualises earnings from shorter periods, excludes those with very variable earnings
  - GS includes drop outs and excludes non borrowers
  - GS includes those who move abroad (zero earnings)
  - GS includes the self employed, LFS does not
  - LFS respondents may include pension contributions

Key differences between administrative data and LFS

- Share on zero earnings is higher in the administrative data for men
- Share on low earnings is much higher in the administrative data

LFS earnings higher even conditional on earnings above £8k

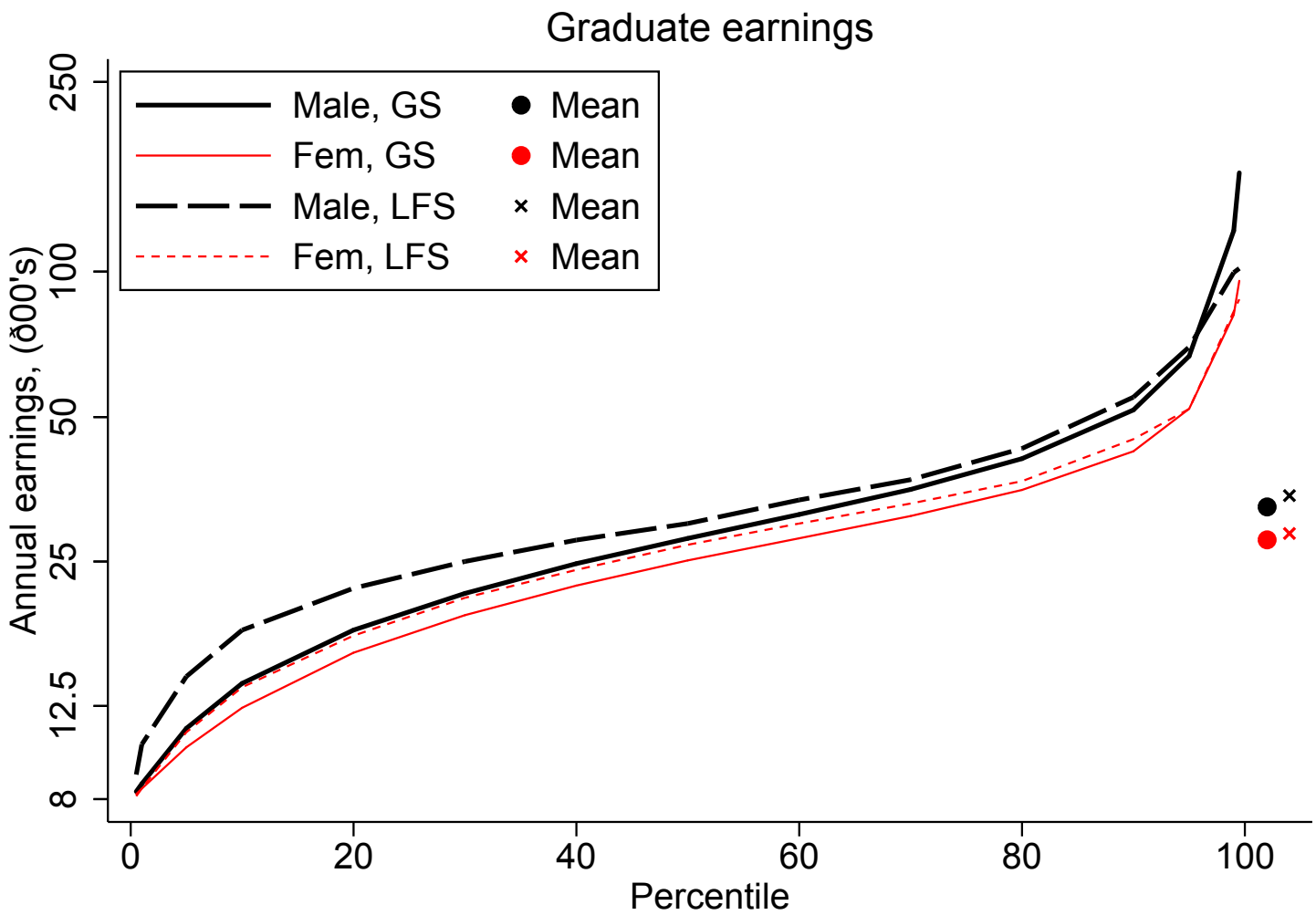
	% Not employed			% Earnings $<$ £8,000 given Earnings$>$£0		
Year	All	M	F	All	M	F
LFS						
2008/09	10.4	8.7	12.0	4.6	3.2	6.1
2009/10	11.3	9.3	13.1	5.3	3.6	7.0
2010/11	11.2	8.0	14.1	5.4	2.8	7.9
2011/12	13.0	8.9	16.5	4.5	1.7	7.4
2012/13	12.5	6.9	17.5	5.9	3.6	8.1
Golden Sample						
2008/09	11.4	11.9	10.9	13.8	13.9	13.8
2009/10	13.8	14.0	13.7	13.8	14.2	13.4
2010/11	13.4	13.6	13.2	13.5	13.3	13.7
2011/12	13.5	13.4	13.5	13.3	12.7	13.7
2012/13	14.6	14.4	14.7	13.9	12.5	15.2

Table: LFS and Golden Sample: graduates not employed and with low earnings overall and by gender for 2008/09 through 2012/13. The 1998-2003 cohort are pooled for each year. LFS population weights are applied (the "pwt" weight for the unemployed share, while the "piwt" weight for the earnings share).

Key differences between administrative data and LFS

- LFS earnings higher through the distribution up to 90th percentile
- Conditional on working, LFS earnings around 20% higher for males, slightly less for women

Graduate Comparison 2011/12, earnings>£8k



2011/12 data, 1998-2003 cohorts pooled

Possible explanations for differences

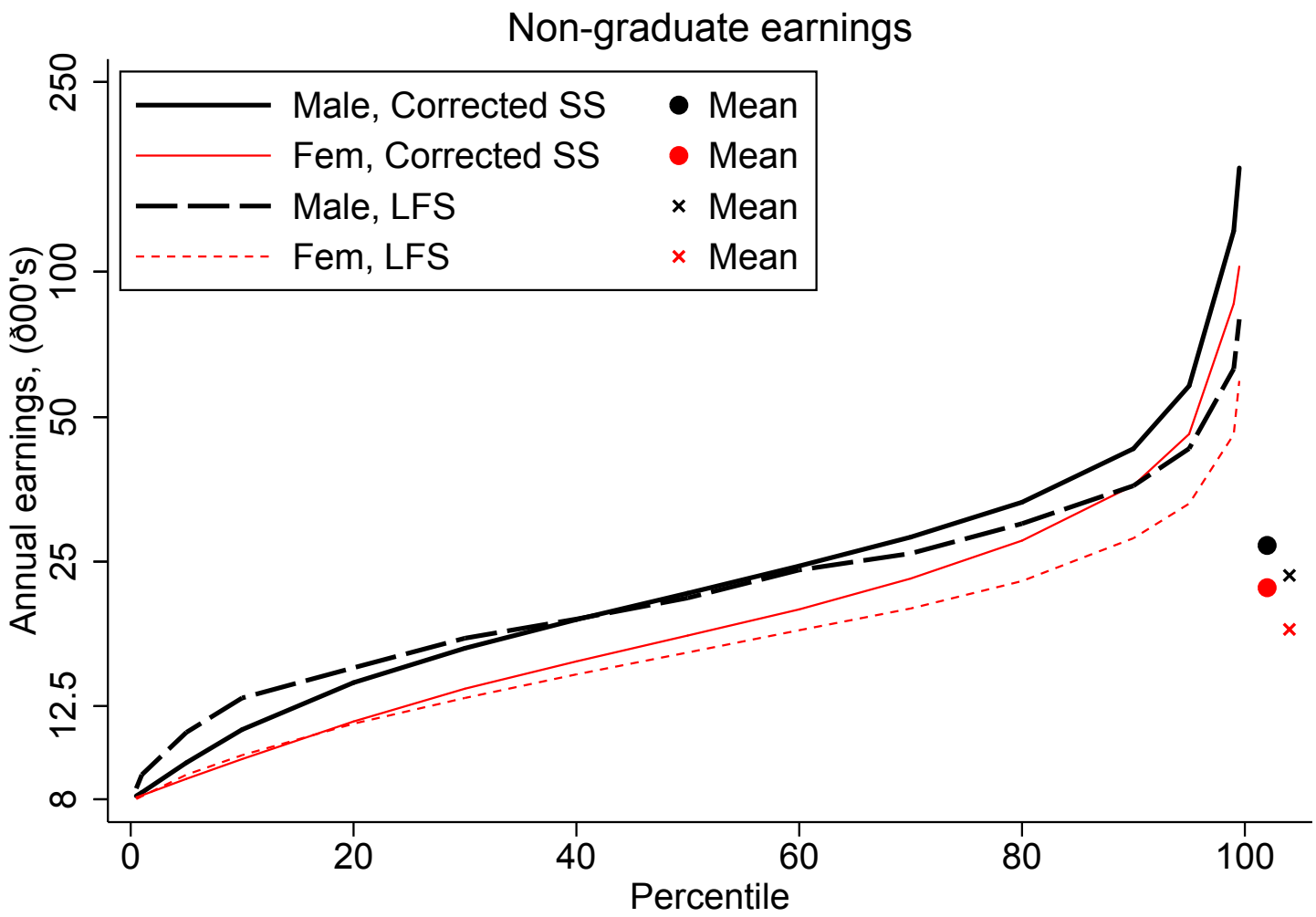
- Lower earners (particularly men) not responding to LFS
- Measurement error in LFS earnings variable
- LFS excludes self employed (but can only explain some of the difference)
- Under-reporting in admin data to avoid tax (esp men in second jobs)

Non-graduate comparison between LFS and administrative data

- Above £8k, the difference between admin data and LFS is similar to those observed in the graduate sample

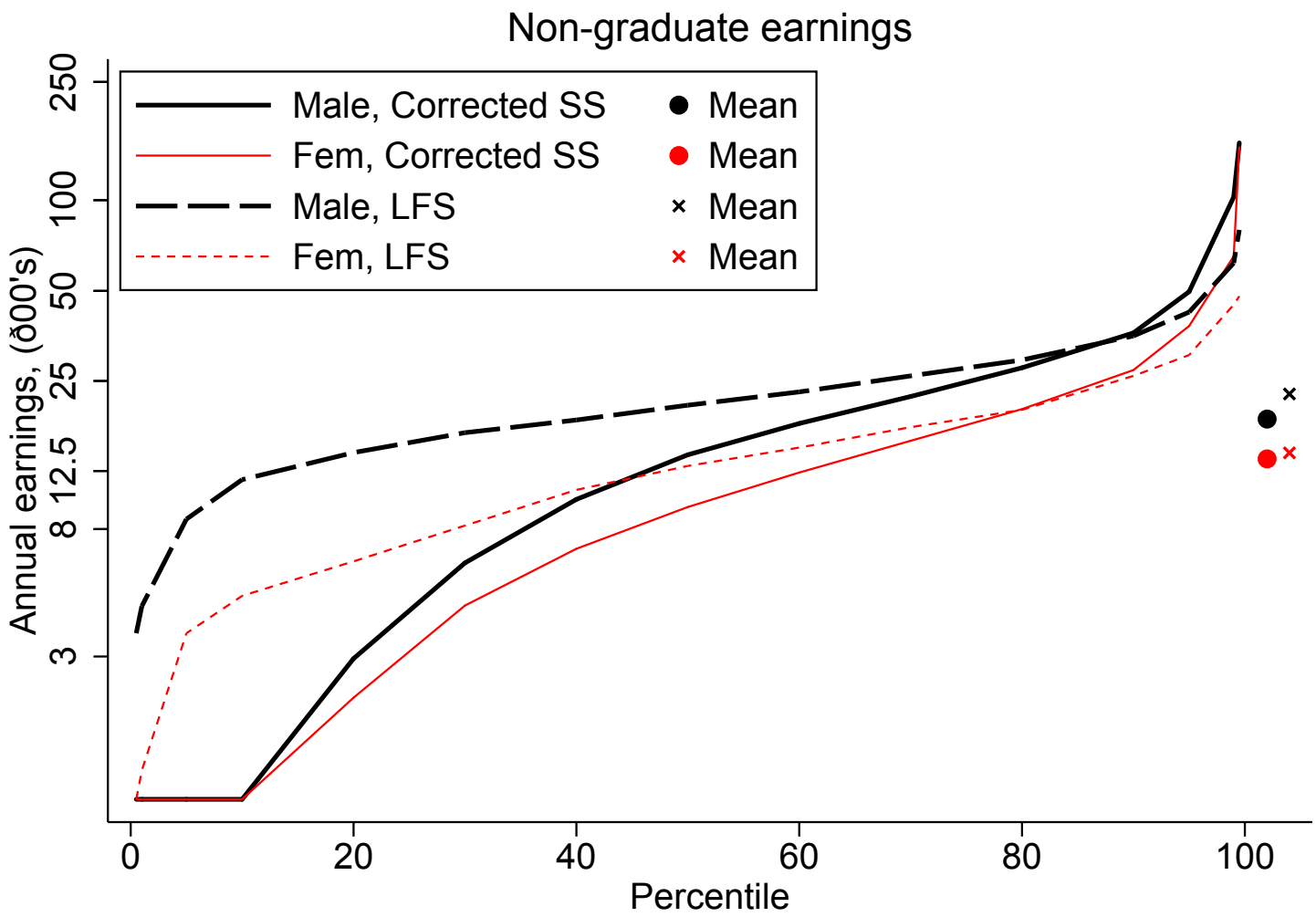
- Focusing on full distribution, LFS earnings are higher for males particularly at the bottom of the distribution

Non-graduate Comparison 2011/12, earnings>£8k



2011/12 data, 1998-2003 cohorts pooled

Non-graduate Comparison 2011/12, earnings>£0

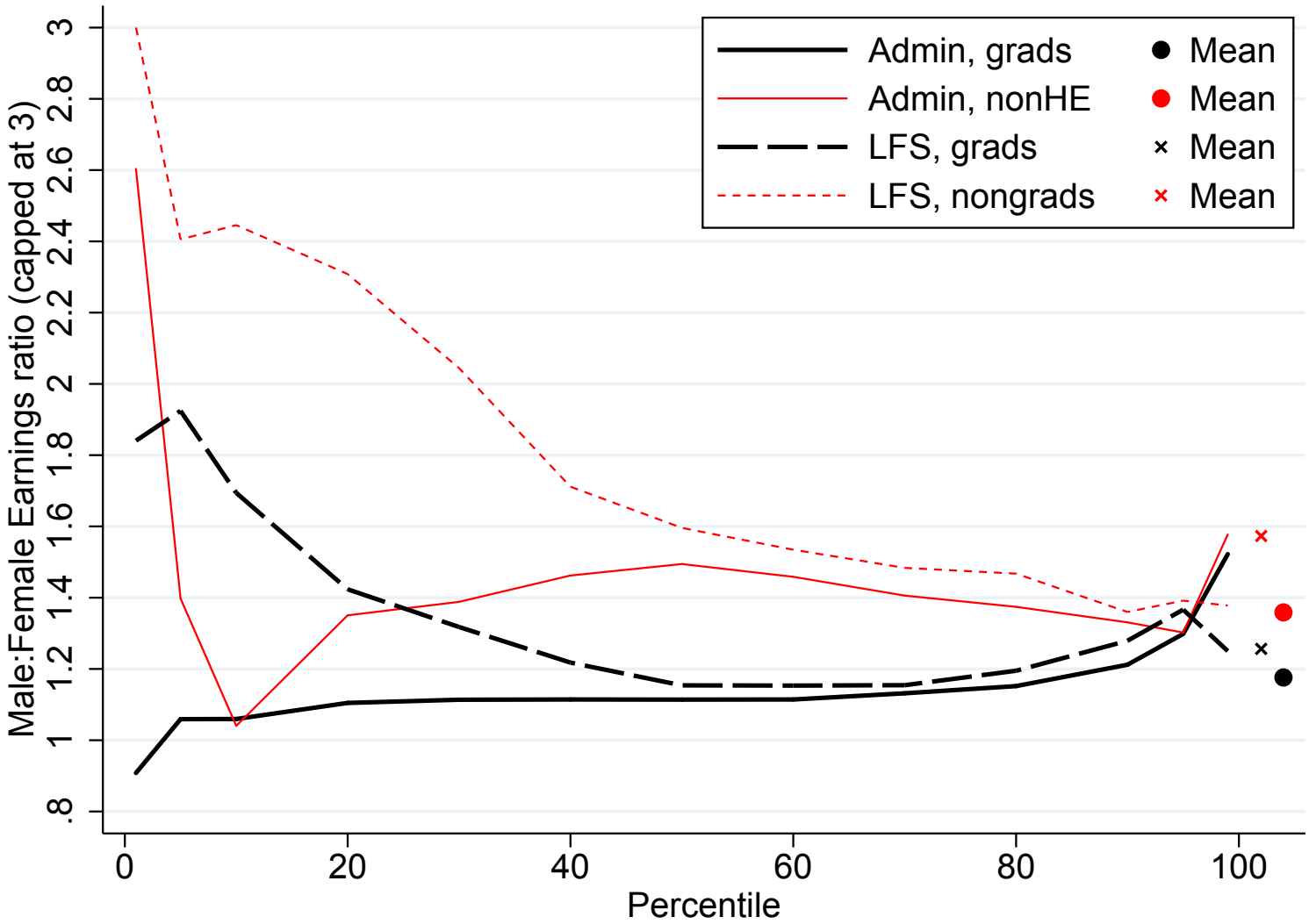


2011/12 data, 1998-2003 cohorts pooled

Gender pay gap

- Pay gap is larger for non graduates
- Administrative data shows far smaller wage gap at bottom of distribution

Gender pay gaps through the distribution, 2011/12 data

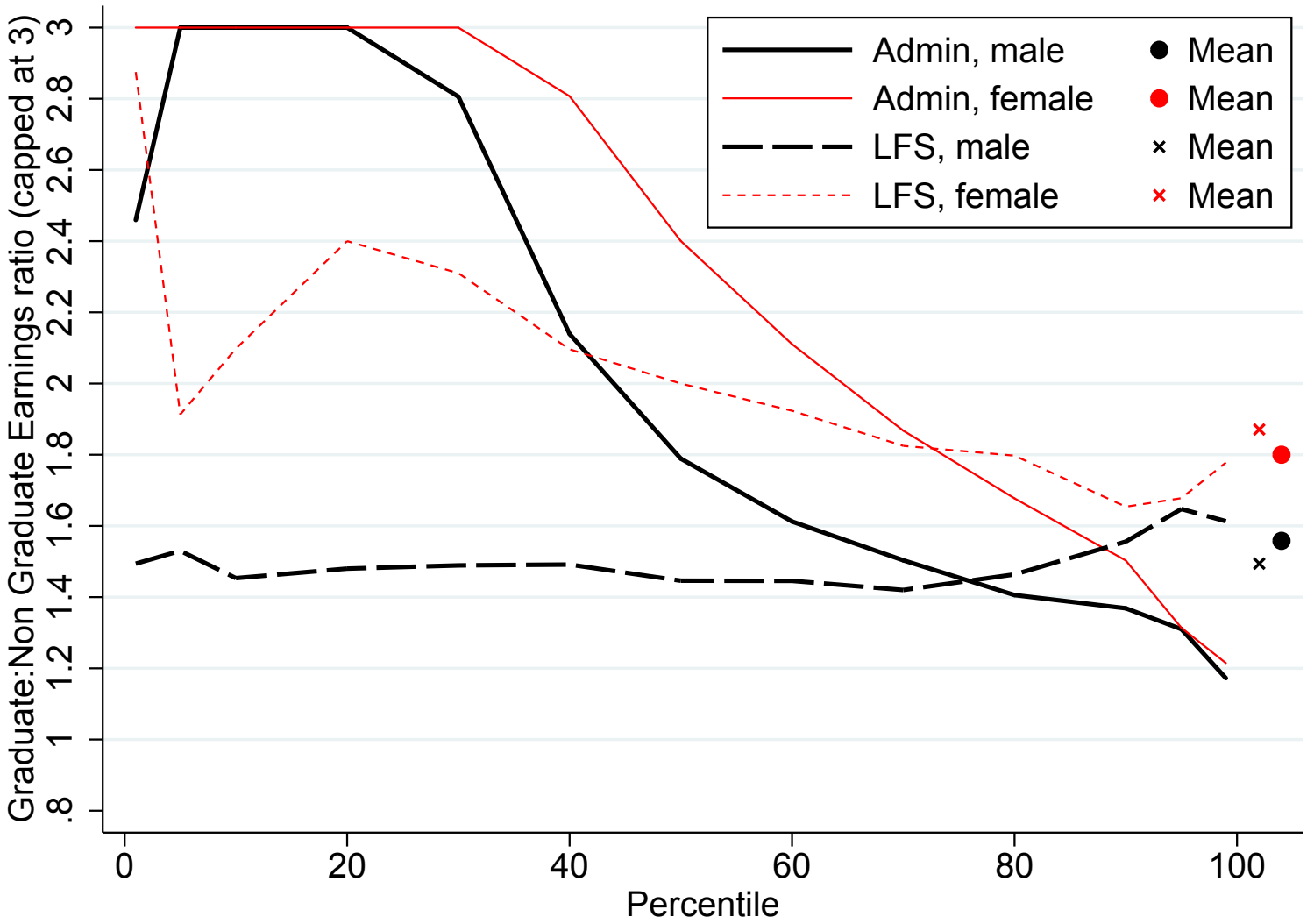


Male:female earnings ratio, 1998-2003 cohorts pooled

Graduate wage premium

- Graduate wage premium is larger in the administrative data at bottom of distribution
- Graduate wage premium is 1.8 for females and 1.5 for males at end of period
- Graduate wage premium increased over period

Graduate wage premium through the distribution, 2011/12 data



Graduate:Non-graduate, 1998-2003 cohorts combined

Earnings inequality

- Earnings inequality increases over period
- Much less earnings inequality in the LFS than in the administrative data
- Earnings inequality much greater for non graduates

Earnings inequality

	Graduates				Non-Graduates			
	Men		Women		Men		Women	
Year	GS	LFS	GS	LFS	CSS	LFS	CSS	LFS
2008/09	0.360	0.259	0.332	0.263	0.505	0.246	0.509	0.304
2009/10	0.372	0.252	0.337	0.268	0.510	0.240	0.515	0.319
2010/11	0.380	0.268	0.347	0.272	0.525	0.255	0.520	0.355
2011/12	0.388	0.262	0.358	0.283	0.523	0.255	0.540	0.340
2012/13	0.395	0.283	0.375	0.287	0.528	0.251	0.540	0.370

Table: Gini coefficients for the administrative data and LFS positive earnings distributions, split by gender and graduate status for 2008/09 - 2012/13. Each observation includes the 1998-2003 cohorts. The 'GS' (Golden Sample) and 'CSS' (Corrected Silver Sample) show the administrative data.

Summary of findings

- Both data sets show similar share of individuals with zero earnings
- Higher share in the administrative data earning below £8k
- LFS has higher earnings through the distribution
- Findings similar for graduates and non-graduates - suggesting problem may be measurement in LFS not graduate sample construction

Conclusion

- Implications for research and policy are significant
- Different data sets suggest different conclusions about the gender wage gap, graduate wage premium and inequality
- Admin record should be more reliable and is official record on which loan repayments are calculated
- Further work on under-reporting in the tax record AND problems in the survey data (sample selection, measurement error)

Appendix

Additional information - those who go abroad

Cohort	% abroad	% been abroad	Of those abroad		Of those been abroad	
			Earn= £0	Earn<£8,000	Earn=£0	Earn<£8,000
1998	1.0	4.4	73.2		40.8	50.6
1999	1.1	4.2	69.8	86.8	42.7	52.2
2000	1.2	4.0	61.7	80.7	42.0	53.9
2001	1.2	4.1	66.2	78.5	42.5	52.2
2002	1.3	3.8	58.4	78.1	39.5	52.1
2003	1.4	3.7	57.0	73.8	40.8	55.1
2004	1.4	3.8	47.7	71.8	36.4	54.9
2005	1.4	3.4	55.9	82.6	39.3	61.9
2006	1.4	2.6	51.2	85.3	43.1	73.5
2007	1.4	1.9	43.8	86.0	44.0	82.9
2008	0.6	0.7	34.1		35.3	

Table: SLC in repayment and living abroad data in 2011/12. Abroad is an indicator for being overseas and in repayment according to SLC records. Been abroad is an indicator for abroad and in repayment *or* have been in this state at some point in the past. Figures are excluded where implied sample sizes are too small.

Additional information - the self employed

		Only partly self-employed						Entirely self-employed					
Median age	Cohort	Of all (%)			Of SE part: % earnings $<$ £8,000			Of all (%)			Of SE only: % earnings $<$ £8,000		
		All	M	F	All	M	F	All	M	F	All	M	F
31	1998	6.4	7.1	5.7	33.4	27.1	40.7	3.6	4.4	2.8	44.9	35.4	58.8
30	1999	6.5	7.3	5.8	34.6	30.3	39.3	3.8	4.5	3.1	46.4	39.4	55.3
29	2000	6.6	7.5	5.8	33.8	31.7	36.1	3.7	4.6	2.9	46.7	42.9	51.9
28	2001	6.2	7.5	5.1	34.3	31.7	37.5	3.5	4.7	2.5	47.6	43.4	54.4
27	2002	5.8	6.9	5.0	35.9	35.5	36.3	3.3	4.3	2.4	47.2	46.3	48.7
26	2003	5.4	6.1	4.8	37.9	33.9	42.1	3.0	3.6	2.5	52.1	46.6	58.9
25	2004	5.2	6.2	4.3	38.8	36.2	41.9	2.8	3.6	2.1	51.7	47.7	57.6
24	2005	4.9	5.9	4.1	41.3	41.5	41.1	2.6	3.3	2.0	58.3	55.3	62.6
23	2006	4.3	5.1	3.7	47.6	46.6	48.8	2.2	3.0	1.5	63.1	59.7	68.5
22	2007	3.8	4.4	3.4	54.7	50.7	58.8	1.9	2.5	1.5	68.2	61.2	77.7
21	2008	3.1	3.6	2.7	67.9	68.3	67.5	1.7	2.4	1.2	78.1	78.0	78.1
21	2009	3.4	4.0	3.0	62.5	63.3	61.5	1.8	2.2	1.4	85.2	86.4	83.7
20	2010	2.8	3.4	2.4	61.9			1.3			87.5		

Table: Golden Sample self-employment: cohort who are only partially self-employed (not those fully self-employed) and those entirely self-employed. Also given are the corresponding % who have low earnings. Earnings is all earnings from work, not just from the self-employed part. Results are for the 2011-12 tax year. See footer to Table 2 to explain pattern for median age.

Additional information - reweighting the non graduate sample

$$F_S(y) = \omega F_{HE}(y) + (1-\omega)\, F_{nonHE}(y), \quad \omega \in [0,1].$$

Under these assumptions, for $y \in R_{\geq 0}$,

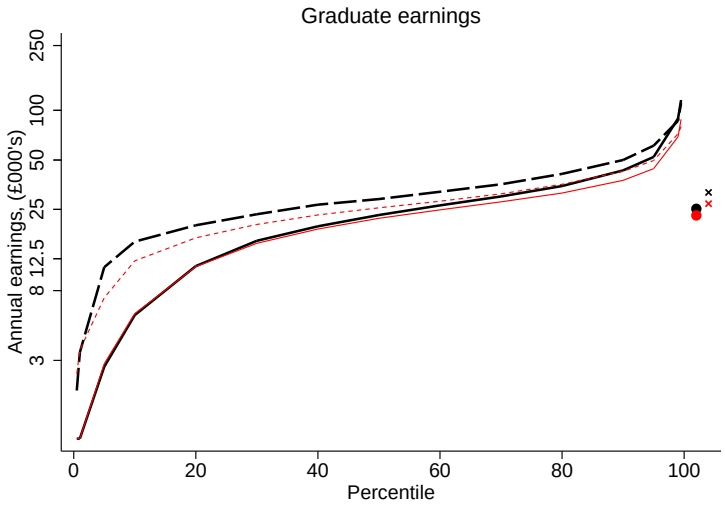
$$F_{nonHE}(y) = \frac{F_S(y) - \omega F_G(y)}{(1-\omega)}, \quad \mathrm{E}_{nonHE}(Y) = \frac{\mathrm{E}_S(Y) - \omega \mathrm{E}_G(Y)}{(1-\omega)}, \quad f_{nonHE}(y) = \frac{f_S(y) - \omega f_G(y)}{(1-\omega)}.$$

Table of differences between data sets

	LFS graduates	Golden Sample
Definition of graduates	Those whose highest qualification is at graduate degree level. For majority this is “higher degree” or “first degree”.	Those who borrowed from the SLC. Includes those who borrowed and failed to complete degree. Excludes those who did not borrow.
Population	Graduates living in England at the point of survey who are surveyed and respond. Those with ‘variable’ earnings and those not in households excluded.	10% sample of English-domiciled (on application) borrowers from the SLC. Includes those never in contact with HMRC and those living outside England.
Definition of cohort	Allocated based on age on August 31 in a given year.	Observed year started borrowing.
Earnings	Gross weekly earnings in first and second job combined, multiplied by 52. Weekly earnings are imputed in the survey based on a response period chosen by the individual.	PAYE & SA reported annual labour income. Individuals are legally required to report.
Pensions	Employer contributions usually excluded. Employee contributions usually included.	Employer & employee contributions excluded.
Proxy responses	Included (although this has limited impact on the qualitative conclusions of the paper).	Not applicable
Self-employment	Included, but with no earnings data.	Included.

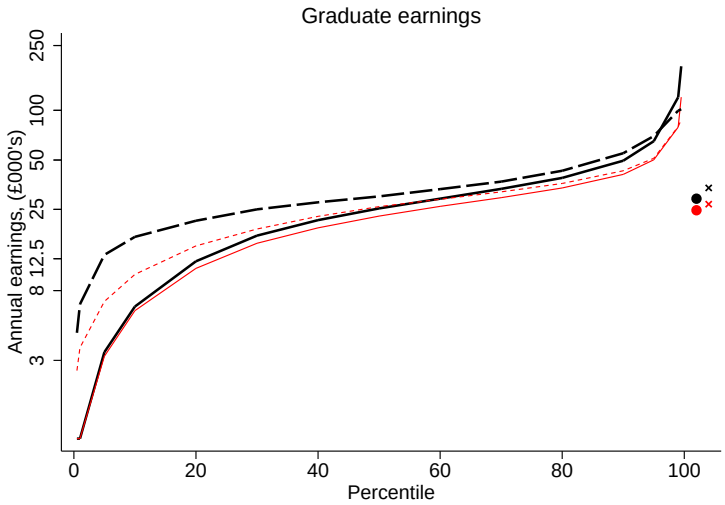
Table: Summary of differences between the LFS graduate and the Golden Sample datasets

Additional information - full earnings distribution



2008/09

data, all earnings



201

data, all earnings

Figure: Mimics Figure 22 but with the full earnings distribution, not including zeros. See Figure 22 for legend. Precise numbers given in the Online Appendix **??**, along with other years of data.

Additional information: SS vs LFS

	% Not employed						% Earnings < £8,000 given Earnings>£0					
	Sample size			Share			Sample size			Share		
Year	All	M	F	All	M	F	All	M	F	All	M	F
LFS												
2008/09	16,326	7,759	8,567	27.0	17.2	37.3	2,942	1,484	1,458	11.7	4.2	20.9
2009/10	14,920	7,021	7,899	30.1	21.0	40.0	2,594	1,326	1,268	13.2	4.8	24.5
2010/11	14,382	6,925	7,457	28.2	17.5	40.1	2,430	1,261	1,169	13.0	4.1	25.5
2011/12	14,041	6,828	7,213	28.3	18.1	39.8	2,481	1,278	1,203	13.9	4.3	28.8
2012/13	10,495	5,082	5,413	27.6	16.3	40.1	1,778	946	832	15.4	5.4	29.9
HMRC non-HE												
2008/09	243,099	132,522	110,577	29.1	28.5	30.0	172,245	94,816	77,429	37.4	33.7	42.0
2009/10	243,099	132,522	110,577	31.4	30.1	32.9	166,805	92,588	74,217	37.6	34.5	41.5
2010/11	243,099	132,522	110,577	29.9	28.8	31.2	170,451	94,392	76,059	38.0	34.7	42.0
2011/12	243,099	132,522	110,577	29.0	27.7	30.6	172,581	95,870	76,711	38.9	34.7	44.3
2012/13	243,099	132,522	110,577	28.8	27.6	30.2	173,102	95,966	77,136	38.7	33.9	44.7

Table: LFS and Corrected Silver Sample: non-graduates not employed and with low earnings overall and by gender for 2008/09 through 2012/13. The 1998-2003 cohort are pooled for each year. Note the share not employed in the Corrected Silver Sample is equal to the share no employed in the Silver Sample as the econometric correction only corrects the positive earnings distribution.