Britton, Jack; Shephard, Neil; Vignoles, Anna

Working Paper

Comparing sample survey measures of English earnings of graduates with administrative data during the Great Recession

IFS Working Papers, No. W15/28

Provided in Cooperation with:
Institute for Fiscal Studies (IFS), London

Suggested Citation: Britton, Jack; Shephard, Neil; Vignoles, Anna (2015) : Comparing sample survey measures of English earnings of graduates with administrative data during the Great Recession, IFS Working Papers, No. W15/28, Institute for Fiscal Studies (IFS), London, https://doi.org/10.1920/wp.ifs.2015.1528

This Version is available at:
https://hdl.handle.net/10419/145443

Institute for Fiscal Studies

Comparing sample survey measures of English earnings of graduates with administrative data during the Great Recession

IFS Working Paper W15/28

Jack Britton
Neil Shephard
Anna Vignoles

Nuffield Foundation

1965 - 2015 50 50 YEARS SHAPING SOCIETY E·S·R·C

Comparing sample survey measures of English earnings of graduates with administrative data during the Great Recession*

Jack Britton
Institute for Fiscal Studies, London
jack_b@ifs.org.uk

Neil Shephard
Department for Economics and Department of Statistics, Harvard University
shephard@fas.harvard.edu

Anna Vignoles
Department of Education, University of Cambridge
av404@cam.ac.uk

September 23, 2015

Abstract

This paper compares survey based labour earnings data for English graduates, taken from the UK's Labour Force Survey (LFS), with the UK Government administrative sources of official individual level earnings data. This type of administrative data has few sample selection issues, is substantially longitudinal and its large samples mean the earnings of subpopulations can be potentially studied (e.g. those who study a specific subject at a specific university and graduate in a specific year). We find that very broadly the LFS and administrative data show a similar distribution of graduates' earnings. However, the administrative data has considerably less gender disparity, higher high quantiles and more time series persistence. We also report on how the distribution of graduate and non-graduate earnings fell during each year of the great recession.

Keywords: Administrative data; Graduate earnings; Human capital; Labour earnings; Labour force survey; Quantile regression; Student loans.

*Many civil servants and policy makers have helped us gain access to the data which is the core of this paper. Although it is difficult to pick out a small group who helped most, we must thank in particular Daniele Bega, Dave Cartwright, Nick Hillman, Tim Leunig and David Willetts for this work would not have seen the light of day without each of their contributions. In addition, we also thank Anthony Atkinson, Raj Chetty, Jonathan Cribb, Mark Gittoes, Chuka Ilochi and Jonathan Waller for their comments on a previous draft, and our advisory group, Alison Allden, Nick Barr, Danny Dorling, Josh Hillman, Robin Naylor, Kate Purcell and Ian Walker. Of course we solely are responsible for any errors. We are grateful to the Nuffield Foundation for their financial support. The Nuffield Foundation is an endowed charitable trust that aims to improve social well-being in the widest sense. It funds research and innovation in education and social policy and also works to build capacity in education, science and social science research. The Nuffield Foundation has funded this project, but the views expressed are those of the authors and not necessarily those of the Foundation. More information is available at www.nuffieldfoundation.org. HM Revenue & Customs (HMRC) and Student Loans Company (SLC) have agreed that the figures and descriptions of results in the attached document may be published. This does not imply HMRC's or SLC's acceptance of the validity of the methods used to obtain these figures, or of any analysis of the results. The research datasets used may not exactly reproduce HMRC or SLC aggregates. The use of HMRC or SLC statistical data in this work does not imply the endorsement of either HMRC or SLC in relation to the interpretation or analysis of the information.

1 Introduction

1.1 The core of the paper

A rich literature has shown the power of big data, particularly administrative tax records, to help better understand the level and distribution of earnings of subpopulations in developed economies (e.g. Chetty et al. (2014a,b)). Such data have comprehensive coverage, clearly defined income categories and individual (or household) level data that stretches over significant periods of time.

In this paper we build and document a new database of tax records for individual English graduates. We compare the database's summary statistics with corresponding results from several well established Government funded labour market sample surveys. We demonstrate that this new data source gives sensible country level results and find that it would be credible for subsequent researchers to look at the earnings of subpopulations of graduates (e.g. those who study a specific subject at a specific university and graduate in a specific year) where existing survey data is too sparse to deliver useful results without employing ad hoc and very tightly parametrized statistical models.

Very broadly, we find that established survey data and this new administrative data source show a remarkably similar distribution of graduates' earnings. However, the administrative data has considerably larger very high quantiles of earnings, less gender disparity and more time series persistence. Each of these three findings is important from policy viewpoints. We also report how the distribution of graduate real earnings changed during the great recession, showing a very substantial fall in cohort adjusted average earnings, and compare their earnings changes with non-graduates, whose age adjusted earnings fell proportionally even more. Graduate and non-graduate women were particularly badly hit by the recession suffering proportionally more than the corresponding men. We show that median earnings of English women around 10 years out from higher education is roughly 3.5 times that of the median for those who did not attend higher education. The corresponding male ratio is roughly 2.2. These multiples, for both genders, are lower for higher quantiles (the higher paid) and higher for lower quantiles (the lower paid). This shows that graduates have not just much higher average earnings, but have much less individual level earnings risk and subpopulation inequality. These differences are particularly stark for women.

1.2 Existing survey data

In the UK context, there are a number of well established surveys that include sufficient numbers of graduates to be used to examine the graduate earnings distribution within a country. Many papers that have analyzed graduate earnings have relied on the UK's "Labour Force Survey" (LFS). For example, a number of influential studies have used LFS data to analyze how graduate earnings vary

by subject of degree (e.g. Sloane and O'Leary (2005) and Walker and Zhu (2011)).

In this paper we compare the graduate earnings distribution obtained from the LFS with administrative data on graduate earnings, exploring the relative strengths and weaknesses of both of these sources of data. Much more briefly we will also compare the administrative data to the longitudinal version of the "Destination of Leavers from Higher Education" (DLHE) survey, which is a specialist survey organized by the Higher Education Statistics Agency (HESA) to track the former students around 3.5 years after they leave HE. HESA's mission to is produce statistics on higher education in the UK. The DLHE survey is funded by the UK Government.

Understanding how both the LFS and the longitudinal DLHE compare to administrative data on graduate earnings is also important from a policy perspective. The extensive modelling that has been undertaken of the English higher education funding system is largely underpinned with data on graduate earnings from the LFS (e.g. Department for Business, Innovation and Skills (2010) and Chowdry et al. (2012)). Further, the UK Government's website `unistats` provides DLHE based earnings data to potential HE students in order to aid students and their families make decisions about higher education. Improving our understanding of the quality of these data sets is therefore paramount.

1.3 Administrative data

The administrative data we use in this paper is from a novel database we have built. It provides longitudinal data on graduates' annual official earnings in the United Kingdom. This database was constructed by using a unique identifier to link three complex administrative data sets, namely data from the Student Loans Company (SLC) and from Her Majesty's Revenue and Customs (HMRC) Pay As You Earn (PAYE) and self-assessment (SA) databases. This provides us with a very large database of UK earnings data on individuals domiciled in England upon application to HE who received loans from the SLC.

The SLC is a state enterprise which provides loans to English domiciled students to cover their maintenance costs and tuition fees for higher education, on more favourable terms than obtained from commercial lenders. The loans they provide have historically had relatively low real rates of interest and have income-contingent repayments, resulting in low earning graduates not repaying, or repaying more slowly than high earnings graduates. For recent cohorts the SLC data set covers approximately 90 percent of English domiciled students who attend UK universities (we discuss the potential for selection bias from this below).[1]

This paper does not estimate the economic return to a degree. Instead the analysis will provide a more accurate and complete picture of graduates' earnings in the UK than hitherto has been

[1] See `http://www.slc.co.uk/media/855703/slcsfr052014.pdf`

possible. Further, in later studies, we will use it to study the earnings of subpopulations (e.g. those who study a specific subject at a specific university and graduate in a specific year).

1.4 Literature

There is a significant literature which has discussed the problems associated with using sample survey data, comparing their results to some administrative data. Here we review that work.

There are a number of problems when measuring earned income using self-reported surveys, as well as some specific issues relating to the way in which the data is collected in the LFS (e.g. Skinner et al. (2002) in relation to measuring low income in the LFS). Bound et al. (2001) discuss the many sources of error in self-reported income data. Sources of error include the complexity of data being asked of respondents and social desirability bias that may cause some to under or overstate their income. With the LFS, an additional source of bias is introduced by the use of proxy reporting of household members' income.

Bound et al. (2001) conclude that self-reported annual earnings tend to have less error than more disaggregated measures, such as hourly or weekly earnings. This is partly because hourly earnings in particular have measurement error in both earnings and hours. These findings are consistent with previous studies that have found only moderate measurement error in annual earnings but considerable error in estimates of hours of work and hourly pay (e.g. Duncan and Hill (1985)).

Bound et al. (2001) also found evidence that errors are mean-reverting, which is a particular problem when considering the earnings paths of graduates. There was mixed evidence on whether graduates or individuals with more human capital were more likely to report their earnings with error, though some individual studies that have compared survey measures with administrative records have found a positive correlation between true earnings and error in earnings (e.g. Rodgers et al. (1993)). Bound et al. (2001) found limited empirical evidence on considerable social desirability bias but did find non-negligible measurement error in measures of schooling and highest education level: this too may contribute to measurement error in estimates of graduate earnings.

A comprehensive review of studies by Moore et al. (2000) that focused on sources of error in earnings measures in official surveys suggested a wide range of different sources of both random and systematic bias. Non-response is an issue, though less so with earned income than other sources of income, such as from assets. Another important factor is that asking respondents about their income is cognitively complex. Respondents may not completely understand the different definitions of income being used (e.g. in the LFS they are asked for earnings both "before deductions" and net pay "after deductions"). Questions may not be precise about excluding or including pension contributions and childcare allowances and individuals may have recall problems depending on the period being asked about. Whilst it is well known that income data collected with a single question

are subject to extensive measurement error (e.g. Micklewright and Schnepf (2010)), even when more complex survey designs are used it remains a challenge to design high quality instruments with which to measure income in surveys, particularly if (as is the case with the LFS) one is often asking household members to report on the income of others.

Annual earnings have considerable volatility, in the US context at least, in both administrative and survey data. But this volatility does appear to differ markedly according to the data source (e.g. Dahl et al. (2011)), suggesting that measurement error in survey data is a particular issue.

This paper will contribute in a number of ways. First, it follows the tradition of validation studies for surveys in comparing the "true" distribution of official earnings against the earnings data in the LFS. Second, in widening the scope of administrative data, and revealing its advantages and disadvantages, as discussed in Webber (2009) and Card et al. (2010). Third, documenting the substantial variation in graduate earnings, which has increased over time (e.g. Blundell et al. (2005), Bratti et al. (2005), Chevalier (2011), Hussain et al. (1999), Sloane and O'Leary (2005), Smith and Naylor (2001) and Walker and Zhu (2011)). Fourth, documenting the impact of the Great recession on the distribution of graduate and more generally young peoples' earnings (e.g. Jenkins et al. (2012), Gregg et al. (2014) and Bell and Blanchflower (2010)). Specifically, we will provide empirical evidence on the variation in graduate earnings using higher quality data than previously.

1.5 The structure of the paper

In Section 2 we detail our data sources for Administrative data and how we linked them. The Section also has various summary statistics of the databases and details our eventual "Golden sample" (GS) of yearly earnings of individual borrowers who were domiciled in England at the start of their HE careers. In Section 3 we discuss the UK's LFS, focusing on graduate earnings. We describe how this database is built, provide sample sizes and discuss its main features.

In Section 4 we review three other data sets. The first is the DLHE survey. The second is the "Silver sample" (SS), which is our second linked Administrative database. The SS is the cohort of people who are not English borrowers, and hence who are likely to be non graduates, but who have the same age profile as the GS (specifically we randomly assign some older individuals to younger cohorts to reflect the fact that the GS cohorts include people who started university older than 18). We use this to very roughly approximate the population of non-graduates. Our third data set is the corrected Silver sample, which uses econometric methods to adjust for the fact that some of the individuals in the SS will be graduates. Specifically we attempt to adjust for former English domiciled students who do not borrow and for former students who were domiciled in Wales, Scotland and Northern Ireland when they started in HE and thus who are not eligible to

borrow from the English part of the SLC. This is our best, but tentative, estimate of the earnings of non-graduates in England. We compare our results for non-graduates with those obtained by using the LFS.

Section 5 compares the LFS and GS results. This ranges over the cross-sectional features of the data and the dynamics. Section 6 uses the results from the data summary to empirically assess the impact on earnings of the recession. Section 7 concludes. There is a lengthy Appendix which contains additional results, typically for more cohorts than covered in the main text.

2 Our administrative database: the Golden Sample

2.1 UK tax forms

The UK runs an individual tax filing system — there is no option to file as a household. This means UK administrative data will be good at studying individuals' earnings but not the earnings of households[2]. Since our aim here is to examine variation in graduates' earnings, we will entirely focus on individual earnings rather than attempting to calculate household earnings.

The UK has two types of tax forms. The significant majority of tax payers use the "Pay As You Earn" (PAYE) system, which is operated by employers who withhold income and other employment taxes and report the earnings and deductions made to HMRC. This means the majority of UK citizens do not themselves file taxes. Pope and Roantree (2014) report that around 90% of UK income tax is collected through the PAYE system.

For those with more complicated tax affairs (e.g. high incomes, self-employed, owning a business, having significant investment accounts, being in a professional partnership) HMRC requires them to file a set of "self-assessment" (SA) forms. Individual taxpayers can also opt to submit SA forms.

Once submitted to the HMRC, UK tax forms are highly confidential and access to them is restricted by Parliamentary statutes. We have been given access to an anonymized version for a study of individual level measurements of earnings and human capital. Our work, reported here, was carried out in a highly secure data enclave in a HMRC office. All outputs from our work have been checked by officials to ensure they cannot be disclosive of any individual's information. As an example, we are not allowed to report or plot any individual's earnings, even though the individual is anonymized. This makes some attractive scatterplots unavailable to our readers.

[2]Guvenen et al. (2014) use U.S. Social Security data to look at earnings (recorded in W2 filings) by gender. In principle HMRC also has address information, which would allow us to fuzzy link individuals into households. However, we have no access to that information so we have no ability to map our individual earnings data into household data.

2.2 HMRC's research databases

For internal research purposes HMRC has a database which is a random subsample of the taxpayer population. It includes individuals with specific digits in their randomly allocated identification number (this is a "National Insurance Number", known as a NINO). The digits are a run of 10 numbers out of a possible 100 and so are thought of statistically as a 10% random panel sample. This research sample is used by HMRC researchers to model PAYE returns. By using a 10% sample the sample sizes are thought to be easily manageable whilst being large enough to draw robust conclusions[3]. We have access to this database from 2002/03 through 2012/13, noting that in the UK the tax year runs from April 6th to April 5th each year. Hence with this database we can track individuals' earnings paths for a decade.

2.3 Student Loan Company data

2.3.1 Background

Table 1 shows the number of 18 year olds[4] and Table 2 the number of new HE students in each year. Both tables are split by country of domicile, gender and cohort. Around 80% of UK HE students are domiciled in England, but England also has the lowest participation rate of the four countries. Although there are marginally more male 18 year olds in the population in England, substantially more women attend HE than men in all countries and in all cohorts. HE is defined as any institution whose students are eligible to receive a loan from the SLC. Our dataset includes named institutions where more than 1,000 English-domiciled students have received a loan from the SLC, of which there are 170. The several hundred institutions with fewer than 1,000 students receiving a loan are not named in our dataset and are instead classified as "other" institutions. These consist largely of smaller Further Education Colleges. The vast majority (around 97%) of borrowers attend named institutions, however.

The Student Loan Company has been making income contingent loans available to English domiciled HE students since 1998[5]. The take-up rate in more recent times is around 90%[6].

The SLC kindly deposited a copy of some anonymized elements of their loan book data onto the secure HMRC computers, so it could be analyzed by us after being linked together with the

[3] We are discussing with the HMRC access to the other 90% of the PAYE data. This would have little impact on this paper's country wide results but would ease extensions of our work on subpopulations of former students.

[4] It is difficult to define what a satisfactory population size might be for those who could start HE as all ages can potentially go in any year. Government definitions have changed through time and vary by country. Consequently we decided to report this very simple well defined but imperfect measure. It is used here to set the scene and only impacts our subsequent results in the calculation of the corrected Silver sample estimates.

[5] At the start of this project we wrote to the appropriate civil servants in Scotland and Wales to ask them to allow the SLC to release data to HMRC for us about students domiciled in their countries. We have yet to receive a reply.

[6] Not all people receiving a loan from the SLC will be studying for first degrees. Some are carrying out foundation degrees, HNDs and smaller undergraduate qualifications. The dataset we received from SLC does not have any indicators to split up the borrowing populations into these different groups.

	England			Scotland			Wales			N. Ireland		
	All	M	F	All	M	F	All	M	F	All	M	F
1998	604	305	298	64	32	32	36	18	18	24	12	12
1999	600	304	296	65	33	32	36	18	18	24	12	12
2000	586	299	287	62	31	30	36	18	18	24	12	12
2001	593	303	290	60	30	30	35	18	17	25	13	12
2002	611	315	296	60	31	29	36	18	18	25	13	12
2003	637	328	309	64	33	31	38	20	18	25	13	12
2004	635	324	311	65	33	32	37	19	18	27	14	13
2005	641	321	320	64	33	31	38	19	19	25	13	12
2006	661	335	326	66	34	32	39	20	19	25	13	12
2007	658	335	323	64	33	31	39	20	19	23	12	11
2008	668	342	326	64	33	31	39	20	19	24	12	12
2009	692	353	339	66	34	32	41	21	20	24	12	12
2010	682	346	336	66	34	32	40	20	20	24	12	12
2011	671	343	328	66	33	33	39	20	19	24	12	12
2012	670	344	326	65	33	32	39	20	19	24	12	12
2013	649	333	316	63	32	31	37	19	18	24	12	12

Table 1: Total population indicators (in 1,000s). The number of 18 year olds in each mid-year in the different countries, so includes both those in and not in HE (mid-year is a reasonable approximation to the start of the HE year, e.g. 1998 will be taken to correspond to 1998/99). Source: Office of National Statistics, "Estimated Resident Population Mid-Year by single year of age" series.

HMRC databases. The SLC database covers around 2.6M former borrowers who are qualified to be in repayment, which happens the first April of the year after they leave HE[7]. We have no data on those who are in HE but are yet to qualify for repayment, which explains the drop off in loan numbers in our database in more recent student cohorts.

The linkage we carry out is at the individual level. This was possible as the SLC and HMRC databases are indexed by National Insurance Numbers (NINOs), a personal account number used in the UK in the administration of the National Insurance system. The NINOs themselves are anonymized to ensure that researchers cannot observe them: HMRC coded them using the same scrambling device applied to all databases. This enabled us to link the databases on the basis of these anonymized NINOs. The important information that we gleaned from this SLC data is whether or not a taxpayer took a loan for their higher education, which we use as a proxy for whether or not they are a graduate. During this period the drop out rate from UK universities for those who enroll is low at around one in ten, including mature entrants. We might expect drop-outs being included in the SLC dataset means our estimator of graduate earnings will be downward biased.[8]

[7] A person in the 1998/99 cohort who takes 3 years to graduate in 2001 will start being be in repayment 2002/03. Some students leave HE early, so enter repayment earlier. Others become in repayment later. We have synchronized students via the year they started HE, which we have called their cohort, not by age or by when they left HE.

[8] See `www.hesa.ac.uk/dox/performanceIndicators/0405/t3a_0405.xls`. SLC does have information on drop outs, but it was not in the information they made available to us as it was viewed as being potentially disclosive. Note: HESA measure of drop out only includes those who attended for at least 90 days before dropping out, while SLC will include all those who started at HE but dropped out.

	Newly recorded Higher Education enrollment (HE) (in 1,000s)												%Participation											
	England			Scotland			Wales			N.I.			England			Scotland			Wales			N.I.		
	All	M	F	All	M	F	All	M	F	All	M	F	All	M	F	All	M	F	All	M	F	All	M	F
1998				*31*	*14*	*17*				*10*	*4*	*6*				47	43	52				41	*34*	*48*
1999	238	113	125	*32*	*14*	*18*				*11*	*4*	*6*	*39*	*37*	*42*	48	43	54				44	*37*	*51*
2000	238	112	128	*32*	*14*	*18*				*11*	*5*	*6*	*40*	*37*	*44*	51	46	57				46	*39*	*53*
2001	244	112	132	*31*	*14*	*17*	*21*	*10*	*11*	*11*	*5*	*6*	*41*	*37*	*45*	51	45	57	*61*	55	67	45	*38*	*52*
2002	255	117	138	*29*	*13*	*16*	*22*	*10*	*13*	*11*	*5*	*6*	*41*	*37*	*46*	48	42	55	*62*	55	70	46	*39*	*53*
2003	257	116	141	*31*	*14*	*17*	*24*	*11*	*13*	*11*	*5*	*6*	*40*	*35*	*45*	48	43	54	*63*	54	72	44	*37*	*51*
2004	261	118	143	*30*	*14*	*16*	*23*	*10*	*13*	*12*	*5*	*7*	*41*	*36*	*46*	46	41	51	63	53	73	45	*38*	*52*
2005	281	127	155	*30*	*14*	*16*	*23*	*10*	*13*	*12*	*5*	*7*	*43*	*39*	*48*	47	41	53	61	52	71	47	*39*	*55*
2006	284	127	156	35	16	19	*24*	*10*	*13*	*11*	*5*	*6*	42	37	47	53	46	59	61	50	71	46	*39*	*53*
2007	294	132	162	34	15	19	*23*	*10*	*13*	*11*	*5*	*6*	43	38	48	52	44	59	59	49	68	49	*41*	*57*
2008	311	141	171	36	16	20	*22*	*10*	*12*	*12*	*5*	*7*	45	40	50	54	47	61	57	48	67	48	*40*	*56*
2009	322	146	176	37	17	20	*24*	*11*	*13*	*12*	*5*	*7*	46	41	51	55	49	62	59	50	68	50	*42*	*58*
2010	324	149	175	37	17	20				*11*	*5*	*7*	46	41	50	55	49	61				48	*40*	*56*
2011	341	158	183	37	17	20				*12*	*5*	*7*	49	45	54	56	49	62				48	*40*	*56*
2012	294	135	159							*11*	*5*	*6*	43	38	47							45	*38*	*52*
2013										*12*	*5*	*7*										49	*41*	*57*

Table 2: Estimates of numbers in Higher Education (HE). Italics numbers have been imputed. Imputed rates use HE numbers and populations from Table 1 and are presented when we were unable to source official data. The participation rates themselves are not used in this paper, but the HE numbers moderately impact our estimates of quantiles and mean for non-graduates. Official participation rates times Table 1 results are used to impute HE numbers. We have no data on the gender split in participation rates for Northern Ireland. We assume a 15% spread to proxy the other countries. Using the population size at age 18 is quite coarse. Sources: https://www.gov.uk/government/collections/statistics-on-higher-education-initial-participation-rates

2.3.2 High quality linking

Throughout this paper we treat the administrative data as being an accurate representation of graduates' earnings. There are a number of reasons to believe that this is the case. The most compelling reason is that there is a legal requirement for earnings to be reported accurately to HMRC and that for the majority of individuals this reporting comes from their employer. The second reason is that these data arguably do not suffer from some of the weaknesses of other linked administrative data sets (e.g. Chetty et al. (2014a) report linkage rates close to 90% using fuzzy matching, based on date of birth, state of birth, names and gender, between school reports and tax records and just under 98% for matching parents to children). For example, the linkage between the SLC and the HMRC data is on the basis of a hard link (based on an individual's unique identification number, namely their NINO) the quality of which has been checked many times.

An applicant for a loan must supply a NINO, their name and date of birth. SLC checks the consistency of this information with the Department for Work and Pensions (DWP), the UK Government's sole issuer of NINOs. A loan is never issued unless the NINO is validated by DWP. We can expect the number of mismatches to be minimal. When the student finishes at the HEP (Higher Education provider) and the SLC contacts HMRC with the NINO, name and date of birth. Again, the link is checked, if it fails an investigation is launched. When former students become non-resident for UK tax purposes, then HMRC may lose contact with them and generally will only record earnings from UK sources as these are their UK taxable earnings. SLC has methods for

chasing up these students for repayment, but we have not investigated those here. We will express their earnings as 0 in our reports if HMRC records it as 0, which is their UK taxable earnings but not necessarily their true earnings.

2.3.3 Concerns over bias caused by selection

There are three main sources of potential bias associated with the administrative data. First, not all English domiciled graduates are in the SLC data, only those who borrow. Second, although there is a legal obligation to accurately report earnings to HMRC, the self-employed have a higher propensity to under-report their earnings (this is quantified in Section 2.6). Third, as described above, SLC borrowers who are non-resident for tax purposes have only their UK earnings recorded which will often be zero. We cannot identify these individuals in our dataset, so we are unable to drop them out of our sample[9].

The first of these sources of bias is the most important. The dataset therefore excludes foreign students, even those who remain in the UK to work, as well as students doing some tertiary level courses in Further Education colleges who do not qualify for loans. The latter students are likely to have relatively low earnings on graduation. Finally some students eligible for a loan choose not to take one (just over 10% of UK domiciled students do not take out a loan by the end of this period).

We might expect that the students who do not take out a loan are either wealthier than average or more averse to taking on income-contingent debt. We cannot sign the biases arising from this. Students who are wealthier are, conditional on their educational achievements at 18, still likely to earn more as graduates (e.g. Crawford and Vignoles (2014)), hence we may be underestimating graduates' earnings by excluding these individuals. Conversely, Callender and Jackson (2005, 2008) have suggested poor students are more debt averse and these students are likely to earn a lower return to their higher education[10]. This would cause bias in the opposite direction.

The two remaining sources of bias are easier to sign: under-reporting of income and graduates moving abroad being treated as having zero earnings will both result in underestimating earnings. Despite these potential biases, the accuracy gained from using tax records leads us to assume that the administrative data, whilst imperfect, is superior to survey data.

2.3.4 Structure of the databases

Table 3 describes the variables in the SLC databases on borrowing and repayment. This paper will not report any subject or institutional analysis. The only aspects used here will be the scrambled

[9] The tax authorities do have this information, but it is not in the databases we have available to us.

[10] Income contingent loans should reduce debt aversion, but some may not be aware of the differences.

Database name	Details	Missing Data
NINO anon	HMRC's scrambled NINO	
Gender	Female, Male	
First academic year	Date first went to any HEP: 1998 onwards	
Last HEP name	Last HEP attended	Small institutions are grouped together and labelled 'Other HEP'
Subject code	First letter of JACS code	Censored if the n in that year group in that subject was less than five.
Subject group	LEM, Other, STEM. LEM denotes law, economics & management STEM denotes science, tech, eng, math	
Amount borrowed	Given in cash aggregated through time, no interest rate applied	
Borrowed first year	Given in cash, no interest rate applied	
PAYE Flag	In Golden sample: Y,N	
Domicile	Almost always England	
Region at application date	Government region of address when the SLC application was first made.	
Voluntary loan repayments	Voluntary repayments of student loan. Cash repaid that year.	Complete record of years in which repayments are made.
Scrambled Universal Tax Record	Scrambled UTR used for SA records. e.g. if UTR2012 is present, a SA tax form for 2011/12 was returned.	Very high levels of selection effects.

Table 3: Main variables in the Student Loan Company database

NINO, gender and first academic year, which we will think of as a cohort. The other variables will be used in subsequent papers.

2.4 Definition of the Golden sample

Most of our results are derived from the "Golden sample" (GS). We defined the GS as the 10% of borrowers in the SLC database whose NINO qualifies them to be in the HMRC panel sample. HMRC include identifiers in the SLC data so that we know which borrowers fall in the 10% sample. Identifiers are also included so we can match borrowers in the 10% sample to the SA data. Hence we can also track the tiny number of individuals who never file with HMRC (specifically these will be individuals in the 10% sample who have no PAYE or SA records at any time).

The GS has 263,052 members, covering cohorts from 1998 to 2011. This is detailed in Table 4. Each individual potentially has a SA and a PAYE tax record in each tax year, but may have neither. By construction, we are able to state that if they have neither a SA nor a PAYE record then they are recorded as having no UK tax return at all — note that unlike the US, in the UK it is not legally necessary to file a tax form if your income is indeed zero – but it is required for any amount above 0. We will record such non-filers as having zero earnings, recognizing of course that there will be some measurement error and some individuals may not declare earnings to avoid taxation. We end up with the GS for whom we have earnings data from the PAYE database, the SA database or both.

	All				Male				Female			
Cohort	Golden	PAYE	SA	Either	Golden	PAYE	SA	Either	Golden	PAYE	SA	Either
1998	14,487	11,646	2,310	12,226	6,927	5,528	1,351	5,875	7,560	6,118	959	6,351
1999	22,621	18,410	3,447	19,354	10,590	8,529	1,912	9,063	12,031	9,881	1,535	10,291
2000	23,506	19,214	3,425	20,176	10,853	8,761	1,908	9,322	12,653	10,453	1,517	10,854
2001	23,924	19,921	3,108	20,818	11,025	9,060	1,759	9,625	12,899	10,861	1,349	11,193
2002	23,891	20,104	2,814	20,906	11,060	9,156	1,576	9,642	12,831	10,948	1,238	11,264
2003	23,972	20,387	2,447	21,097	11,024	9,315	1,314	9,726	12,948	11,072	1,133	11,371
2004	23,577	20,367	2,266	20,997	10,767	9,163	1,251	9,526	12,810	11,204	1,015	11,471
2005	25,103	21,800	2,085	22,397	11,439	9,822	1,141	10,183	13,664	11,978	944	12,214
2006	25,383	22,149	1,864	22,589	11,340	9,749	992	10,024	14,043	12,400	872	12,565
2007	25,352	22,303	1,527	22,694	11,292	9,746	774	9,981	14,060	12,557	753	12,713
2008	20,847	18,154	1,039	18,430	8,990	7,704	531	7,872	11,857	10,450	508	10,558
2009	6,510	5,386	426	5,485	3,029	2,452	215	2,509	3,481	2,934	211	2,976
2010	2,993	2,477	152	2,511	1,334	1,082	72	1,101	1,659	1,395	80	1,410
2011	851	721		724	360	291		294	491	430		430
All	263k	223k	27k	230k	120k	100k	15k	105k	143k	123k	12k	126k

Table 4: 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. Golden denotes the Golden sample. Either denotes being in either PAYE or SA or both. Cohort denotes the first year the former borrower received a loan from the SLC.

In the main text we focus on the 2011-12 tax year. The Appendix contains results for other years. We will only comment on results which broadly hold in all our tax years. Following the Great Recession, 2011-12 was a financially difficult period as the economy was still recovering from the financial crisis. That tax year ran from 6th April 2011 to 5th April 2012 inclusive.

2.5 Government earnings data

We define our measure of labour income as the aggregate of various components of income which are reported to HMRC for tax reporting purposes. As has been said, our data on earnings comes from the two distinct HMRC databases - PAYE and SA databases. The SA databases contain information on various types of income. Since we are interested in labour income, we construct this as the sum of employment income, profits from partnerships and profits from self-employment. Clearly some aspects of the returns from a partnership are due to the associate capital risk a partner is exposed to, however we cannot break that component out here and so take profits from partnerships as earnings.

The SA databases also contain information on trust income, profits on share transactions, profits from land and property, foreign employment[11] and savings, UK dividends, pension income, life policy gains, "other" income, bank and building society interest and total income. Since these variables measure non-employment income, they are excluded from our earned income calculation.

We do not make a record of any deductions tax payers make, e.g. capital losses on invest-

[11] We would have liked to have included foreign income but the calculation involved various delicate deductions and so we decided not to include it.

Variable name	Database	Details
PAYE Income	NPS1112	Aggregate pay through PAYE
SA Employment Income	F1112	Pay from employment plus benefits plus net expenses.
SA PTR Income	F1112	Profits from a partnership.
SA TRLL Income	F1112	Profits from self-employment
SA Total Income	F1112	Total income, including investment income

Table 5: Variables from PAYE (NPS) and Self-Assessment (F) HMRC databases. Labour earnings are the sum of employment income, profits from partnerships and profits from self-employment.

ments, nor of any tax free allowances individuals may have. We also do not account for pension contributions[12]. This is compatible with LFS data which asks for information gross of deductions.

When we have both PAYE and SA earnings we prioritize the SA data, as HMRC regard the SA records as definitive. If an individual has no reported earnings then we take their earnings as zero. This is likely to miss some earnings for very low earners who do not have to return a PAYE form and who may not be asked to complete a SA form (note however that technically they have a legal responsibility to report this income). This research decision is unlikely to be of major importance given that graduate earnings tend to be higher on average. In our GS dataset all earnings are converted into October 2012 prices using the Consumer Price Index (CPI).

Table 5 names the different HMRC databases we draw information from. The majority comes from the "NPS1112" PAYE database. It has the aggregate of all PAYE returns for each individual in the 2011-12 tax year. We do not know the number of different employers the person has, only their annual taxable income. The other data comes from "F1112", which has SA records.

2.6 Basic summaries of Golden sample

Table 4 provides a basic summary of the cohort sizes of the GS and how the HMRC data for each former borrower breaks down. It shows a significant majority of borrowers are female for all cohorts. Notice the 1998 cohort is quite a lot smaller than the later ones as it took a little while for students to adjust from the move from mortgage style loans into the less risky income contingent loans and so their takeup was lower in their first year.

In more recent cohorts there is very little SA data since it is higher earners and the self-employed that are more likely to use SA, but as the cohorts mature this data becomes much more important. There are some people, mostly self-employed, who appear only in the SA data (e.g. in 1998 of the 14,487 individuals in the GS, 12,226 have tax files for that year and 580 only had SA records). There is a considerably higher rate of SA for males, reflecting their greater likelihood of being

[12]We would have liked to have included employer's and employee's tax free pension contributions as labour earnings. These are likely to be a significant fraction of graduate earnings. However, UK tax forms only record pension income not pension contributions so this is not possible. We did consider imputting pension contributions as a function of recorded labour earnings but decided not to do this to make the results compatible with the LFS data.

			% No tax form			% Earnings < £1			% Earnings < £8,000		
Median age	LFS age	Cohort	All	Male	Female	All	Male	Female	All	Male	Female
31	30-31	1998	13.0	12.6	13.3	15.6	15.2	16.0	27.3	26.7	27.9
30	29-30	1999	11.7	11.4	11.9	14.4	14.4	14.5	26.2	25.7	26.7
29	28-29	2000	11.4	11.2	11.5	14.2	14.1	14.2	26.1	25.7	26.5
28	27-28	2001	10.1	9.9	10.3	13.0	12.7	13.2	25.0	24.5	25.5
27	26-27	2002	9.6	9.9	9.3	12.5	12.8	12.2	25.3	25.5	25.0
26	25-26	2003	9.0	8.9	9.0	12.0	11.8	12.2	25.8	25.4	26.1
25	24-25	2004	8.0	8.3	7.7	10.9	11.5	10.5	25.9	26.8	25.2
24	23-24	2005	7.5	7.4	7.5	10.8	11.0	10.6	29.1	30.3	28.2
23	22-23	2006	7.5	7.8	7.2	11.0	11.6	10.5	34.3	36.3	32.6
22	21-22	2007	7.0	7.8	6.3	10.5	11.6	9.6	43.2	45.1	41.8
21	20-21	2008	8.4	9.1	7.8	11.6	12.4	11.0	61.6	63.2	60.4
21	20-21	2009	10.9	11.6	10.4	15.8	17.2	14.5	61.1	64.6	58.0
20	19-20	2010	11.0	12.0	10.2	16.1	17.5	15.0	67.9	72.0	64.6
18	17-18	2011	10.1	13.1	7.9	14.9	18.3	12.4	90.6	90.6	90.6

Table 6: Golden Sample for 2011-12. Shows percentage of individuals with no filed income tax form. Also shows numbers with no or low earnings. Median age does not increase by one each year for later cohorts in the GS because of small sample sizes and variation in the ages of university leavers (since individuals only enter our dataset once they have left university).

self-employed and having higher earnings.

Table 6 maps cohorts into ages using the median age in each cohort. This is to allow us to compare with LFS data. The Table also shows the percentage of individuals who file no tax form at all during 2011/12, which, as has already been discussed, we take as the individual having zero taxable income in the UK. The rate of not filing decreases as the cohort matures but then increases. That there is not a great deal of gender difference in the not filing rate, even as the cohort reaches their early 30s, is surprising given evidence on the unequal split of childcare responsibilities.

We separately record the percentage of people who have incomes below one pound or not filing — using this as a measure of the percentage of former borrowers who are not in the labour market. Note that there are a sizeable group of people in the databases who do have PAYE returns of 0 income, e.g. from an employer they have left in the previous tax year who are filing that they did not pay this former employee in this tax year. Again there is very little difference by gender.

We also record the percentage of borrowers with incomes below £8,000 in October 2012 prices. This low income cut point was selected since it is approximately equal to the level of earnings at which individuals start to pay National Insurance Contributions and income tax (Pope and Roantree (2014)), meaning our HMRC data is more reliable above this level. It is also just below the full-time minimum wage in 2011 (i.e., assuming unpaid holiday, $£6.08 \times 35 \times 48 = £10,214$). Around a quarter of former borrowers are in this situation, with again only a relatively small difference between genders. By way of comparison, according to HMRC and ONS figures, around 38% of 18-65's do not pay income tax in England. We would expect this to be a lower proportion for graduates.

Table 7 quantifies the degree of self-employment in this dataset, showing how it varies with cohort and gender. Self-employment tax records are more vulnerable to potential under recording of earnings than those in employment and hence may be a source of weakness for administrative earnings based data. Although it is a legal responsibility for the self-employed to accurately report their taxable income, the individual has a strong incentive to under report their income. For the self-employed there is no employer based filing which can be used to independently verify the amount of income earned.

Her Majesty's Revenue and Customs (2014) have estimated the amount of uncollected tax caused by the underreporting of income. They estimate a tax gap of around 17% for self-assessed taxes (with around 25% of SA taxpayers underreporting their earnings) and 1.5% for PAYE taxes. The vast majority of our data comes from PAYE sources, and the majority of those with SA reports also have most of their earnings recorded through employer based PAYE records (i.e. the "P60" form). Our main vulnerability is to the underreporting of the non-employment labour earnings of those who are fully or partially self-employed. This is around 10% of our sample. We have not made any correction to the raw HMRC data in our analysis to take this into account though we would obviously expect this to downward bias our estimates for this group.

			% Only partly self-employed						% Entirely self-employed					
Median age	LFS 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	30-31	1998	6.4	7.1	5.7	33.4	27.1	40.7	3.6	4.4	2.8	75.6	82.0	70.9
30	29-30	1999	6.5	7.3	5.8	34.6	30.3	39.3	3.8	4.5	3.1	78.3	81.1	75.9
29	28-29	2000	6.6	7.5	5.8	33.8	31.7	36.1	3.7	4.6	2.9	77.8	83.7	72.0
28	27-28	2001	6.2	7.5	5.1	34.3	31.7	37.5	3.5	4.7	2.5	78.0	86.2	69.5
27	26-27	2002	5.8	6.9	5.0	35.9	35.5	36.3	3.3	4.3	2.4	73.7	82.2	63.6
26	25-26	2003	5.4	6.1	4.8	37.9	33.9	42.1	3.0	3.6	2.5	76.8	81.1	73.0
25	24-25	2004	5.2	6.2	4.3	38.8	36.2	41.9	2.8	3.6	2.1	71.8	76.4	67.0
24	23-24	2005	4.9	5.9	4.1	41.3	41.5	41.1	2.6	3.3	2.0	74.0	74.9	72.8
23	22-23	2006	4.3	5.1	3.7	47.6	46.6	48.8	2.2	3.0	1.5	67.8	76.0	59.0
22	21-22	2007	3.8	4.4	3.4	54.7	50.7	58.8	1.9	2.5	1.5	62.9	68.5	57.8
21	20-21	2008	3.1	3.6	2.7	67.9	68.3	67.5	1.7	2.4	1.2	62.6	75.2	49.5
21	20-21	2009	3.4	4.0	3.0	62.5	63.3	61.5	1.8	2.2	1.4	70.0	75.0	64.1
20	19-20	2010	2.8	3.4	2.4	61.9			1.3			67.3		

Table 7: Golden Sample self-employment percentages. Percentage of cohort who are only partially self-employment (does not include those fully self-employed) and those entirely self-employed. Also given are the corresponding percentages who have low earnings. Earnings means all earnings from work, not just from the self-employed part. Results are given for the 2011-12 tax year.

The proportion of graduates who only have earnings from self-employment is modest at around 3%, with a slightly higher rate for men than women. Of these, around 80% of men who are entirely self-employed report having low labour earnings. A higher rate of partial self-employment is recorded, again with males having higher incidence than females. For the partially self-employed, women have a moderately higher chance of having a low income.

	# Any answers sample size			# Earnings answers			# Unemployed and missing earnings answers			LFS earnings total sample size		
Cohort	All	M	F	All	M	F	All	M	F	All	M	F
1998	3,739	1,604	2,135	857	351	506	243	95	148	996	403	594
1999	3,760	1,644	2,116	878	379	499	274	105	169	1,038	440	599
2000	3,772	1,703	2,069	881	388	493	279	131	148	1,044	463	581
2001	3,510	1,482	2,028	783	328	455	280	122	158	939	396	544
2002	3,401	1,418	1,983	693	274	419	271	114	157	831	329	502
2003	3,329	1,506	1,823	724	318	406	249	102	147	859	372	488
2004	3,266	1,437	1,829	706	296	410	338	152	186	889	374	514
2005	3,124	1,475	1,649	561	267	294	353	172	181	719	345	375
2006	2,975	1,341	1,634	547	231	316	358	182	176	712	309	401
2007	2,802	1,288	1,514	473	196	277	424	199	225	652	272	380
2008	2,310	1,079	1,231	352	153	199	408	216	192	507	230	277
All	35,988	15,977	20,011	7,455	3,181	4,174	3,477	1,590	1,887	9,186	3,933	5,255

Table 8: Labour Force Survey graduates for the 2011 and 2012 waves. The first set of three columns gives the sample size for individuals who provided at least some response to at least one LFS question in at least one wave. The second set of three columns gives the sample size for individuals who also provided responses to the questions on earnings. The third set coumns provides the sample size for individuals who were recorded as being unemployed in waves 1 and 5 and who did not provide earnings data. The final set of columns provides our final usable sample size with imputed earnings data for those recorded as unemployed in waves 1 and 5.

3 The Labour Force Survey

Table 8 shows the sample size for the LFS in their 2011 and 2012 surveys (which is carried out in quarterly waves) for different cohorts and by gender[13]. This double calender year will be compared to the Administrative data for the tax year 2011/12. An individual is included if they answer at least one question in the LFS during the 2011 and 2012 period — they are not in the sample if they are approached and refuse to answer any questions. Such non-response is modelled here as missing at random and entirely ignored, as is typical in labour economics.

The LFS has a five wave design, with a new sample of individual respondents sampled every quarter with four follow up interviews each quarter. This means that five waves of data may be available for one person with the 1st and 5th wave one year apart. Importantly for us, earnings questions only appear in waves 1 and 5. Many people will take the survey but, as is often the case, not provide information on earnings while answering other questions. Typically people will reveal if they are unemployed or in employment, but many responders are reluctant to give their level of individual earnings or the earnings of family members.

Unlike the GS, the LFS does not have information on when graduates started university. We therefore assign individuals to cohorts based on their date of birth, which we observe in the LFS special license access dataset. Individuals who are assigned to cohort based on the year they were 18 on September 1. For example, individuals who turned 18 on September 1, 1998 are assigned to the 1998 cohort.

A high percentage of LFS earnings data is missing altogether. Indeed this is one justification for our hypothesis that administrative sources of earnings data for graduates are both likely to be

[13]We have included proxy earnings responses in our database as we find that the proxy responses make very little difference to our earnings distributions.

	# Answers in both waves			# Earnings both waves			# Unemployed 1st wave			# Unemployed 2nd wave			# Unemployed both waves			Time series Total sample size		
Cohort	All	M	F	All	M	F	All	M	F	All	M	F	All	M	F	All	M	F
1998	374	145	229	197	77	120	7	<3	5	19	8	11	27	8	19	225	87	138
1999	342	144	198	183	76	107	5	<3	5	15	6	9	20	3	17	204	81	124
2000	326	134	192	190	81	109	4	<3	3	7	<3	5	20	7	13	208	87	121
2001	303	118	185	158	66	92	4	<3	4	11	3	8	13	<3	11	173	69	103
2002	252	86	166	129	45	84	7	<3	5	16	6	10	12	5	7	147	52	95
2003	268	102	166	139	54	85	8	<3	6	9	<3	9	19	6	13	158	58	99
2004	248	99	149	104	42	62	7	3	4	14	6	8	19	6	13	121	48	72
2005	180	88	92	72	32	40	12	7	5	5	3	<3	21	9	12	87	39	48
2006	211	97	114	79	38	41	10	5	5	11	6	5	17	8	9	93	45	48
2007	140	60	80	39	13	26	16	9	7	14	<3	12	15	8	7	51	17	34
2008	88	37	51	25	10	15	10	4	6	4	<3	<3	10	4	6	32	13	19
all	2732	1110	1622	1315	534	781	90	35	55	125	44	81	193	66	127	1500	596	901

Table 9: LFS sample sizes for the time series data, following the pattern of reporting described in the footnote of the previous table. Includes LFS graduates who answer in both the 2011/12 waves. LFS disclosure rules require sample sizes of ≤ 2 to be specified. These cases are indicated by <3.

higher quality and may have a somewhat different distribution of earnings. To attempt to deal with missing data, we impute zero earnings for those who report that they are not employed in the LFS.[14] Table 8 records this value as the imputed earnings sample size in the final three columns of the table.

The fact that earnings data are collected in waves 1 and 5 means we have data for groups of individuals one year apart. We call this time series data, for which we follow a similar earnings imputation for the unemployed to that described above.[15] This mechanism delivers the right percentages of non-employment and persistent non-employment in the database if we believe that the survey respondents all answered the employment questions correctly. Notice the sample sizes will be small, roughly two orders of magnitude less than the GS data. This too is a major advantage of the administrative data.

Table 10 also shows the degree of self-employment recorded in the LFS. Consistent with the administrative data, the rate of self-employment is relatively low and gendered, skewed towards men. However, in the LFS the estimated incidence is even lower than in the administrative data.[16]

[14]For example, for the 1998 cohort we have 1,604 male individual interviews during the 2011 and 2012 waves (some are interviewed five times, others once). If all answered the earnings questions then roughly 2/5 interviewees should give earnings answers — corresponding to 642 answers. But only 351 actually did give earnings answers. 95 people said they were unemployed out of 642 (we define not employed here as the individual indicating that they are not employed and not responding to the earnings question. For individuals stating they are not employed but with non-zero earnings, we use their reported earnings and do not record them as unemployed here). We will set the earnings of these 95 people to 0 and then randomly select a fraction (351/642) of these responders (52 people) to add to the group of 351 who gave earnings responses to create a sample with the appropriate fraction of unemployed. This gives a total imputed earnings sample size of $351 + 52 = 403$ (of whom 52 have exactly zero earnings). When we report the unemployment fraction then this is 52/403 (around 13%). In our results for the 1998 male cohort, earnings will be based on this 403 person population. The same method is used for all cohorts and genders.

[15]Imputing the time series data follows the same lines, but is more complicated. The results are in Table 9. The core data are the people who answered some questions in both waves 1 and 5. For the 2004 cohort 99 males did this. 42 answered both earnings questions, 3 were unemployed in the first year but employed in the second, 6 were unemployed in the second year and employed in the first, while 6 were unemployed in both years. We now add to the 42 people: $3 \times (42/99)$ who were unemployed in the first but answered the earnings question in the second year, $6 \times (42/99)$ who were unemployed in the second year but had earnings answers in the first and $6 \times (42/99)$ who were unemployed in both. This delivers a time series database sample size of $42 + 15 \times (42/99) = 48$ who were observed twice.

[16]We investigate whether our finding are sensitive to the inclusion of self-employed individuals in the GS, finding that they are not.

	% Unemployment			% Earnings < £8,000			% Self-employed		
Cohort	All	M	F	All	M	F	All	M	F
1998	14.0	12.9	14.8	19.6	15.6	22.4	6.8	8.7	5.5
1999	15.4	13.9	16.7	21.2	15.9	25.2	6.7	8.5	5.3
2000	15.6	16.2	15.1	20.0	17.5	22.0	5.6	7.7	3.9
2001	16.6	17.2	16.4	22.0	19.4	24.1	5.8	7.2	4.7
2002	16.6	16.7	16.5	21.8	18.2	24.1	5.7	8.0	4.0
2003	15.7	14.5	16.8	20.8	18.0	23.2	4.9	6.2	3.8
2004	20.6	20.9	20.2	24.7	24.3	24.9	4.2	5.6	3.1
2005	22.0	22.6	21.6	28.7	28.1	29.3	3.7	4.8	2.7
2006	23.2	25.2	21.2	32.2	35.6	29.2	3.3	5.1	1.8
2007	27.5	27.9	27.1	38.3	39.7	37.4	2.7	3.3	2.2
2008	30.6	33.5	28.2	47.7	47.8	47.7	2.3	2.9	1.7
all	18.8	19.1	18.7	25.5	23.7	26.8	4.9	6.4	3.7

Table 10: Labour Force Survey: percentage of graduates recorded as unemployed, with low earnings & self-employed in the 2011/12 waves. LFS asks if they are fully or partially self-employed the week of the interview.

This may be because some individuals may not see themselves as self-employed (e.g. those with small amounts of self-employed income in addition to full-time employment). A comparison of Table 6 with Table 10 suggests that the LFS data has a lower proportion of graduates who are recorded as having low income (less than £8,000) than the GS. The difference is mostly in terms of men, who the LFS measures as being substantially less likely to have low incomes (roughly 16% against 27% in the administrative data for the 1998 cohort). The difference between the data sources for women is modest (22% against 28%).

These differences in the incidence of low pay across data sets are material, but it is hard to know what to make of them. The quality of LFS data is weakest for low earners, as we have had to impute some of the data in order not to oversample the unemployed. On the other hand the GS results of no gender effects on low pay is not the expected result (throughout we will see weaker gender effects than others have estimated elsewhere). However, the £8,000 threshold is above the national insurance threshold and so we would expect tax data to yield relatively accurate estimates of the percentage below that threshold. So we have reason to believe the GS results more than the LFS results on low pay. As discussed previously, most of the results we focus on here will look at individuals with earnings above £8,000 and so for these results we should be robust to the issues we just discussed.

4 Other databases

4.1 DLHE longitudinal survey

We will also briefly compare the GS results with the results from longitudinal DLHE survey. The DLHE survey attempts to follow UK and EU domiciled former students between three and four

years after they graduate and is currently carried out every two years. We assume that each student had three years of study at the higher education provider and so started four calendar years before they left (starting in the autumn, leaving in the summer).

The DLHE is quite a complicated survey. We focus on the 2010 results. It relies on a population established by an earlier six month out Early Survey (of 453,880 leavers eligible to take part in this census in 2006/07, of which 332,110 (73.2%) were contacted). This Earlier Survey had 70,960 responders, all of which were then contacted for the longitudinal DHLE survey[17]. 29,340 responded and a further 153,630 from the original census (over-sampling of some sub groups) were contacted yielding an additional 19,725 responses. This provides 49,065 responses with a response rate of around 21%. The census for the DLHE was 29th November, 2010, but the survey was taken up to 10 weeks after the census date. It covered people who left on average 3.5 years earlier in academic year 2006/7. We include only England domiciled students at the time of application to HE, those who had taken a first degree (as distinct from other higher education qualifications) and, for those reporting earnings, we restrict the sample to those working in the UK.

The former students provide information on their employment status (part time, full time, self employed), whether they are in full time education, their earnings and for those who report hourly pay, their hours of work. 16% of all respondents are not in employment. Within the sample, 21,780 students are English domiciled, working in the UK or not in employment with zero earnings, which is the set of individuals we work with.

Respondents are asked for their earnings and the period over which they are paid. [18] We work with the derived salary variable provided by HESA which builds on these fields to provide pre tax annual salary information.

4.2 Silver sample: all but the Golden sample

One of the advantages of the HMRC and SLC linking is that we can also use it to build a sample of people who did not take out English loans. The significant majority of these UK people are non-graduates. We call this database the "Silver sample" (SS).

The Silver sample is built by first looking at the 10% NINO sample and then removing all the borrowers who appear in the SLC database[19]. For each person in this population we know their age and gender. Then for each cohort and gender we have sampled this new population to

[17] See www.hesa.ac.uk for a full description

[18] See `https://www.hesa.ac.uk/includes/C06019_resources/3808%20IFF%20HESA%20Research%20IFF%20HESA%20Research%20Questionnaire_2010_16pp.pdf?v=1.3`

[19] We construct the NINO sample by finding the union of 10% sample NINOs in the SA and PAYE databases from 2007/08 to 2012/13. This misses people who have no tax record at all (including a filing which has zero income) in any of our databases. This creates a very small bias by missing a set of individuals who are persistently not in contact with HMRC.

	All				Male				Female			
Cohort	Silver	PAYE	SA	Either	Silver	PAYE	SA	Either	Silver	PAYE	SA	Either
1998	27,019	16,253	3,401	19,654	14,724	8,490	2,298	10,788	12,295	7,763	1,103	8,866
1999	41,911	25,491	4,808	30,299	22,849	13,543	3,227	16,770	19,062	11,948	1,581	13,529
2000	42,996	25,979	4,756	30,735	23,504	13,833	3,336	17,169	19,492	12,146	1,420	13,566
2001	43,783	26,599	4,423	31,022	23,667	14,017	3,115	17,132	20,116	12,582	1,308	13,890
2002	43,694	26,735	3,964	30,699	23,586	14,139	2,762	16,901	20,108	12,596	1,202	13,798
2003	43,697	26,912	3,710	30,622	23,675	14,376	2,612	16,988	20,022	12,536	1,098	13,634
2004	43,473	26,980	3,415	30,395	23,506	14,489	2,423	16,912	19,967	12,491	992	13,483
2005	46,550	29,668	3,225	32,893	25,026	15,888	2,300	18,188	21,524	13,780	925	14,705
2006	46,403	30,122	2,816	32,938	24,745	16,074	2,010	18,084	21,658	14,048	806	14,854
2007	46,580	30,891	2,538	33,429	24,760	16,517	1,829	18,346	21,820	14,374	709	15,083
2008	37,810	25,585	1,721	27,306	20,063	13,691	1,277	14,968	17,747	11,894	444	12,338
2009	10,298	7,103	476	7,579	5,460	3,702	361	4,063	4,838	3,401	115	3,516
2010	4,836	3,481	175	3,656	2,529	1,813	131	1,944	2,307	1,668	44	1,712

Table 11: Number of Silver sample borrowers and tax data in 2011-12. PAYE is the Pay As You Earn database and SA denotes the self-assessment database. Silver denotes the Silver sample. Either denotes being in either PAYE or SA or both. Cohort denotes the equivalent cohort these individuals would have been in had they borrowed from the SLC.

			% No tax form			% Earnings < £1			% Earnings < £8,000		
Median age	LFS age	Cohort	All	Male	Female	All	Male	Female	All	Male	Female
31	30-31	1998	22.1	21.5	23.0	27.3	26.7	27.9	46.3	43.3	49.9
30	29-30	1999	22.6	21.3	24.2	27.7	26.6	29.0	47.5	43.8	51.9
29	28-29	2000	23.5	21.8	25.5	28.5	27.0	30.4	48.8	45.2	53.2
28	27-28	2001	24.3	22.4	26.5	29.1	27.6	31.0	49.7	46.1	54.0
27	26-27	2002	24.8	23.1	26.8	29.7	28.3	31.4	51.2	47.9	55.1
26	25-26	2003	25.0	23.2	27.2	29.9	28.2	31.9	51.9	48.5	55.8
25	24-25	2004	24.9	22.7	27.5	30.1	28.1	32.5	52.9	49.8	56.6
24	23-24	2005	24.2	21.8	27.0	29.3	27.3	31.7	53.8	51.2	56.9
23	22-23	2006	23.7	21.4	26.4	29.0	26.9	31.4	55.8	53.4	58.6
22	21-22	2007	22.8	20.3	25.6	28.2	25.9	30.9	58.6	55.7	61.9
21	20-21	2008	21.6	19.4	24.1	27.8	25.4	30.5	61.6	59.0	64.5
21	20-21	2009	20.4	19.5	21.3	26.4	25.6	27.3	64.2	62.0	66.7
20	19-20	2010	18.4	17.1	19.9	24.4	23.1	25.8	68.8	66.0	71.8

Table 12: Silver Sample database for 2011-12. Shows percentage of individuals with no filed income tax form. Also shows numbers with no or low earnings.

produce a database with the same age profile as the SLC database. Typically we have generated two members of the SS for every one in the GS.

The SS is a sample of graduates and non-graduates of Scotland, Wales and Northern Ireland and non-graduates from England and graduates from England who did not borrow from the SLC. As England represents about 80% of the population of the UK and we have removed most English graduates, the SS will be mostly made up of English non-graduates.

Summaries of the characteristics of the SS are given in Table 11. This shows the SS over sampling men, a reflection of the GS over sampling women, consistent with HE participation being higher for women. The rate of SA is lower in the SS than the in GS (e.g. in 1999 the GS SA rate is about 15%, while for the SS it is about 11%).

In a moment we will show how to correct for some of these biases, but before that Table 12

	UK Male							UK Female						
	#18	HE	%Part	Gold	Silver	mHE	%ω	#18	HE	%Part	Gold	Silver	mHE	%ω
2001	364	141	39	110	254	31	12	349	166	48	129	220	37	17
2002	377	145	38	111	266	34	13	355	173	49	128	227	45	20
2003	394	146	37	110	284	36	13	370	177	48	129	241	48	20
2004	390	147	38	107	283	40	14	374	179	48	128	246	51	21
2005	386	156	40	114	272	42	15	382	191	50	136	246	55	22
2006	402	158	39	113	289	45	16	389	194	50	140	249	54	22
2007	400	162	41	112	288	50	17	384	200	52	140	244	60	25

Table 13: Quantifying the adjustment needed to allow for English domiciled graduates who did not borrow and who therefore are in the SS. All non-percentages are in 1,000s. 18 is the number of UK domiciled 18 year olds in mid-year. HE is our estimate of the number of individuals entering HE in that year. % Part is our estimate of UK participation rate. Gold is the number of loans in our SLC database given to English domiciled students. Silver is the number of people not getting an English loan, mHE is the number of students who are not in our loan database (as they are from Wales, Scotland or N. Ireland or are from England and declined a student loan). ω is our estimator of the percentage of former students in the Silver sample (that is people who were students but did not have an English loan and are in the Database for non-former English domiciled borrowers). ω will be used to correct the Silver sample in order to compare to the LFS results for non-graduates.

reports some basic summaries of the low paid in the Silver sample. These rates are roughly twice as high as we saw for the GS. More comparisons with the GS will be given in a moment.

4.3 Correcting the silver sample: estimating non-HE population

Table 13 provides approximate estimates of the percentage of graduates in the SS. It is built out of data from Tables 1, 2 and 4. For each cohort and gender it shows the number of 18 year olds in the UK, the number entering HE, the resulting participation rate and the number of English loans issued (labelled as Gold). mHE=HE-Gold, is the number in HE who did not have English loans, which consist of English students declining loans plus those in HE from Scotland, Wales and Northern Ireland. Finally, Silver=#18-Gold. Of course, all of these estimates are very rough built out of somewhat unsatisfactory data.

The fraction ω =mHE/Silver is an estimate of the proportion of the SS who are graduates. For women the number is typically around 21%; for men it is around 14%. So the SS, which has been stripped of the English borrowers in the GS, still has a substantial number of graduates in it. Around a half of these are non-borrowers from England and the rest are all of the graduates from Scotland, Wales and Northern Ireland. Hence the SS will typically overestimate the distribution of earnings for non-graduates, yielding a large bias if we use it to learn about the upper tail or mean of earnings for non-graduates, but at the center of the distribution and in the left hand tail it is likely to be pretty accurate. It should be more accurate for men than for women because the estimated share of graduates in the SS is lower for men than for women.

We now detail the method we use for correcting the SS to allow for the fact that it includes some

graduates. The result is called the "corrected Silver sample" or the imputed "non-HE" population. Inevitably this is somewhat technical and the rest of this section could be skipped on first reading without loss of continuity if your interest is solely in our results.

4.3.1 Econometrics of correcting the silver sample

Let $F_S(y) = \Pr(Y \leq y)$ be the distribution function of SS earnings Y for a specific cohort and gender. F_{HE} will be the corresponding result for the subset that went into HE and F_{HE^C} is the result for the others. We write ω as proportion of graduates in the Silver sample, then by construction

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

We will now state a working assumption.

Assumption 1: For all $y \in R_{\geq 0}$, then

$$F_{HE}(y) = F_G(y),$$

where F_G is the distribution function from the GS.

Assumption 1 says that the distribution of earnings of the graduates in the SS matches the distribution of earnings in the GS - i.e. the GS well represents all graduates, not just English borrowers. It is important to note that F_G is likely to underestimate earnings for English graduates who do not borrow (because we might expect those who do not borrow to come from wealthy families, and we might expect those individuals to have higher earnings themselves than average), but it is difficult to quantify this underestimation.

Under Assumption 1

$$F_{HE^C}(y) = \frac{F_S(y) - \omega F_G(y)}{(1-\omega)}.$$

We have estimated F_S and F_G from the data. So under Assumption 1, we can estimate F_{HE^C}. We will call this the distribution function of the "corrected Silver sample" or the imputed "non-HE" population. Likewise the mean and density of income Y is, respectively,

$$\mathrm{E}_{HE^c}(Y) = \frac{\mathrm{E}_S(Y) - \omega \mathrm{E}_G(Y)}{(1-\omega)}, \quad f_{HE^c}(y) = \frac{f_S(y) - \omega f_G(y)}{(1-\omega)}, \quad y \in R_{>0}.$$

Further, if we see an individual in the SS with income in the range $(y_l, y_u]$, the chance they are a graduate is, under Assumption 1, $\omega \left\{F_G(y_u) - F_G(y_l)\right\} / \left\{F_S(y_u) - F_S(y_l)\right\}$.

As remarked earlier, Table 13 gives a simple measure of ω. In our work below we will take ω as 0.14 for men and 0.21 for women.

5 Survey and Golden Sample comparison

5.1 Outline

In this section we investigate cross sectional and time series differences in graduates' earnings, comparing data from LFS and the GS. We look separately at six cohorts, consisting of individuals who started university between 1998 and 2003 and live in England at the point they are surveyed or when the loan was first issued, respectively.[20] At the end of this section will also briefly compare the GS with the results from .

5.2 Levels of earnings for each cohort

We start by looking at the cross-sectional differences. Tables 6 and 10 reported results for proportions of individuals with earnings below £8,000, where the results were broadly comparable but the GS had about a third more low earners and less gender difference. As described previously, we will now focus on those with incomes above £8,000.

The distribution of earnings for the GS and LFS using 2011-12 data are reported in Table 14 for males and Table 15 for females. The results for the 1999 cohort are displayed graphically in Figure 1, which shows quantiles up to 99.5%. We give the corresponding Tables for 2008/09, 2009/10, 2010/11 and 2012/13 in the Appendix.[21] For another point of comparison, our Appendix also contains our full results for the Silver sample of non-borrowers as well as the corrected version for the non-HE population. Tables 14 and 15 extract key elements, giving the median and mean to help the reader calibrate the levels of earnings in comparison with the rest of the population.

The results are broadly consistent across the different cohorts and tax years. The GS and LFS data are to first order similar, with male earnings being higher for the LFS. Women have higher average earnings under the GS, with the quantile where there is a cross over between the GS and LFS being around 20%. In both cases GS has higher earnings for high quantiles. If we treat the GS as more reliable, the LFS underreports graduate inequality and over estimates the differences between genders.

When we turn to the Silver sample, the Tables show a moderate but important difference between the results for the SS and the corrected version, here denoted Non-HE. The median results for the Non-HE and LFS are roughly the same for men, but the means are not. Similar results hold for women.

Overall the results suggest earnings rates above £8,000 for female graduates and non-graduates are higher than recorded by the LFS, in particular high quantiles are substantially under-reported.

[20] This is to get as close to the GS as possible, which covers those living in England at the time they applied to HE. Since we do not observe this in the LFS, we use country of residence upon being surveyed instead.

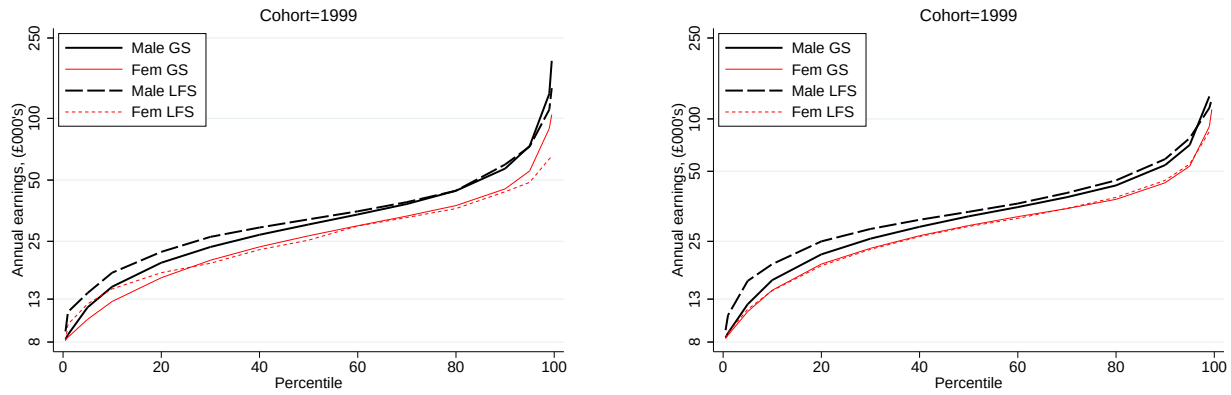
[21] Our comparison with the LFS focuses on the 2011/12 tax year. Our results hold for the other tax years, however.

Cohort	Male borrowers earnings (£000's)										Silver		Non-HE	
	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Mean	Q5	Mean	Q5	Mean
1998	15.2	20.2	24.6	28.6	32.2	36.1	40.7	47.2	63.2	39.6	24.4	31.4	23.5	30.2
1999	15.0	19.6	23.4	26.9	30.3	33.9	38.1	44.3	56.9	35.9	23.6	29.0	22.3	27.8
2000	14.5	19.0	22.7	26.0	29.3	32.7	36.7	42.3	53.0	33.3	22.7	28.8	21.6	26.6
2001	14.1	18.3	21.8	24.9	28.0	31.1	34.8	39.9	49.5	31.6	21.6	27.7	20.7	25.5
2002	13.7	17.3	20.5	23.5	26.3	29.2	32.6	37.5	46.0	29.4	20.9	26.1	20.1	24.4
2003	12.8	16.2	18.9	21.5	24.3	27.0	30.2	34.8	42.6	27.1	20.1	24.1	19.6	23.5
2004	12.0	15.2	17.6	20.1	22.7	25.2	28.4	32.3	39.0	25.1	19.1	22.9	18.7	22.5
2005	11.2	14.0	16.2	18.3	20.5	22.7	25.2	28.5	33.6	22.1	18.1	21.5	18.0	21.5
2006	10.2	12.5	14.4	16.2	18.1	20.1	22.6	25.7	30.4	19.9	17.4	20.3	17.5	20.4
2007	9.4	11.0	12.4	13.9	15.3	16.9	18.9	21.8	26.8	17.1	16.6	19.3	16.7	19.8
2008	8.7	9.6	10.5	11.6	12.6	14.0	15.7	18.0	23.3	14.7	15.8	18.3	16.1	18.6
	Male LFS earnings (£000's)										Other			
1998	18.5	24.4	29.2	34.1	36.1	40.9	46.8	53.6	64.3	40.0	20.9	24.9		
1999	17.5	21.4	26.3	29.2	31.2	35.1	38.5	43.8	57.0	36.8	21.4	24.2		
2000	17.5	23.4	26.3	29.2	30.2	34.1	37.6	40.9	49.1	34.5	21.4	26.3		
2001	18.7	22.2	24.6	26.8	29.2	31.6	34.1	39.0	48.5	32.7	20.3	22.5		
2002	15.2	18.9	21.9	23.9	26.3	27.9	30.4	34.1	42.1	30.9	20.4	22.0		
2003	16.8	18.7	21.4	23.4	26.8	29.2	31.0	34.1	39.0	27.8	19.5	21.3		
2004	14.6	16.6	20.5	23.4	26.3	27.3	31.2	34.1	40.9	26.9	20.4	21.5		
2005	12.2	16.6	18.5	21.0	23.4	25.7	27.3	32.2	40.9	25.0	18.0	20.3		
2006	11.7	14.1	16.4	18.4	20.5	23.4	25.3	29.1	35.1	22.6	17.7	18.9		
2007	11.6	12.7	14.7	17.0	18.5	19.9	21.3	23.0	26.8	18.8	17.2	18.5		
2008	10.5	12.7	13.5	14.6	15.4	16.7	19.0	21.4	24.4	17.3	16.2	16.9		

Table 14: Male GS and LFS earnings above £8k. Quantiles & mean of the earnings (in thousands) reported from the administrative and LFS data for those with earnings above £8k. Q1 denotes 10% quantile, Q2 the 20% quantile, etc. Uses the returns from 2011/12. Also give results from the Silver sample of non-borrowers who have the same age profile as the GS and "others" for the LFS, which correspond to the earnings of non-graduates in the LFS. LFS earnings are weighted using the LFS population weights. The differences in recent cohorts are likely due to drop out effects in the SLC data being exaggerated by many of the students not having finished at their HEP. The final column (non-HE) provides data using the correction to the Silver Sample described earlier.

Cohort	Female borrowers earnings (£000's)										Silver		Non-HE	
	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Mean	Q5	Mean	Q5	Mean
1998	12.8	17.0	20.8	24.2	27.7	31.2	34.8	39.2	48.3	30.3	19.9	24.3	18.2	23.7
1999	12.7	16.6	20.2	23.5	26.6	29.8	33.3	37.5	45.2	29.1	19.4	23.4	18.1	22.7
2000	12.8	16.5	19.8	23.1	26.2	29.2	32.4	36.2	43.5	28.4	18.8	22.7	17.4	21.9
2001	12.4	16.4	19.6	22.6	25.4	28.2	31.1	34.9	41.3	27.0	18.4	21.9	17.4	21.3
2002	12.4	15.9	18.9	21.6	24.4	26.9	29.4	33.0	39.4	25.9	18.2	21.2	17.2	20.8
2003	12.0	15.4	18.2	20.8	23.1	25.5	27.6	30.7	36.3	24.1	17.6	20.2	16.7	19.9
2004	11.7	14.6	17.0	19.2	21.5	23.6	25.8	28.7	33.9	22.5	16.9	19.4	15.8	19.0
2005	11.1	13.5	15.8	17.8	19.8	21.8	23.7	26.1	30.3	20.6	16.2	18.3	15.7	18.7
2006	10.4	12.4	14.2	16.0	17.6	19.5	21.3	23.6	27.1	18.6	15.6	17.4	15.2	17.7
2007	9.5	11.0	12.2	13.5	14.9	16.4	18.3	21.0	24.7	16.2	14.7	16.3	14.6	17.0
2008	8.8	9.5	10.3	11.2	12.1	13.2	14.8	16.8	21.1	13.7	14.0	15.6	14.6	16.8
	Female LFS earnings (£000's)										Other			
1998	12.9	16.6	21.6	24.6	27.3	30.2	33.3	38.0	43.8	30.3	15.6	17.9		
1999	14.6	17.5	19.5	22.4	25.3	29.2	32.2	35.8	43.8	27.8	16.6	18.8		
2000	12.9	16.6	20.3	23.4	25.3	29.2	32.4	35.1	43.8	29.1	16.4	18.0		
2001	12.2	17.5	20.1	23.4	27.1	29.2	31.8	35.1	40.9	28.1	15.0	17.3		
2002	14.0	17.8	21.9	24.4	26.9	28.6	31.2	35.1	39.8	27.6	15.8	17.4		
2003	12.9	15.9	17.5	20.9	23.4	25.7	27.8	30.8	35.1	24.2	15.2	16.5		
2004	13.4	16.6	18.7	21.0	22.4	24.4	26.6	29.2	34.1	23.4	14.1	15.5		
2005	14.1	16.6	18.7	20.5	22.4	24.4	26.3	29.2	33.1	23.5	14.6	15.8		
2006	11.7	14.0	15.6	18.5	20.4	21.6	23.4	24.6	29.2	20.7	14.6	15.5		
2007	10.5	12.6	14.0	16.2	17.5	19.4	20.9	23.0	25.3	17.8	14.1	15.1		
2008	10.1	12.3	13.5	15.2	16.7	18.7	20.6	21.4	23.9	17.6	14.6	14.9		

Table 15: Female GS and LFS earnings above £8k. Quantiles & mean of the earnings (in thousands) reported from the administrative and LFS data for those with earnings above £8k. Q1 denotes 10% quantile, Q2 the 20% quantile, etc. Uses the returns from 2011/12. Also give results from the Silver sample of non-borrowers who have the same age profile as the GS and "others" for the LFS, which correspond to the earnings of non-graduates in the LFS. LFS earnings are weighted using the LFS population weights. The differences in recent cohorts are likely due to drop out effects in the SLC data being exaggerated by many of the students not having finished at their HEP. The final column (non-HE) provides data using the correction to the Silver Sample described earlier.



Nonparametric estimation

Model based estimation

Figure 1: Non-parametric and model based estimates of the LFS and GS conditional earnings distributions, for earnings > £8k for the 1999 cohort for 2011/12. y-axis shows earnings on a log scale. LFS earnings are weighted using the LFS population weights.

On the other hand male graduate earnings disappoint. From a policy perspective these observations are important. Earnings much above the lowest tax rates seem underreported by LFS and this is particularly the case for earnings above the income contingent repayment threshold (£21,000). This tentatively suggests repayments of student loans should be materially more robust than predicted by the modelling based on the LFS.

Section 6 will draw out comparisons of the distribution of corrected Silver and Gold samples without the £8,000 truncation, which will provide another perspective on these issues.

5.2.1 Conditional earnings distribution: model based

As the LFS sample size is small we have also built model based versions of this analysis. Figure 1 shows fitted conditional earnings distributions for the earnings above £8,000 for the LFS and GS data for the 1999 cohort 2011/12. This time the estimation of the quantiles is based upon a model with controls for cohort, gender and year. The model structure we use has

$$\tau = \Pr(Y_{i,t} < q_{it}(\tau)|\mathcal{Z}_i) \tag{1}$$

where $Y_{i,t}$ is the earnings for person i at time t, $\mathcal{Z}_i$ are conditioning variables known about person i from the SLC database at time of first application for a loan (e.g. cohort, gender, year).

Here τ is the quantile level and $q_{it}(\tau)$ is the model based quantile, where

$$q_{i,t}(\tau) = \beta_0(\tau) + \beta_1(\tau)Fem_i + \beta_2(\tau)Cohort + \beta_3(\tau)Cohort^2 \tag{2}$$

$$+ \beta_4(\tau)Cohort_i \times Fem_i + \beta_5(\tau)Cohort_i^2 \times Fem_i + \gamma(\tau)'t \tag{3}$$

and Fem is a female dummy, and $Cohort$ is set equal to 0 for individuals who first went to university in 1998, increasing by 1 with each year. t has a set of year dummies $\gamma(\tau)$.

This model is estimated using a quantile regression at the $100\tau \in \{0.5, 1, 5, 10, 20, ..., 80, 90, 95, 99, 99.5\}$, percentiles, and the plots show the predicted averages at each percentile.

We again observe a reasonably good match between the LFS and GS data, and a similar pattern as in Figure 1, although the model may have smoothed away some of the high earnings effects we see in the non-parametric approach. Specifically, the LFS data continues to over estimate values at the low end of the earnings distribution and under estimate at the high end of the earnings distribution. Once more earnings are truncated at £8,000 and hence these patterns are not explained by the issue of very low earnings being under recorded in the administrative data. Hence whether we use observed values or fitted values that condition on gender, cohort and year, we find similar differences between the LFS and the GS distributions suggesting that the LFS will in particular not measure the earnings of higher paid graduates particularly well.

5.2.2 Measuring inequality: Lorenz and Gini results

Here we plot the Lorenz curves and report Gini coefficients for different cohorts for the GS and non-HE sample in various tax years.

The empirical Lorenz (1905) curve first sorts n individuals by their earnings $Y_{[1]}, Y_{[2]}, ..., Y_{[n]}$. Then it plots $L_n(s)$, the cumulative share of income against the population fraction $s \in [0, 1]$, where

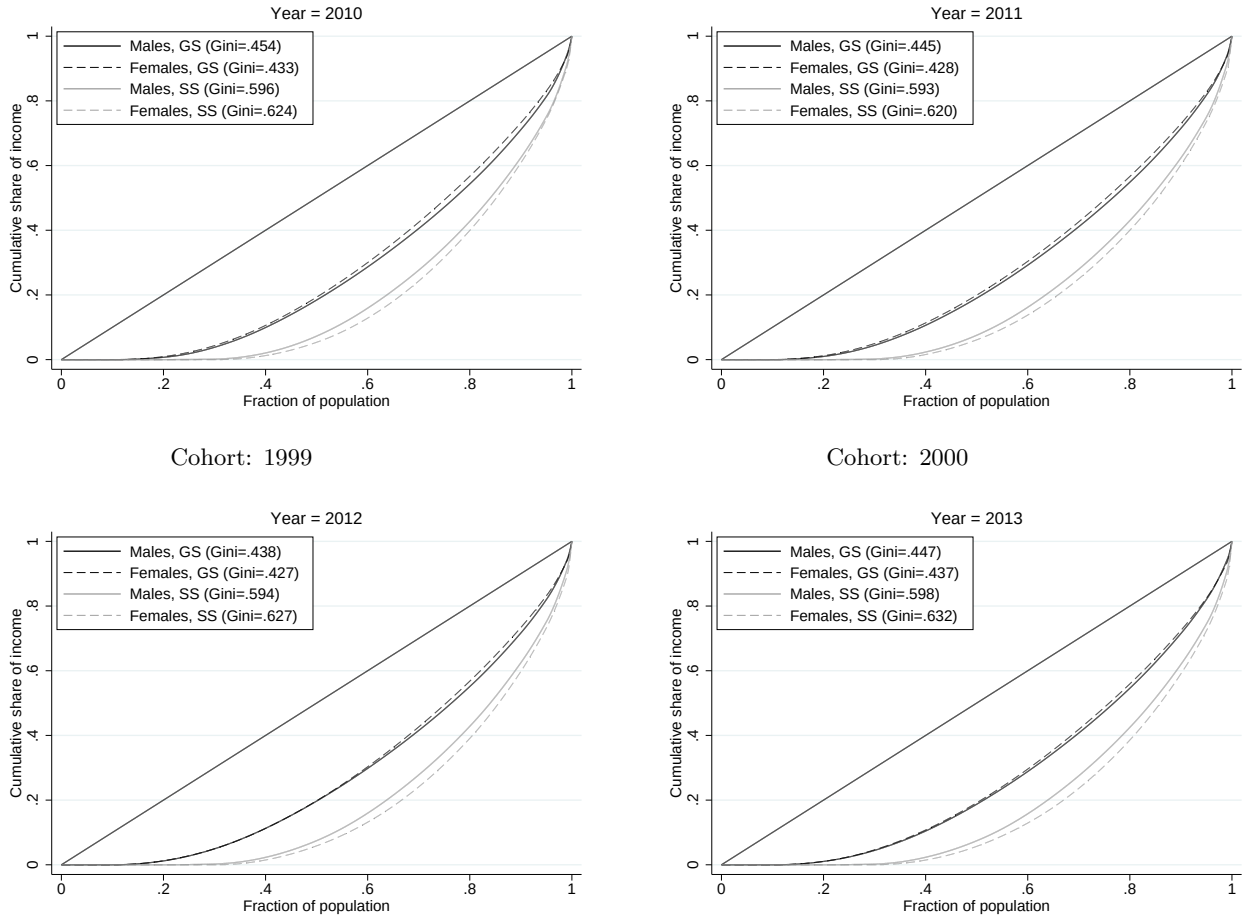
$$L_n(s) = \frac{\sum_{j=1}^{\lfloor ns \rfloor} Y_{[j]}}{\sum_{j=1}^{n} Y_{[j]}},$$

where $\lfloor x \rfloor$ generically denotes the integer part of x. Curves further to the right are regard as representing groups with more inequality. Figure 2 draws the empirical Lorenz curve for our data. It shows that graduate inequality is lower than for the rest of the population, with female non-graduates being particularly unequal. The differences between the graduate and non-graduate Lorenz curves are large.

A well known scalar summary of the curve is the Gini coefficient

$$G_n = 2 \int_0^1 \{s - L_n(s)\} \, \mathrm{d}s \in [0, 1],$$

which is twice the area below the 45° line (alternative measures include the Atkinson (1970) index). Figure 2 includes those numbers in the legend. The value for the non-HE populations is quite high, particularly for women. It does not change very much through the different years.



Cohort: 1999

Cohort: 2000

Cohort: 2001

Cohort: 2002

Figure 2: Empirical Lorenz curves for GS and non-HE samples. A 45% degree line indicates no income inequality. As the curves go to the right there is more inequality. Cohorts and tax years are selected so each picture shows the results for a group of roughly the same age, but at different times during the recession.

5.2.3 Measuring inequality: tail behaviour of earnings

The tail thickness of earnings is important from many different statistical and economic perspectives. The LFS does not have enough data to be able to usefully estimate tail thickness, but here we report some results for the GS. Again we split the analysis by cohort and gender and report separate results for the 2010/11 and 2011/12 data.

To assess the thickness of the tails of the distribution of earnings we follow a conventional approach of fitting a power law distribution to earnings Y:

$$\Pr(Y > x|Y > K) = \left(\frac{x}{K}\right)^{-\alpha} \qquad x > K, \qquad \alpha > 0,$$

to the extremes in our earnings databases (e.g. Embrechts et al. (1997)), where the extremes are about some high threshold K. This follows, for example, Saez (2001), Atkinson et al. (2011) and Jones and Kim (2014) who use the power law of labour earnings and who also map out the implications of α for optimal taxation.

Within each subpopulation of cohort, year and gender, we regard the earnings data as i.i.d., so the observed extremes over an extreme threshold should be roughly i.i.d. with a power law distribution using extreme value theory. The key parameter α tells us that at most α moments exist for the earnings data. In practice we have first sorted our earnings data, which is written as $Y_{[1]}, Y_{[2]}, ..., Y_{[n]}$. We select the K-th largest earnings for each subpopulation of administrative data. The Hill (1975) estimator of α is then

$$\widehat{\alpha} = \left[\frac{1}{K}\sum_{i=0}^{K}\left\{\log\left(Y_{[n-i]}\right) - \log(Y_{[n-K]})\right\}\right]^{-1}.$$

Under standard conditions $\sqrt{K}(\widehat{\alpha} - \alpha) \xrightarrow{d} N(0, \alpha^2)$. Knight (2007) provides an incisive discussion of this estimator and various potential improvements. Table 16 shows the results by gender for each cohort using the 2010/11 and 2011/12 data. Throughout, K was taken to be $0.005n$, so we are looking at 1 in 200 extreme events. Hence we are applying the Hill estimator to the very highest incomes in the sample where the extreme value theory should be a reasonable guide.

The results suggest that graduate earnings are quite heavy tailed, with the variance existing for students recently out of HE but probably not by the time they are 10 years into the labour market. There is a marked and systematic difference in the genders, with men having thicker tails in earnings (Atkinson et al. (2015), Guvenen et al. (2014) and Bertrand et al. (2010)). The values of estimated α we report here are not unusual in the literature on income inequality. Of course it has many econometric implications, for example regressions of earnings on past earnings do not make any conventional sense as the population moments will not exist, at least for men.

			$\widehat{\alpha}$			Standard Error			K		
Median age	LFS age	Cohort	All	M	F	All	M	F	All	M	F
					2011/12						
31	30-31	1998	1.908			0.225			72		
30	29-30	1999	2.070	1.942	2.941	0.196	0.269	0.383	112	52	59
29	28-29	2000	2.577	2.315	2.833	0.240	0.318	0.363	115	53	61
28	27-28	2001	2.646	2.660	3.268	0.244	0.362	0.412	118	54	63
27	26-27	2002	2.584	2.439	3.205	0.239	0.332	0.404	117	54	63
26	25-26	2003	2.950	2.985	5.025	0.273	0.406	0.633	117	54	63
25	24-25	2004	2.899	2.611	4.651	0.268	0.359	0.586	117	53	63
24	23-24	2005	3.390	2.841	4.484	0.303	0.376	0.544	125	57	68
23	22-23	2006	3.322	2.618	5.556	0.296	0.350	0.669	126	56	69
					2010/11						
30	29-30	1998	1.531			0.180			72		
29	28-29	1999	1.972	1.767	2.618	0.186	0.245	0.341	112	52	59
28	27-28	2000	2.188	2.451	2.639	0.204	0.337	0.338	115	53	61
27	26-27	2001	2.513	2.710	3.236	0.231	0.369	0.408	118	54	63
26	25-26	2002	2.475	2.646	2.976	0.229	0.360	0.375	117	54	63
25	24-25	2003	2.959	2.825	4.525	0.274	0.384	0.570	117	54	63
24	23-24	2004	2.439	2.151	4.348	0.225	0.295	0.548	117	53	63
23	22-23	2005	3.731	2.907	5.025	0.334	0.385	0.609	125	57	68
22	21-22	2006	3.546	2.667	5.435	0.316	0.356	0.654	126	56	69

Table 16: Hill estimator $\widehat{\alpha}$ for the GS earnings data, by gender and cohort for 2010/11 and 2011/12. Throughout, the K of the most extreme datapoints are used in the estimation for each cohort and gender subset. K was taken to be $0.005n$, so we are looking at 1 in 200 extreme events. The standard error is an asymptotic one for i.i.d. data and takes on the form $\alpha/\sqrt{K}$.

We have not run the extreme value methods on the Silver sample as these methods will focus on the weakest aspect of the SS, the extreme right hand tail of the earnings distribution.

5.3 Dynamics of earnings

We now investigate the time series properties of the Administrative Data and the LFS. Table 9 showed the times series sample size of the LFS is tiny, so their results will be subject to enormous uncertainty. All the GS cohorts split by gender have very large time series sample sizes covering 6 years. The smallest sample size is 6,927 corresponding to males in the 1998 cohort.

Since earnings are only observed a maximum of twice in the LFS, this restricts us to compare the data to the GS over snippets of two year periods.

We follow a traditional labour economics modelling strategy for the dynamics of using a low income threshold, here taken as £8k, around which to build the model. The model's structure is summarized in Table 17 which shows the four parts of the model. The bulk of the data is in the "Steady" category where incomes are above £8k in both years. There is a small group of labour market "Inactive" people who fall below the threshold twice. Finally, we have "Joiners" and "Leavers" for the job market, who are transitioning between low and non-low earnings.

	Earnings >8k at time t	Earnings <8k at time t
Earnings >8k at time $t-1$	Steady	Leavers
Earnings <8k at time $t-1$	Joiners	Inactive

Table 17: Structure of Markov dynamic model for earnings from time $t-1$ to time t.

		Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Mean	N
		Golden Sample Borrowers										
1999	Male	0.94	0.97	0.98	0.99	0.99	0.98	0.97	0.95	0.89	0.90	7,404
		(0.01)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.01)	(0.01)	(0.01)	
	Female	0.83	0.94	0.97	0.98	0.97	0.94	0.91	0.86	0.79	0.84	8,312
		(0.02)	(0.01)	(0.01)	(0.00)	(0.00)	(0.00)	(0.01)	(0.01)	(0.01)	(0.01)	
		Labour Force Survey Graduates										
1999	Male	0.71	0.83	0.90	0.92	0.94	0.96	0.90	0.91	0.94	0.81	71
		(0.22)	(0.13)	(0.06)	(0.04)	(0.04)	(0.05)	(0.06)	(0.10)	(0.12)	(0.05)	
	Female	1.08	0.97	0.96	0.95	0.93	0.93	0.88	0.89	0.78	0.91	99
		(0.28)	(0.06)	(0.04)	(0.02)	(0.03)	(0.03)	(0.03)	(0.07)	(0.23)	(0.06)	

Table 18: Coefficients from quantile and mean autoregressions of log-earnings (2011/12) on lagged log-earnings (2010/11) for the Golden Sample and LFS, 1999 cohort. Figures in brackets are estimated standard errors for the persistence coefficient $\beta_1(\tau)$, where τ is the quantile level. Results for other cohorts are given in Table 26.

5.3.1 Steady: earnings at t and earnings at $t-1$

We start by modelling those who have non-low earnings in both years. Throughout this paper we will focus on earnings Y_t, where $t = 2010/11$ and $t = 2011/12$ tax years and then look at changes in log earnings quantile-regressed on the level of lagged log earnings. Initially we focus on the 1999 cohort, the Appendix contains the results for the other cohorts. As usual we report separately for different genders.

The basic model structure will be

$$\tau = \Pr\left\{\log(1k) \leq \log Y_t < q(\tau) | Y_{t-1} > 8k\right\},$$

where we assume a linear quantile

$$q(\tau) = \beta_0(\tau) + \beta_1(\tau)\log(Y_{t-1}). \tag{4}$$

We report the estimates for the persistence coefficient $\beta_1(\tau)$, for a small number of quantiles in Table 18. For the GS results the coefficients have tiny standard errors due to the large sample sizes. At every quantile level the GS male earnings are more persistent than those of females. Further, for both genders the GS data has more persistence near the centre of the distribution than in the tails. The LFS data has a very small sample size and so the coefficients exhibit quite some scatter and are poorly determined which clearly makes modelling the dynamics of graduates' earnings from this data set problematic.

Due to privacy constraints we are not allowed to show you scatter plots of GS or LFS earnings. To produce scatter plots of GS type data we have resorted to simulation. To do this we simulate

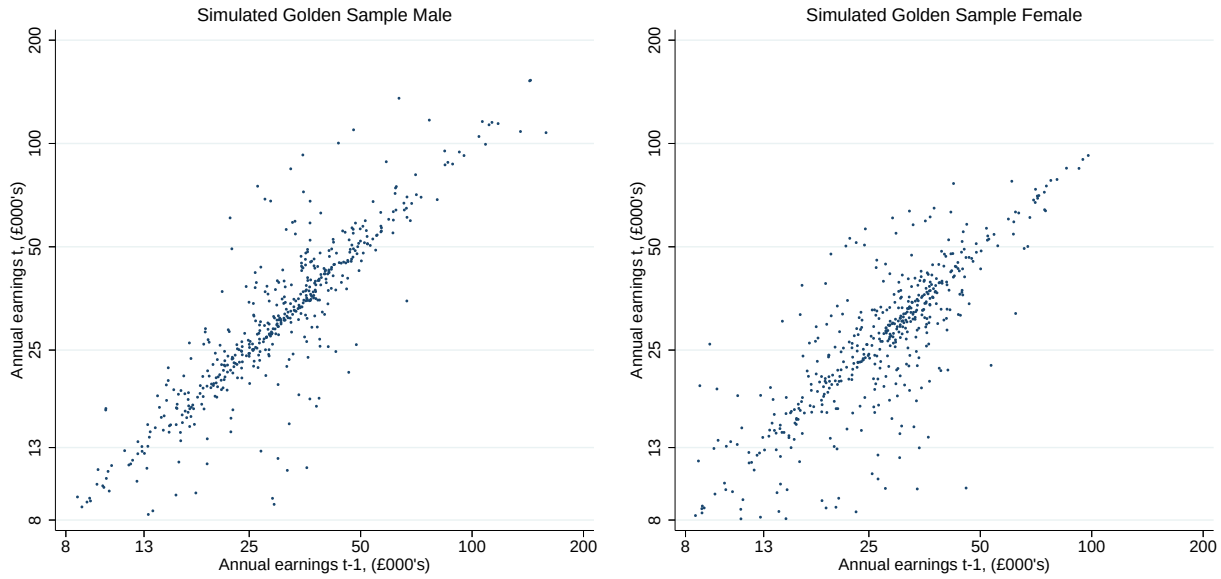


Figure 3: Steady: Simulated Golden Sample earnings above £8,000 at t on earnings above £8,000 at $t-1$, Males and Females, 1999 cohort. t is 2011/12, while $t-1$ is 2010/11. 500 individuals are simulated in each picture.

off a model for earnings above the threshold $Y_{t-1}|Y_{t-1} > 8k$ and then simulate from $(Y_t|Y_{t-1} > 8k)$ by simulating standard uniforms and then using that drawn value as the quantile level to invert the quantile regression model. The only challenge in carrying this out is that we need to be able to fit the model for all quantile levels.

To implement this we have fitted a quantile process regression model where we fit smooth functions $\beta_0(\tau)$, $\beta_1(\tau)$, $\tau \in [0,1]$, so we can compute (4) for any permissible τ. To estimate these functions we first run the quantile regression for Q different discrete quantiles levels τ_i, $\widehat{\beta}_0(\tau_i)$, $\widehat{\beta}_1(\tau_i)$, $i = 1, 2, ..., Q$. We then use least squares to place a R order polynomial in τ

$$\beta_j(\tau) = \beta_j + \sum_{k=1}^{R} \beta_{k,j}\tau^k, \quad \tau \in [0,1), \quad j \in \{0,1\},$$

through these estimates. Throughout we have taken $R = 9$ and $Q = 1,000$. This means we have summarized all the possible quantile linear regression models using $2(R+1) = 20$ parameters.

The results are shown in Figure 3 for 500 draws. We have compared these simulated results to the actual data in the secure Datalab and the results are substantively similar.

5.3.2 Joiners: earnings at t given low earnings at $t-1$

Here we quantify the earnings dynamics for those who at $t-1$ had earnings of less than £8,000 but then at time t had incomes above the threshold. We have called these "Joiners" to the labour market. For the LFS the sample size is tiny and so the results will be very ragged, while for the

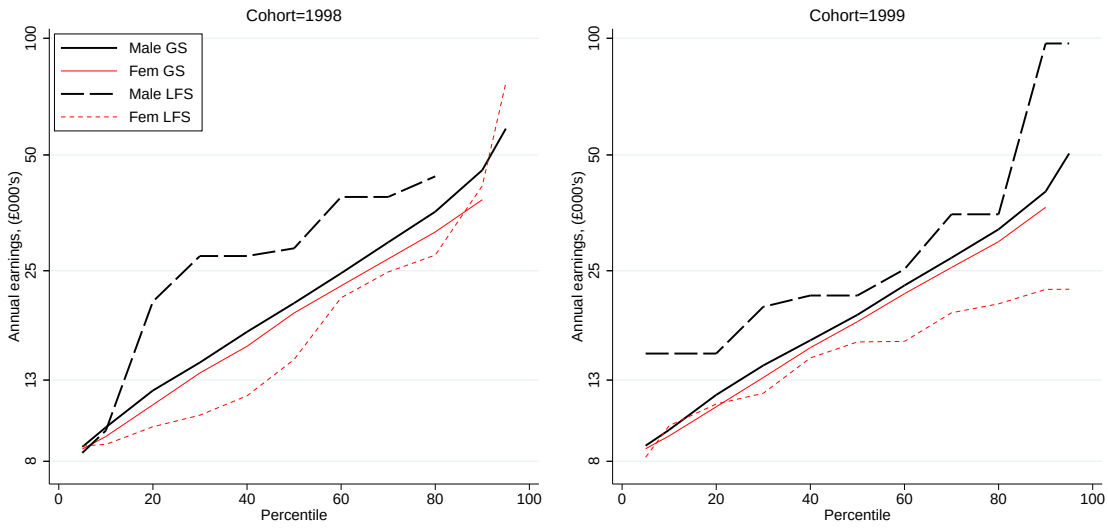


Figure 4: Joiners: LFS & GS earnings distributions t given low earnings $t-1$, 1998 & 1999 cohorts. LFS earnings are weighted using the LFS population weights.

GS data is again plentiful.

Figure 4 shows the quantiles of earnings at t given low earnings in $t-1$ for the 1998 and 1999 cohorts. Due to the small sample sizes in the LFS, we pool LFS data over more years so that t ranges from 2008 to 2012, while $t-1$ ranges from 2007 to 2011. We look at the 5th, 10th, ..., 90th, 95th percentiles only as it is not possible to go further into the tails due to sample size issues.

Other cohorts are in the Appendix as Figure 13. Overall, the distribution of the joiners changes modestly with gender in the GS, with female joiners always having lower earnings. Typically the gap is around 10% at most quantiles.

5.3.3 Leavers: earnings at $t-1$ given low earnings at t

We now look at the opposite set of people, those who had earnings above the threshold at time $t-1$ but then went below at time t. We call these people "Leavers" and give results for the 1998 and 1999 cohorts. Again all pooled LFS results are very ragged and unreliable.

Figure 5 shows the quantiles for the GS and again the results for male are consistently and modestly above the results for females. The curves in Figures 4 and 5 are very similar so the quantiles of those going in and out of the labour market seem roughly the same. This may reflect graduates coming in and out of post-graduate education, but we have no data to check this.

The Appendix contains Figure 14 which give results for more cohorts.

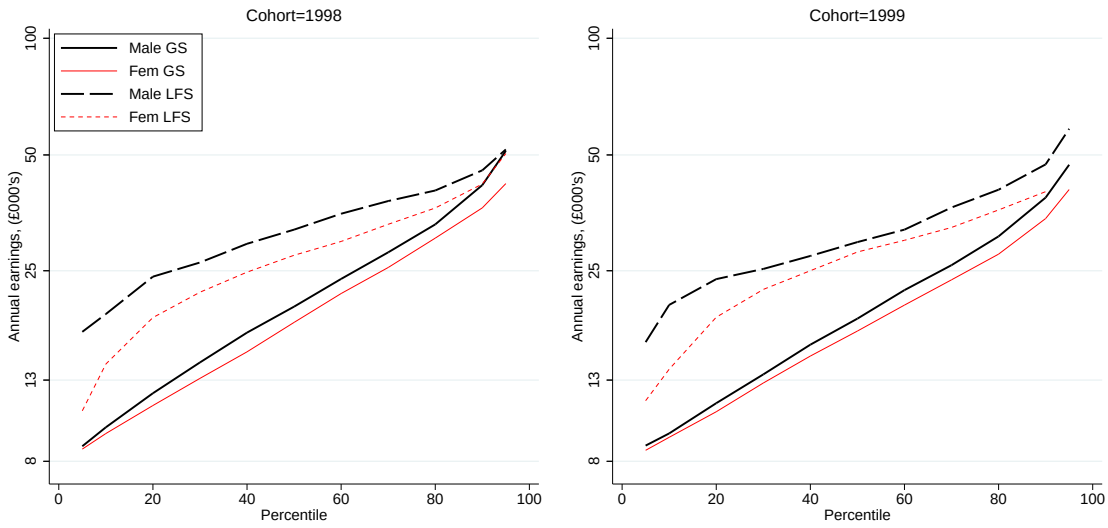


Figure 5: Leavers: LFS & GS earnings distributions $t-1$ given low earnings t, 1998 & 1999 cohorts. LFS earnings are weighted using the LFS population weights.

	Pr(Low Earners)				Pooled N over 07/08-11/12			
	Male		Female		Male		Female	
Cohort	GS	LFS	GS	LFS	GS	LFS	GS	LFS
1998	0.75	0.46	0.75	0.65	9,668	13	10,222	51
1999	0.73	0.64	0.73	0.59	14,386	11	15,926	39
2000	0.71	0.64	0.71	0.60	15,374	11	17,246	42
2001	0.68	0.60	0.68	0.39	16,291	10	18,280	33
2002	0.66	0.50	0.65	0.50	18,441	10	19,811	30
2003	0.65	0.60	0.63	0.52	22,356	15	24,781	27

Table 19: Inactive: Probability of remaining in the low earnings state. Based on pooled data from 2007-08 to 2011-12 tax years. LFS and GS results together with their sample sizes. Here N denotes the number of inactive people in these databases. Roughly the HMRC database is around 50-100 times larger for men and 20-50 times larger for women than the LFS data.

5.3.4 Inactive: low earnings at t given low earnings at $t-1$

Table 19 shows the proportion of individuals earnings under £8,000 in t given they are earning less than £8,000 at $t-1$. We have pooled the data for the different tax years to boost the sample for the LFS. Even so the resulting sample sizes are tiny, particularly for men where the sample is close to useless, and so the results are highly speculative.

We observe that the LFS underscores these probabilities of persistently low income for women, but for the 1998-2000 the differences are not dramatic. The increase in the probabilities with maturing cohorts occurs both for men and women in the GS. There is some evidence of this for women in the LFS data, but this is speculative.

Gender	DLHE 2010/11 Earnings (£000's)									
	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	mean
Male	12.5	17.3	20.8	24.1	26.3	29.1	31.5	35.3	41.7	27.5
Female	11.6	15.9	18.9	21.4	23.4	25.3	27.3	29.8	34.2	24.2
	Golden Sample 2003 Cohort 2010/11 Earnings (£000's)									
Male	12.4	15.7	18.2	20.6	23.2	25.7	28.4	32.5	39.1	25.5
Female	12.1	15.0	17.7	20.1	22.3	24.4	26.5	29.3	34.5	23.1
	Golden Sample 2004 Cohort 2010/11 Earnings (£000's)									
Male	11.4	14.2	16.5	18.8	21.2	23.5	26.3	29.7	35.0	23.2
Female	11.4	14.0	16.3	18.3	20.5	22.4	24.5	27.0	31.4	21.2

Table 20: Quantiles of the earnings reported from the Long DLHE in 2010/11 and 2010/11 earnings for the equivalent cohort (2003 and 2004) from the Golden sample. Earning are truncated at £8,000.

5.4 DLHE and Golden sample comparison

We can now return to the long DLHE. The results for earnings above £8,000 are given in Table 20 for the 2010/11 tax year, together with two cohorts from the GS which most closely match the DLHE data. Although direct comparison is difficult, overall DLHE estimates of earnings seem substantially above the HMRC results. Our conclusion is that the DLHE data is somewhat over optimistic about graduate earnings and this could be due to a number of factors, including errors in reporting, sample selection and problems estimating the annual earnings of workers who do not report an annual salary.

5.5 Comparing the HMRC population and the LFS sample

To put the GS and Silver sample in context, it is helpful to take a step back to compare the 1999 cohort population seen in the LFS with the corresponding HMRC population of those who have some kind of tax record in at least one of the five years from 2008/09 to 2012/13.

Figure 6 shows the distribution of earnings for those with earnings above £8,000. Overall it shows considerably higher HMRC earnings for men and women than observed for the LFS, with this high level seen at all percentiles above about 20%. This is consistent with the high earnings figures we saw for the SS than the LFS results for non-graduates.

6 Graduates through the recession

We now turn to using the summary results presented in the previous Sections. Here we answer the question: how were English graduates affected by the great recession?

Figure 7 shows the quantiles of real earnings for each tax year from 2008/09 to 2012/13 split by gender, for students who are in the cohort in which the bulk of students are 29 years old (e.g. in 2012/13, we use the 2001 cohort). The most recent data is displayed using a cross. This picture is designed to take out cohort effects. Importantly these graphs show all GS members, not those with incomes above £8,000. Consequently, there is a substantial group with 0 earnings. Also given

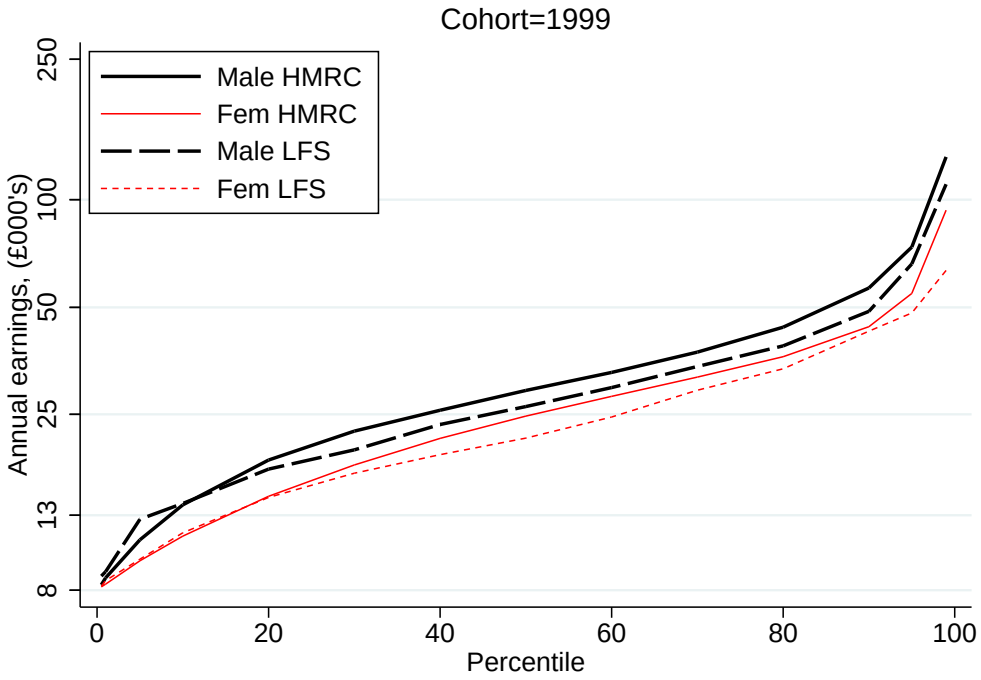


Figure 6: Comparison of HMRC and LFS based population (graduates and non-graduates) estimates of earnings > £8k for the 1999 cohort in 2011/12. LFS earnings are weighted using the LFS population weights.

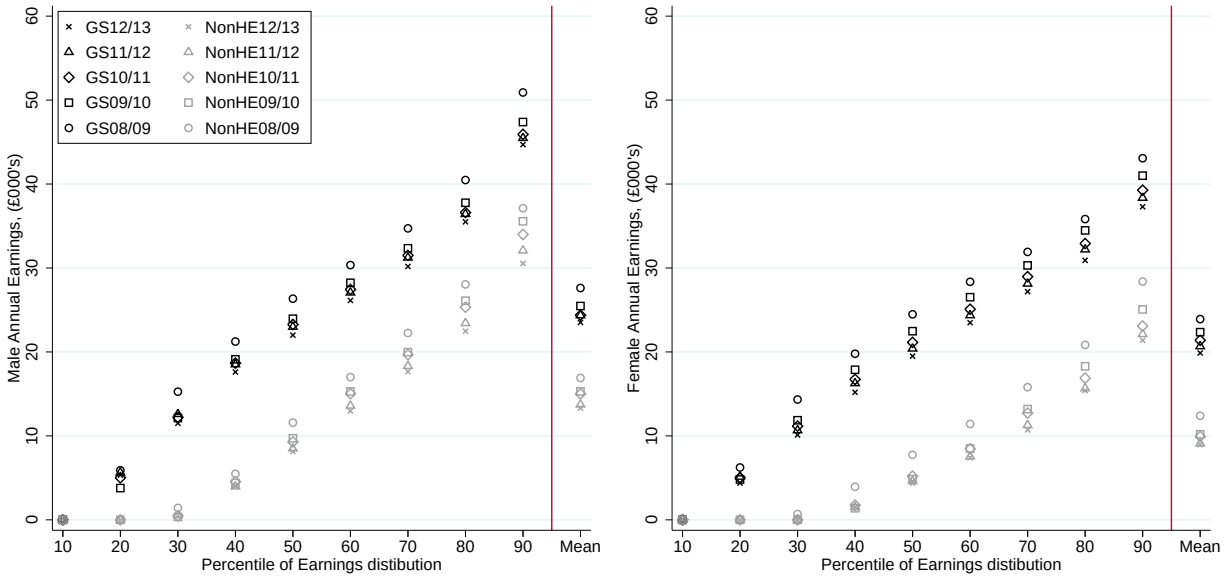


Figure 7: Impact of recession: removing cohort effects. Quantiles of GS earnings of 29 year old borrowers by gender during 5 years of the great recession: 2008/09-2012/13. The same analysis is also reported for non-students using the corrected silver sample.

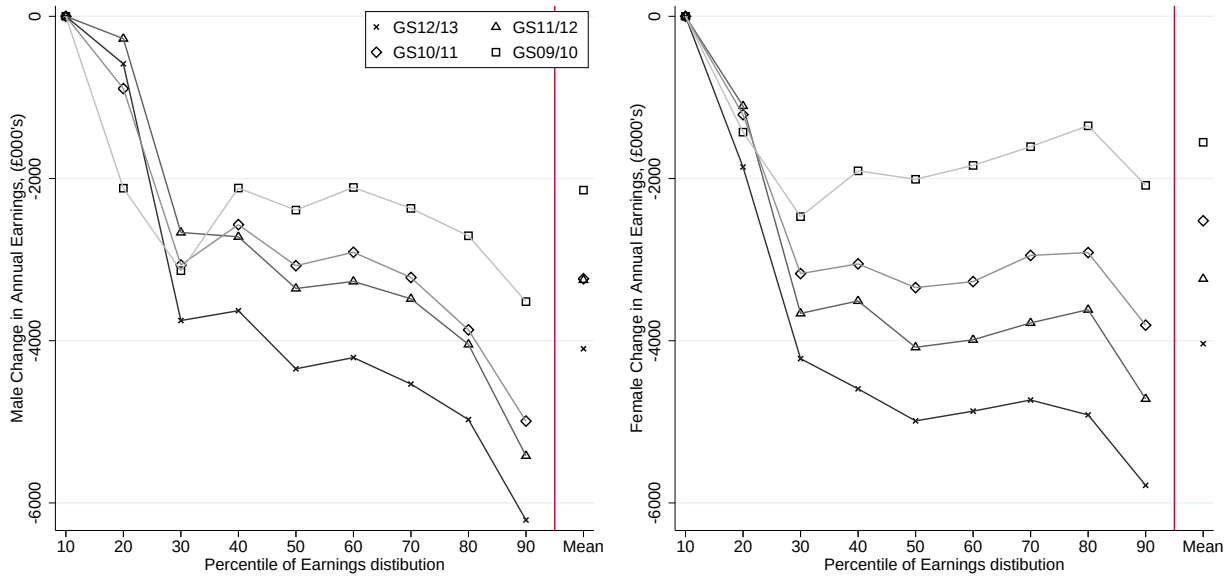


Figure 8: Impact of recession on graduates: removing cohort effects. Changes in the Quantiles of earnings of 29 year old borrowers by gender during 5 years of the great recession: 2009/10 to 2012/13. This is based on the Golden Sample.

in these figures is the results for the corrected Silver sample — our best estimates for non-HE UK individuals.

Around 70% of non-HE women have incomes below £10,000 in 2012/13, while for non-HE men it is just over 50%. For GS women it is 30%, for men a little below this. Almost all of the very low paid in the UK are in the non-HE group and most of them are women. The earnings prospects for non-HE women are almost uniformly very bad — only around 10% of them have earnings above £20,000 in 2012/13. An important by-product of these Figures is to show how much less inequality there is amongst graduates, particularly women, than there is amongst the non-graduates.

We also give the means for these different groups. In the non-HE sample this is somewhat fragile as we may have not fully adjusted for the presence of some high paying graduates which can impact the right hand tail and so the mean. But they do show the usual result that the GS mean is far higher than that for the non-HE, and this is particularly the case for women whose non-HE average earnings are below £10,000.

Figure 8 shows the corresponding temporal changes for the GS, explicitly showing the reductions in the year group's real earnings compared to the corresponding cohort in 2008/09, while Figure 9 shows the equivalent for the Non-HE sample. These reductions effect all quantiles beyond the 10% level. They show the dramatic fall in male cohort earnings in the first year of the recession (this does not necessarily imply salaries were actually cut: rather graduates would expect very rapidly

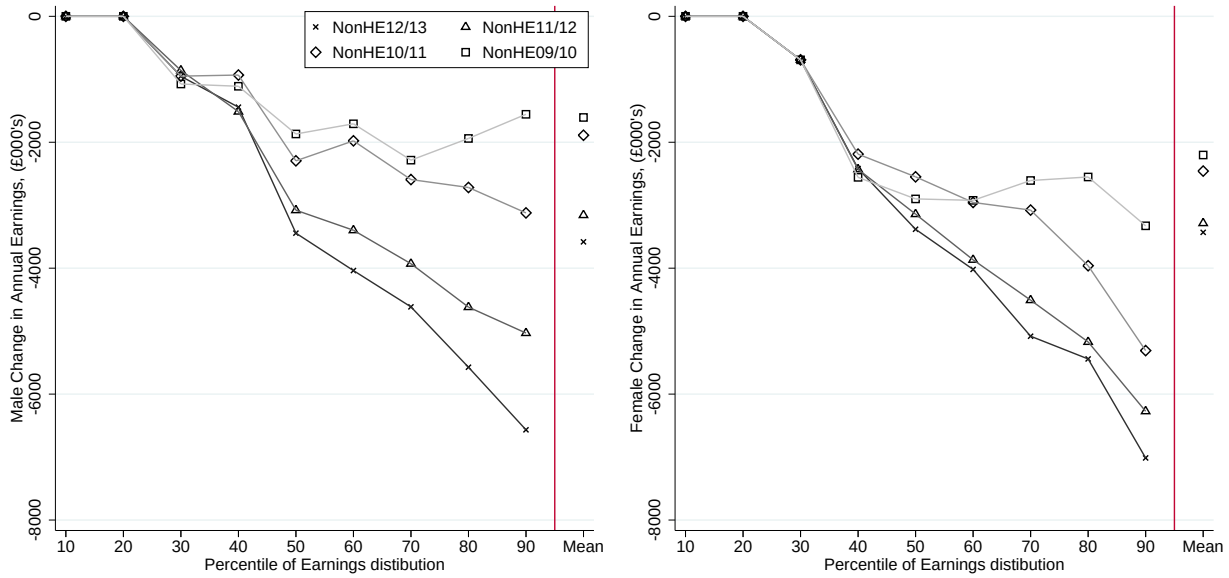


Figure 9: Impact of recession on non-HE: removing cohort effects. Changes in the Quantiles of earnings of 29 year old borrowers by gender during 5 years of the great recession: 2009/10 to 2012/13. This is based on the Non-HE Sample.

rising wages during this period in their lifecycle and this group of individuals had rises which were very disappointing compared to earlier cohorts), followed by a fall of similar magnitude, but spread over the next 3 years. For women the fall was initially less severe but through time the reduction has had roughly the same impact and worse for most quantiles.

To quantify these we now focus on median graduate real earnings in 2012 prices. For men the falls were cumulatively over the 4 years we look at here £2,400, £3,100, £3,400 and £4,300, which corresponds to 9%, 12%, 13% and 14% falls in earnings compared to what we would have expected for that age group. For women the corresponding results are £2,000, £3,000, £4,100 and £5,000 and 8%, 14%, 17% and 20%, although we should note that higher female quantiles did considerably less badly than this as a percentage.

These are very large negative earnings shocks, compared to previous cohorts, and provide evidence that the relatively young absorbed a large share of the reduction in UK real earnings seen during the great recession by having much less fast wage growth than you would expect for people in that part of their lives. The figures also show that proportionally, women were worse affected than men. The corresponding figures for our estimates of non-HE men are £2,600, £2,600, £2,600 and £3,500 which correspond to 22%, 22%, 23% and 30% of their median earnings. For women the falls were £2,300, £2,000, £3,000, £2,900 which correspond to 31%, 28%, 40% and 40% of their median earnings. Hence it suggests graduate earnings were less disappointing than the earnings

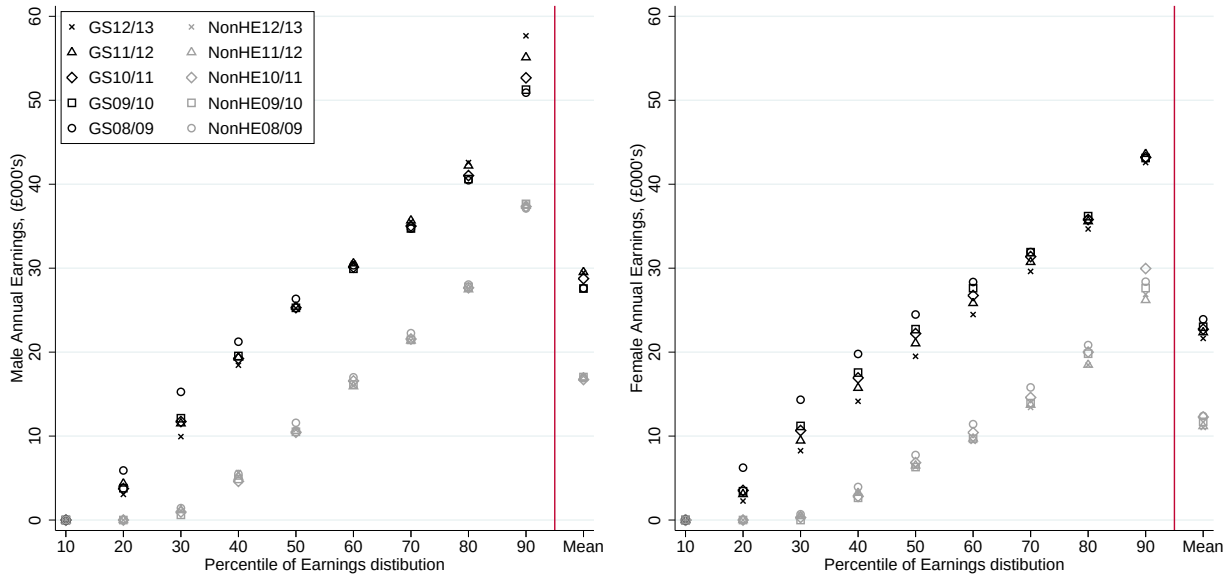


Figure 10: Impact of recession: the real earnings of the 1998 cohort through the recession. Quantiles of earnings for the 1998 cohort of borrowers by gender during 5 years of the great recession: 2008/09 to 2012/13. This is based on the GS. The result combines cohort effects and general economic conditions. The same analysis is also reported for non-borrowers using the corrected silver sample.

of non-HE individuals. That is HE seems to have some insurance value during very bad times. Earnings insurance is very valuable.

To place these numbers in context, Figure 7 of Office of National Statistics (2014) shows the mean gross real earnings per hour for full and part-time employees in high-skilled occupations fell around 11% during this period, while the mean for low-skilled occupations fell by around 14%.

The uniformity of Figure 7 does not imply that the distribution of their earnings is stable as each cohort ages. Figure 10, which shows real earnings for the 1998 cohort through time, indicates this is not the case. Higher earning men have rapidly increasing incomes during the great recession, but their advancement was not as fast as previous generations due to the recession. All but the highest earning women have an actual real earning decline during this period — with their earnings falling considerably in real terms. This is due to the recession combined with the life cycle effects of unequal childcare responsibilities hitting hard during this period in their lives. Within the five years we plot here, we go from men and women graduates having reasonably equal earnings to distributions which are starkly different — this comparative change is due to life cycle effects. But the fact that female real earnings actually fall so much during this period is due to both cohort and recession effects.

The long term effects of this difficult start for this generation of graduates are as yet unclear.

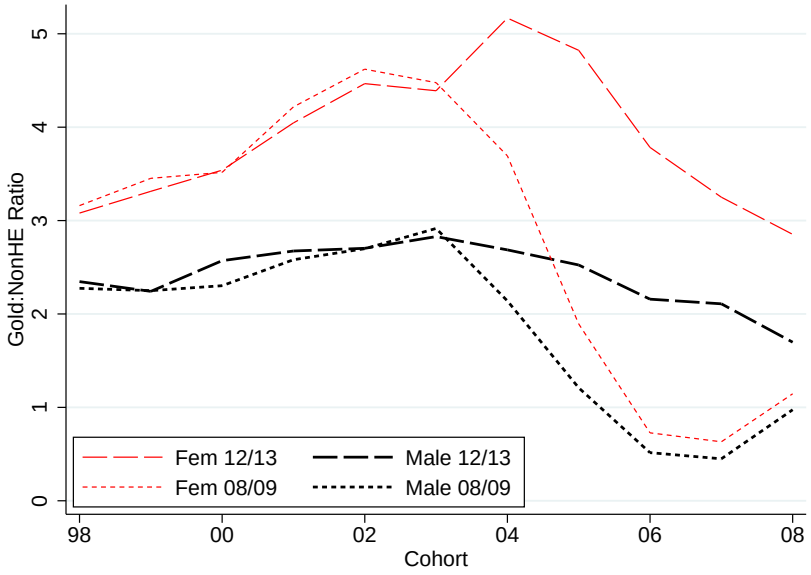


Figure 11: Ratio of the median earners for the Gold and Corrected Silver samples. Shown drawn against cohort, split by gender. Each line represents a tax year. Figures below 1 for recent cohorts for 2008/09 tax year are due to early drop-out biasing the sample.

We do know however, that early unemployment spells in particular have long lasting scarring effects (e.g. Arulampalam et al. (2001), Burgess et al. (2003) and Gregg and Tominey (2005)).

Having discussed these falling incomes, it is important to keep this in perspective: graduate women's earnings are so much higher than the corresponding incomes of non-graduates, suggesting a high economic return to HE for women, consistent with existing literature which has shown a higher return to higher education for women.

Figure 11 shows the ratio of the median earnings for the GS with those from the non-HE sample for different cohorts for different tax years. The GS's earnings for men are roughly 2.3 those in the non-HE sample. For women the ratio is around 3.3. For lower quantiles these multiples are higher, for higher quantiles the ratios are lower. The caveat is that results for the higher quantiles may reflect the inadequacies of the non-HE sample.

7 Conclusion

The distribution of graduates' earnings from the LFS and from the administrative data are very similar, however the survey data appears to underestimate the higher quantiles of earnings and overestimate gender diversity and serial dependence. There are larger differences between the tax data and LFS results for non-graduates, where the survey data is yielding higher earnings for individuals in the lower quantiles of the earnings distribution.

The potential under payment of tax detailed by Her Majesty's Revenue and Customs (2014)

might have suggested that the administrative data would under-report the earnings of higher earners as compared to the survey data. The opposite is true in practice. The LFS may be vulnerable to error in its reporting of high earnings for many different reasons, including individuals having complicated and lumpy earnings, selection bias, proxy response and social desirability bias.

The LFS also underestimates the persistence in earnings over time compared to the tax database. The latter is a particularly strong data source for modelling longitudinal effects due to its massive sample sizes and lengthy time series with little drop-out. This is an important advantage for modelling elements of the HE finance system which requires good longitudinal data on graduates' earnings.

The large administrative dataset also allowed us to quantify the impact of the great recession on graduate and non-graduate earnings distributions, and we find very big falls in real earnings when we control for cohort effects. These falls are proportionally bigger for non-graduates than for graduates, suggesting higher education provided some protection from this major economic shock. This indicates graduate careers carry less undiversifiable risk than non-graduates which means that the spread between graduate and non-graduate human capital is higher than is typically reported in labour economics which often ignores the pricing of this type of risk exposure.

These findings are important for our understanding of the graduate and non-graduate labour market and key policy issues relating to HE funding. First, we find that conventional measures of earnings inequality based on survey data may be too low, for we find high earnings underestimated and low earnings overstated in LFS data. The relatively lower level of inequality amongst graduates is also striking. Second, we find less gender inequality amongst graduates than is indicated by the LFS data which merits further investigation. Third, and perhaps most fundamentally for those seeking evidence of the advantages of higher education, the ratio between graduate earnings and non-graduate earnings is large, typically over 2 for men and over 3 for women. We are however, mindful that we have not calculated a rate of return to a degree in a conventional sense, since we do not compare graduates with a control group of similar non graduates. Fourth, estimates of the country wide distribution of graduate earnings by gender and their time series persistence is highly informative for estimates of the degree to which former students will repay their student loans. In turn these statistical features impact official government financial statements as well as influencing how the HE funding system is designed.

The results on gender are particularly important. First, most of the lowest paid are non-graduate females. Second, women were more negatively impacted by the recession. Although higher earning men increased their earnings during and after the recession period but not as much as previous cohorts and women saw a real decline in earnings over the period, some of which is

attributable to the recession.

We must be mindful of several issues affecting these results. The period we are considering immediately follows the great recession, which may impact our findings[22]. Further, the administrative dataset is not without limitations, specifically the use of SLC borrowers rather than all graduates, the under-reporting of income by individuals for the purposes of avoiding or evading tax, and our necessary assumption that individuals moving abroad have zero earnings. Though we cannot be certain about the earnings of the graduates we are failing to identify, the combination of these factors suggest that the administrative dataset underestimates true graduate earnings. However, despite these imperfections, we view the data as highly informative and our paper makes an important contribution to the literature by highlighting the strengths and limitations of the LFS (and in less detail the DLHE) measures of graduate earnings.

References

Arulampalam, W., P. Gregg, and M. Gregory (2001). Unemployment scarring. *The Economic Journal 111*, 577–584.

Atkinson, A. B. (1970). On the measurement of inequality. *Journal of Economic Theory 2*, 244–263.

Atkinson, A. B., A. Casarico, and S. Voitchovsky (2015). Top incomes and the glass ceiling. Unpublished paper: INET, University of Oxford.

Atkinson, A. B., T. Piketty, and E. Saez (2011). Top incomes in the long run of history. *Journal of Economic Literature 49*, 3–71.

Bell, D. N. F. and D. G. Blanchflower (2010). UK unemployment in the great recession. *National Institute Economic Review 214*, R3–R25.

Bertrand, M., C. Goldin, and L. F. Katz (2010). Dynamics of the gender gap for young professionals in the financial and corporate sectors. *American Economic Journal: Applied Economics 2*, 228–255.

Blundell, R., L. L Dearden, and B. Sianesi (2005). Evaluating the effect of education on earnings: models, methods and results from the national child development survey. *Journal of the Royal Statistical Society, Series A 168*, 473–513.

Bound, J., C. Brown, and N. Mathiowetz (2001). Measurement error in survey data. In J. J. Heckman and E. Leamer (Eds.), *Handbook of Econometrics, Volume 5*, pp. 3705–3843. North Holland.

Bratti, M., R. Naylor, and J. Smith (2005). Variations in the wage returns to a first degree: Evidence from the British cohort study 1970. Unpublished paper: IZA Discussion Paper No. 1631.

[22]The SLC has data on loans taken out before 1998, but we do not have access to that information. This would allow researchers to look further into the life-cycle of the former borrowers. A challenge with that data was the take up of loans was much more modest before 1998 as the loan design was less attractive. Hence selection effects would be stronger. The HMRC has earlier SA and PAYE data than the data we use, and this would allow us to study graduates in non-recession periods and lengthen our data. However, some of the PAYE data is of less high quality before 2008 and so would need extensive use of econometric methods to overcome some significant biases.

Burgess, S., C. Propper, H. Rees, and A. Shearer (2003). The class of 1981: the effects of early career unemployment on subsequent unemployment experiences. *Labour Economics 10*, 291–309.

Callender, C. and J. Jackson (2005). Does fear of debt deter students from higher education? *Journal of Social Policy 34*, 509–540.

Callender, C. and J. Jackson (2008). Does fear of debt constrain choice of university and subject of study? *Studies in Higher Education 33*, 405–429.

Card, D., R. Chetty, M. Feldstein, and E. Saez (2010). Expanding access to administrative data for research in the United States. Unpublished paper: Department of Economics, Harvard University.

Chetty, R., J. N. Friedman, and J. E. Rockoff (2014a). Measuring the impacts of teachers i: Evaluating bias in teacher value-added estimates. *American Economic Review 104*, 2593–2632.

Chetty, R., J. N. Friedman, and J. E. Rockoff (2014b). Measuring the impacts of teachers ii: Teacher value-added and student outcomes in adulthood. *American Economic Review 104*, 2633–2679.

Chevalier, A. (2011). Subject choice and earnings of UK graduates. *Economics of Education Review 30*, 1187–1201.

Chowdry, H., L. Dearden, A. Goodman, and W. Jin (2012). The distributional impact of the 2012-13 Higher Education funding reforms in England. *Fiscal Studies 33*, 211–236.

Crawford, C. and A. Vignoles (2014). Heterogeneity in graduate earnings by socio-economic background. Unpublished paper: Institute of Fiscal Studies.

Dahl, M., T. DeLeire, and J. A. Schwabish (2011). Estimates of year-to-year volatility in earnings and in household incomes from administrative, survey, and matched data. *Journal of Human Resources 46*, 750–774.

Department for Business, Innovation and Skills (2010). *Urgent reforms to higher education funding and student finance: interim impact assessment*. Her Majesty's Stationery Office.

Duncan, G. J. and D. H. Hill (1985). An investigation of the extent and consequences of measurement error in labor – economic survey data. *Journal of Labor Economics*, 508–532.

Embrechts, P., C. Klüppelberg, and T. Mikosch (1997). *Modelling Extremal Events for Insurance and Finance*. Berlin: Springer.

Gregg, P., S. Machin, , and M. FernandezSalgado (2014). Real wages and unemployment in the big squeeze. *Economic Journal 124*, 4008–432.

Gregg, P. and E. Tominey (2005). The wage scar from male youth unemployment. *Labour Economics 12*, 487–509.

Guvenen, F., G. Kaplan, and J. Song (2014). The glass ceiling and the paper floor: Gender differences among top earners, 19812012. Unpublished paper: Department of Economics, Princeton University.

Her Majesty's Revenue and Customs (2014). Measuring tax gaps 2014 edition: tax gap estimates for 2012-13. Issued by Corporate Communications, HMRC.

Hill, B. (1975). A simple general approach to inference about the tail of a distribution. *Annals of Statistics 13*, 331–341.

Hussain, I., S. McNally, and S. Telhaj (1999). University quality and graduate wages in the UK. Unpublished paper: Center for Economics of Education, London School of Economics.

Jenkins, S. P., A. Brandolini, J. Micklewright, and B. Nolan (2012). *The Great Recession and the Distribution of Household Income*. Oxford University Press.

Jones, C. I. and J. Kim (2014). A Schumpeterian model of top income inequality. Unpublished paper: Graduate School of Business, Stanford University.

Knight, K. (2007). A simple modification of the Hill estimator with applications to robustness and bias reduction. Unpublished paper: Statistics Department, University of Toronto.

Lorenz, M. O. (1905). Methods of measuring the concentration of wealth. *Publications of the American Statistical Association 9*, 209–219.

Micklewright, J. and S. V. Schnepf (2010). How reliable are income data collected with a single question? *Journal of the Royal Statistical Society, Series A 173*, 409–429.

Moore, J. C., L. L. Stinson, and E. J. Welniak (2000). Income measurement error in surveys: A review. *Journal of Official Statistics 16*, 331–362.

Office of National Statistics (2014). Economic Review, December 2014. http://www.ons.gov.uk/ons/dcp171766_387913.pdf.

Pope, T. and B. Roantree (2014). A survey of the UK tax system. Technical report. "Institute of Fiscal Studies".

Rodgers, W. L., C. Brown, and G. J. Duncan (1993). Errors in survey reports of earnings, hours worked, and hourly wages. *Journal of the American Statistical Association 88*, 1208–1218.

Saez, E. (2001). Using elasticities to derive optimal tax rates. *Review of Economic Studies 68*, 205–229.

Skinner, C., N. Stuttard, G. B. Durrant, and J. Jenkins (2002). The measurement of low pay in the UK Labour Force Survey. *Oxford Bulletin of Economics and Statistics 64*, 653–676.

Sloane, P. J. and N. C. O'Leary (2005). The return to a university education in Great Britain. *National Institute Economic Review*, 75–89.

Smith, J. and R. A. Naylor (2001). Determinants of individual degree performance: Evidence for the 1993 UK university graduate population from the USR. *Oxford Bulletin of Economics and Statistics 63*, 29–60.

Walker, I. and Y. Zhu (2011). Differences by degree: evidence of the net financial rates of return to undergraduate study for England and Wales. *Economics of Education Review 30*, 1177–1186.

Webber, R. (2009). Response to 'The coming crisis of empirical sociology: an outline of the research potential of administrative and transactional data'. *Sociology 43*, 169–178.

8 Appendix: additional tables and figures

Cohort	Male borrowers earnings (£000's)										Silver		Others	
	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Mean	Q5	Mean	Q5	Mean
1998	15.1	20.4	25.0	28.9	32.5	36.3	41.3	49.0	66.0	40.2	24.2	31.6	23.1	29.7
1999	14.9	19.6	23.6	27.3	30.5	34.2	38.8	45.7	60.1	37.2	23.3	29.2	22.7	28.5
2000	14.5	19.2	23.0	26.5	29.9	33.4	37.4	43.6	56.1	34.3	22.6	28.8	21.7	27.4
2001	13.9	18.4	22.0	25.4	28.7	32.0	35.8	41.5	52.9	32.9	21.8	28.0	21.0	25.9
2002	13.7	17.7	21.0	24.4	27.3	30.2	34.1	39.4	48.9	30.8	21.2	26.5	20.4	25.2
2003	13.1	16.7	19.7	22.6	25.5	28.4	31.6	36.9	45.4	28.6	20.4	24.6	19.7	23.8
2004	12.7	16.0	18.8	21.7	24.1	26.8	30.0	34.4	42.1	26.9	19.5	23.6	19.1	22.9
2005	11.9	15.0	17.4	19.7	22.2	24.6	27.4	31.1	37.4	24.2	18.6	22.3	18.3	21.6
2006	11.0	13.6	15.8	17.9	20.0	22.3	25.0	28.5	33.6	22.1	17.9	21.0	17.6	21.3
2007	10.3	12.6	14.5	16.2	18.1	20.2	22.5	25.7	30.2	19.8	17.1	20.3	16.9	20.1
2008	9.5	11.0	12.6	14.0	15.5	17.2	19.3	22.2	26.6	17.2	16.2	18.9	16.4	19.4
	Female borrowers earnings (£000's)										Silver		Others	
1998	12.2	16.1	19.8	23.4	27.0	30.6	34.1	38.5	48.2	29.9	19.4	24.2	18.1	23.5
1999	12.2	16.0	19.4	23.0	26.3	29.4	33.0	37.1	45.7	28.8	19.1	23.1	17.3	21.9
2000	12.5	16.2	19.5	22.7	26.0	29.0	32.2	36.2	44.4	28.5	18.5	22.7	17.1	21.3
2001	12.2	16.0	19.3	22.4	25.4	28.4	31.5	35.4	42.2	27.4	18.1	21.8	16.6	20.6
2002	12.3	15.8	19.0	21.9	24.7	27.5	30.1	33.8	40.2	26.3	18.1	21.3	16.7	20.5
2003	12.0	15.4	18.4	21.1	23.8	26.1	28.6	31.8	38.3	24.9	17.4	20.2	16.2	19.4
2004	12.0	15.1	17.7	20.1	22.5	24.7	26.9	29.8	35.4	23.5	16.9	19.6	16.0	19.6
2005	11.4	14.1	16.6	18.9	21.1	23.2	25.3	27.8	32.6	21.9	16.4	19.1	15.5	18.7
2006	10.9	13.3	15.4	17.3	19.3	21.2	23.2	25.7	29.7	20.1	15.8	17.8	15.0	17.8
2007	10.3	12.1	13.9	15.6	17.3	19.1	20.9	23.2	26.6	18.2	15.0	16.7	14.5	17.1
2008	9.4	11.0	12.4	13.7	15.1	16.6	18.4	20.8	24.1	16.2	14.2	15.8	14.0	16.4

Table 21: GS earnings above £8k. Quantiles and mean of the earnings reported from the Administrative Data. Uses the returns from 2012/13. To calibrate we also give results from the Silver sample of non-borrowers who have the same age profile as the GS.

Cohort	Male borrowers earnings (£000's)										Silver		Non-HE	
	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Mean	Q5	Mean	Q5	Mean
1998	15.2	20.1	24.5	28.2	31.6	35.4	39.4	45.8	58.9	38.5	24.8	31.5	23.9	30.5
1999	15.1	19.6	23.3	26.6	29.7	32.8	36.9	42.5	53.5	34.5	23.6	29.2	22.8	27.9
2000	14.4	18.7	22.2	25.5	28.5	31.5	35.2	40.2	50.0	32.0	23.0	28.5	22.3	27.9
2001	13.9	17.9	21.2	24.3	27.2	30.0	33.3	38.0	46.5	30.1	21.9	28.0	21.6	26.1
2002	13.1	16.6	19.6	22.4	25.2	27.8	30.8	35.1	43.0	27.6	20.8	25.8	20.0	24.4
2003	12.4	15.7	18.2	20.6	23.2	25.7	28.4	32.5	39.1	25.5	20.1	24.0	19.7	23.6
2004	11.4	14.2	16.5	18.8	21.2	23.5	26.3	29.7	35.0	23.2	19.1	22.5	18.9	22.3
2005	10.4	12.6	14.6	16.5	18.4	20.6	22.8	25.8	30.4	20.0	18.0	20.9	17.8	21.1
2006	9.5	11.0	12.4	13.8	15.3	16.8	18.8	21.9	26.7	17.2	17.2	19.8	17.5	20.7
2007	8.7	9.5	10.3	11.3	12.6	13.9	15.8	18.3	24.2	15.1	16.2	18.8	16.8	19.7
2008	8.6	9.3	10.2	11.4	12.8	14.5	16.6	20.9	26.5	15.6	15.3	17.9	15.7	18.4
	Female borrowers earnings (£000's)										Silver		Non-HE	
1998	13.0	17.6	21.4	24.9	28.3	31.6	34.8	38.9	47.2	30.2	20.5	24.8	19.0	23.6
1999	13.1	17.3	20.8	24.1	27.0	30.0	33.2	37.1	44.5	29.1	19.8	23.6	18.5	22.9
2000	13.0	16.9	20.2	23.3	26.1	29.0	31.8	35.7	42.3	28.0	19.3	23.1	18.2	22.3
2001	12.7	16.6	19.6	22.5	25.2	27.7	30.3	33.7	39.9	26.4	18.9	22.1	17.5	21.4
2002	12.4	16.0	18.8	21.4	23.7	26.0	28.3	31.4	37.1	24.9	18.4	21.1	17.0	20.7
2003	12.1	15.0	17.7	20.1	22.3	24.4	26.5	29.3	34.5	23.1	17.6	20.1	16.6	20.0
2004	11.4	14.0	16.3	18.3	20.5	22.4	24.5	27.0	31.4	21.2	17.1	19.2	16.1	18.7
2005	10.4	12.6	14.4	16.2	18.0	20.0	21.8	24.2	27.6	18.9	16.0	17.9	15.5	18.1
2006	9.5	11.0	12.3	13.6	15.1	16.7	18.7	21.5	25.1	16.5	15.3	17.0	15.4	17.7
2007	8.7	9.5	10.4	11.4	12.3	13.5	14.9	17.1	22.0	14.0	14.4	16.0	15.1	17.3
2008	8.6	9.3	10.2	11.2	12.6	14.0	16.0	19.1	23.6	14.5	13.7	15.3	14.1	16.3

Table 22: GS earnings above £8k. Quantiles and mean of the earnings reported from the Administrative Data. Uses the returns from 2010/11. To calibrate we also give results from the Silver sample of non-borrowers who have the same age profile as the GS.

Cohort	Male borrowers earnings (£000's)										Silver		Non-HE	
	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Mean	Q5	Mean	Q5	Mean
1998	15.6	20.3	24.6	28.2	31.6	35.1	39.3	45.0	56.3	36.8	24.6	31.2	23.9	30.2
1999	15.1	19.5	23.1	26.4	29.4	32.5	36.5	41.6	51.6	33.8	23.9	29.1	23.1	28.6
2000	14.3	18.7	21.9	25.1	28.0	31.1	34.5	39.1	48.0	31.1	22.7	28.3	22.1	26.9
2001	13.4	17.5	20.6	23.7	26.5	29.3	32.3	36.8	44.8	28.8	21.8	26.4	20.9	25.6
2002	12.9	16.4	19.2	21.9	24.7	27.1	29.8	33.5	40.1	26.4	20.9	25.6	20.5	24.5
2003	12.0	15.0	17.5	19.9	22.2	24.5	27.2	30.7	36.6	24.0	19.8	23.4	19.5	23.6
2004	10.8	13.4	15.6	17.6	19.7	22.0	24.5	27.8	32.6	21.4	18.7	22.0	18.5	21.9
2005	9.5	11.2	12.8	14.4	16.1	17.9	19.9	22.9	27.5	17.7	17.7	20.5	17.8	20.7
2006	8.8	9.7	10.6	11.8	13.0	14.5	16.5	19.4	25.1	15.4	17.0	19.2	17.4	20.2
2007	8.6	9.4	10.4	11.3	12.6	14.1	16.1	19.2	26.7	15.2	15.9	18.1	16.3	18.5
2008	8.6	9.2	9.8	10.9	12.1	13.4	15.7	18.8	25.9	15.1	14.8	17.9	15.2	17.6
	Female borrowers earnings (£000's)										Silver		Non-HE	
1998	13.7	18.0	21.9	25.6	28.9	32.1	35.0	39.1	47.2	30.5	20.8	25.0	19.4	24.4
1999	13.6	17.9	21.4	24.5	27.5	30.3	33.4	37.1	44.2	29.3	20.1	23.6	18.6	22.5
2000	13.2	17.1	20.6	23.6	26.3	29.1	31.8	35.2	41.8	27.7	19.2	22.8	18.0	21.8
2001	12.9	16.7	19.6	22.3	24.9	27.4	29.8	32.9	38.5	25.9	18.8	21.5	17.3	20.7
2002	12.4	15.8	18.5	21.2	23.4	25.5	27.6	30.5	35.9	24.1	18.3	20.8	17.3	20.7
2003	11.6	14.6	17.1	19.2	21.4	23.5	25.4	27.8	32.6	22.0	17.3	19.5	16.5	19.6
2004	10.6	13.1	15.1	17.0	19.0	21.1	22.8	25.5	29.1	19.8	16.7	18.6	16.1	18.8
2005	9.6	11.2	12.6	14.1	15.7	17.5	19.7	22.2	26.2	17.1	15.7	17.4	15.8	18.4
2006	8.7	9.5	10.4	11.5	12.6	13.9	15.7	18.3	23.4	14.5	14.9	16.3	15.4	17.5
2007	8.7	9.5	10.4	11.4	12.5	14.0	16.1	19.4	24.8	14.8	14.1	15.4	14.3	16.3
2008	8.5	9.1	9.8	10.8	12.0	13.0	14.6	16.9	21.1	13.6	13.4	14.8	13.6	15.7

Table 23: GS earnings above £8k. Quantiles and mean of the earnings reported from the Administrative Data. Uses the returns from 2009/10. To calibrate we also give results from the Silver sample of non-borrowers who have the same age profile as the GS.

Cohort	Male borrowers earnings (£000's)										Silver		Non-HE	
	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Mean	Q5	Mean	Q5	Mean
1998	15.9	20.5	24.5	28.0	31.0	34.4	38.4	44.2	55.2	35.4	24.5	29.9	23.4	28.9
1999	14.6	19.2	22.7	25.8	28.9	31.9	35.6	40.6	50.5	32.2	23.6	28.1	22.7	27.7
2000	14.2	18.4	21.6	24.5	27.4	30.3	33.6	38.4	46.5	30.0	22.5	26.7	21.9	26.4
2001	13.2	17.1	20.0	23.0	25.7	28.3	31.2	35.5	42.3	27.6	21.4	25.5	21.1	24.8
2002	12.3	15.7	18.3	20.8	23.2	25.7	28.4	31.9	37.8	24.9	20.5	24.2	20.1	24.0
2003	11.4	14.2	16.5	18.7	20.9	22.9	25.5	28.7	33.4	22.5	19.4	22.5	19.3	22.6
2004	9.8	11.9	13.7	15.3	17.1	19.0	21.5	24.7	30.0	19.0	18.2	21.2	18.5	21.7
2005	8.9	9.8	10.9	12.0	13.4	15.0	16.7	19.2	24.5	15.4	17.3	19.7	18.0	20.6
2006	8.8	9.7	10.6	11.7	12.8	14.2	16.0	19.0	25.2	15.4	16.3	18.4	16.9	19.4
2007	8.6	9.2	9.9	10.8	12.1	13.4	15.2	18.2	26.3	14.7	15.0	17.1	15.5	18.1
2008	8.5	9.0	9.7	10.6	11.7	13.4	16.1	20.8	27.4	15.6	14.1	16.7	14.3	16.8
	Female borrowers earnings (£000's)										Silver		Non-HE	
1998	14.3	18.6	22.4	25.8	28.8	31.4	34.4	38.6	46.1	30.2	20.7	24.5	19.2	23.7
1999	13.8	18.2	21.5	24.4	27.2	30.0	32.6	36.1	43.0	28.7	19.9	23.4	18.4	22.6
2000	13.2	17.2	20.3	23.1	25.6	28.0	30.6	34.1	40.6	26.8	19.2	22.1	18.0	22.1
2001	13.0	16.6	19.2	21.9	24.1	26.2	28.5	31.6	37.4	25.0	18.3	21.0	17.0	20.5
2002	12.2	15.4	17.8	20.1	22.3	24.1	26.2	29.0	33.8	22.9	17.7	20.0	17.0	20.3
2003	11.2	13.7	15.9	17.9	19.9	22.0	23.9	26.3	30.3	20.6	16.8	18.8	16.2	18.8
2004	9.9	11.8	13.4	15.0	16.6	18.3	20.3	22.8	26.9	17.8	16.1	17.8	16.0	18.6
2005	8.9	9.8	10.9	11.9	13.1	14.4	16.1	18.4	23.2	14.8	15.1	16.6	15.5	17.8
2006	8.7	9.5	10.4	11.5	12.8	14.2	16.3	19.4	24.2	14.8	14.3	15.8	14.8	17.0
2007	8.5	9.2	9.9	10.8	12.0	13.3	14.7	16.8	20.7	13.7	13.5	14.8	13.9	15.9
2008	8.5	9.0	9.6	10.3	11.3	12.6	14.5	17.3	21.5	13.6	12.9	14.7	13.3	15.6

Table 24: GS earnings above £8k. Quantiles and mean of the earnings reported from the Administrative Data. Uses the returns from 2008/09. To calibrate we also give results from the Silver sample of non-borrowers who have the same age profile as the GS.

Cohort	Non-HE sample										Silver sample									
	Male non-HE estimates (£000's)										Male non-borrowers (£000's)									
	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Mean	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Mean
1998	11.7	14.8	17.8	20.5	23.5	26.9	30.8	36.8	49.0	30.2	11.9	15.2	18.2	21.1	24.4	27.9	32.0	38.1	49.8	31.4
1999	11.4	14.6	17.0	19.5	22.3	25.3	29.4	34.5	45.2	27.8	11.8	15.0	17.7	20.5	23.6	26.7	30.8	36.3	46.8	29.0
2000	11.1	14.0	16.5	19.0	21.6	24.7	28.3	33.2	42.8	26.6	11.5	14.7	17.4	19.9	22.7	26.1	29.9	34.9	44.4	28.8
2001	11.0	13.8	16.2	18.4	20.7	23.4	26.7	31.4	40.1	25.5	11.2	14.2	16.7	19.1	21.6	24.7	28.3	33.1	42.0	27.7
2002	10.8	13.3	15.6	17.8	20.1	22.6	26.0	30.3	38.5	24.4	11.1	13.8	16.2	18.5	20.9	23.7	26.9	31.5	39.5	26.1
2003	10.7	13.1	15.3	17.3	19.6	22.2	25.3	29.4	36.4	23.5	10.8	13.4	15.8	17.9	20.1	22.8	26.0	30.1	37.5	24.1
2004	10.7	13.2	15.0	16.8	18.7	21.0	23.7	27.4	34.3	22.5	10.7	13.2	15.2	17.1	19.1	21.5	24.3	28.2	34.9	22.9
2005	10.3	12.8	14.4	16.2	18.0	20.1	22.8	26.6	32.8	21.5	10.3	12.7	14.5	16.2	18.1	20.3	23.0	26.7	32.7	21.5
2006	10.0	12.2	13.9	15.6	17.5	19.4	21.7	25.3	30.7	20.4	10.1	12.1	14.0	15.7	17.4	19.4	22.0	25.4	30.8	20.3
2007	9.9	11.8	13.4	15.0	16.7	18.7	21.1	24.7	30.0	19.8	9.9	11.8	13.4	15.0	16.6	18.5	20.8	23.9	29.2	19.3
2008	9.8	11.8	13.4	14.7	16.1	17.8	19.9	22.8	27.4	18.6	9.5	11.3	12.9	14.3	15.8	17.4	19.4	22.3	27.2	18.3
	Female non-HE estimates (£000's)										Female non-borrowers (£000's)									
1998	9.7	11.8	13.7	15.8	18.2	21.3	25.0	30.5	39.9	23.7	10.0	12.3	14.7	17.1	19.9	23.1	27.2	32.7	41.4	24.3
1999	9.7	11.7	13.8	15.8	18.1	20.6	24.1	29.2	37.9	22.7	10.0	12.2	14.4	16.8	19.4	22.4	26.4	31.4	39.2	23.4
2000	9.6	11.3	13.3	15.4	17.4	19.8	22.8	27.2	36.3	21.9	9.9	12.0	14.3	16.5	18.8	21.7	25.2	30.2	38.0	22.7
2001	9.7	11.6	13.4	15.4	17.4	19.5	22.4	26.4	33.7	21.3	9.9	12.0	14.1	16.2	18.4	21.1	24.3	28.6	35.5	21.9
2002	9.7	11.5	13.3	15.1	17.2	19.1	21.8	25.4	32.6	20.8	9.9	12.0	14.0	16.0	18.2	20.5	23.5	27.8	34.3	21.2
2003	9.6	11.5	13.2	15.0	16.7	18.4	20.7	24.1	31.0	19.9	9.9	12.1	13.8	15.8	17.6	19.6	22.3	26.2	32.2	20.2
2004	9.5	11.0	12.6	14.2	15.8	17.7	20.0	23.5	29.6	19.0	9.7	11.6	13.3	15.1	16.9	18.9	21.5	25.0	30.6	19.4
2005	9.6	11.3	12.7	14.2	15.7	17.1	18.9	21.9	27.7	18.7	9.7	11.5	13.1	14.6	16.2	17.8	20.0	23.1	28.2	18.3
2006	9.4	11.0	12.6	13.9	15.2	16.6	18.5	21.1	26.0	17.7	9.6	11.3	12.9	14.3	15.6	17.2	19.1	21.9	26.5	17.4
2007	9.3	10.7	12.1	13.3	14.6	16.0	17.5	20.0	24.6	17.0	9.3	10.8	12.2	13.4	14.7	16.1	17.8	20.4	24.7	16.3
2008	9.6	11.0	12.2	13.4	14.6	15.8	17.4	19.4	23.5	16.8	9.2	10.4	11.7	12.8	14.0	15.3	16.8	18.9	22.7	15.6

Table 25: Earnings above £8k. Silver Sample and non-HE (corrected Silver) sample quantiles and mean of the earnings reported from the Administrative Data. Uses the returns from 2011/12.

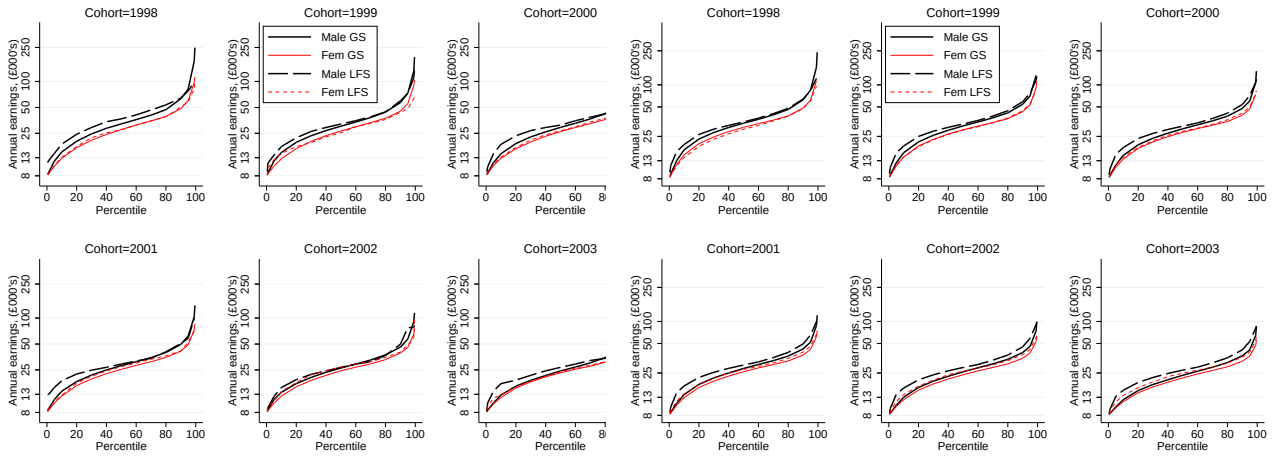


Figure 12: Estimates of the LFS and GS conditional earnings distributions, conditioning on earnings being higher than £8k. The results for the 1998 to 2003 cohorts are shown. Throughout, the y-axis shows earnings on a log scale. We graph our estimates of the quantiles of the earnings for males and females using our two data sources. LFS earnings are weighted using the LFS population weights.

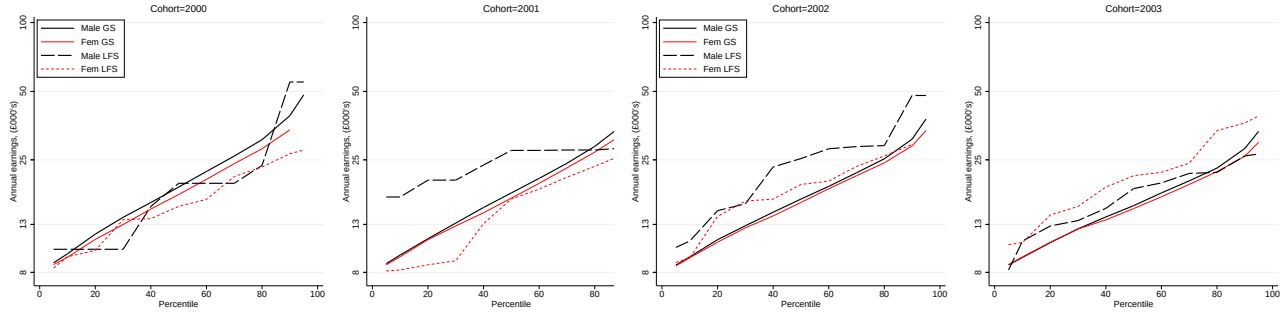


Figure 13: Joiners: LFS & GS earnings quantiles t given low earnings $t-1$, 2000, 2001, 2002 & 2003 cohorts. LFS earnings are weighted using the LFS population weights.

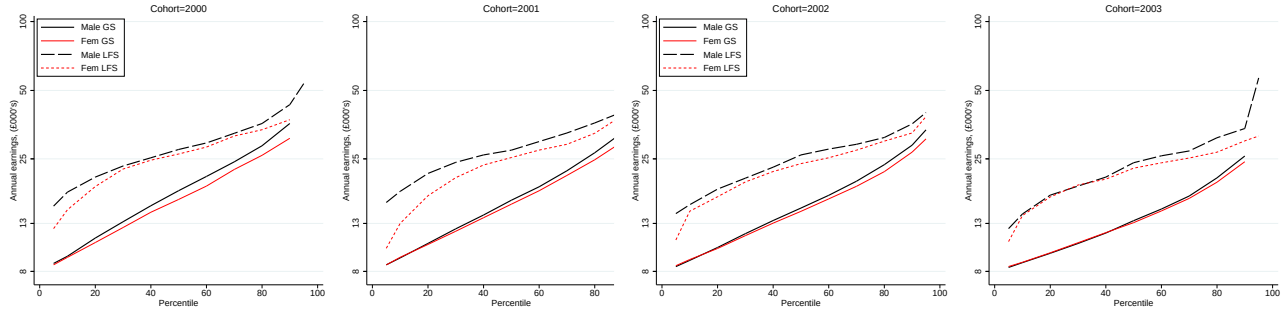


Figure 14: Leavers: LFS & GS earnings quantiles $t-1$ given low earnings t, 2000, 2001, 2002 & 2003 cohorts. LFS earnings are weighted using the LFS population weights.

Cohort	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Mean	N
	Male Golden Sample Borrowers										
1998	0.92	0.97	0.98	0.99	0.99	0.97	0.97	0.94	0.90	0.90	4,764
	(0.02)	(0.01)	(0.00)	(0.00)	(0.00)	(0.00)	(0.01)	(0.01)	(0.02)	(0.01)	
2000	0.93	0.97	0.98	0.98	0.98	0.97	0.95	0.92	0.83	0.86	7,557
	(0.02)	(0.01)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.01)	(0.02)	(0.01)	
2001	0.93	0.97	0.98	0.98	0.97	0.95	0.93	0.88	0.82	0.88	7,762
	(0.01)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.01)	(0.01)	(0.01)	(0.01)	
2002	0.93	0.97	0.98	0.98	0.96	0.94	0.92	0.87	0.79	0.86	7,588
	(0.02)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.01)	(0.01)	(0.01)	(0.01)	
2003	0.93	0.97	0.98	0.97	0.95	0.93	0.89	0.84	0.76	0.85	7,422
	(0.02)	(0.00)	(0.00)	(0.00)	(0.00)	(0.01)	(0.01)	(0.01)	(0.02)	(0.01)	
2004	0.92	0.96	0.96	0.94	0.91	0.88	0.84	0.79	0.73	0.81	6,953
	(0.02)	(0.01)	(0.00)	(0.00)	(0.00)	(0.01)	(0.01)	(0.01)	(0.02)	(0.01)	
2005	0.90	0.94	0.92	0.89	0.84	0.79	0.73	0.67	0.64	0.75	6,618
	(0.02)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.02)	(0.01)	
2006	0.84	0.90	0.85	0.79	0.74	0.68	0.62	0.58	0.59	0.69	5,277
	(0.02)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.02)	(0.02)	(0.01)	
	Female Golden Sample Borrowers										
1998	0.84	0.94	0.97	0.98	0.97	0.95	0.92	0.89	0.82	0.86	5,157
	(0.03)	(0.02)	(0.01)	(0.00)	(0.00)	(0.00)	(0.01)	(0.01)	(0.02)	(0.01)	
1999	0.85	0.96	0.98	0.98	0.96	0.94	0.90	0.86	0.79	0.86	8,711
	(0.03)	(0.01)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.01)	(0.01)	(0.01)	
2000	0.90	0.96	0.98	0.98	0.97	0.95	0.91	0.86	0.78	0.86	8,982
	(0.03)	(0.01)	(0.00)	(0.00)	(0.00)	(0.00)	(0.01)	(0.01)	(0.01)	(0.01)	
2001	0.89	0.98	0.98	0.98	0.96	0.93	0.89	0.83	0.73	0.84	8,944
	(0.03)	(0.01)	(0.00)	(0.00)	(0.00)	(0.00)	(0.01)	(0.01)	(0.02)	(0.01)	
2002	0.88	0.96	0.97	0.97	0.94	0.91	0.86	0.79	0.71	0.81	8,761
	(0.03)	(0.01)	(0.00)	(0.00)	(0.00)	(0.00)	(0.01)	(0.01)	(0.01)	(0.01)	
2003	0.82	0.94	0.95	0.93	0.90	0.86	0.80	0.73	0.65	0.76	8,639
	(0.03)	(0.01)	(0.00)	(0.00)	(0.00)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	
2004	0.85	0.93	0.93	0.88	0.84	0.77	0.69	0.62	0.56	0.72	8,390
	(0.03)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.02)	(0.01)	
2005	0.78	0.88	0.84	0.79	0.72	0.64	0.57	0.50	0.48	0.62	7,147
	(0.02)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.02)	(0.01)	

Table 26: Coefficients from quantile and mean autoregressions of log-earnings (2011/12) on lagged log-earnings (2010/11) for the Golden Sample. Figures in brackets are estimated standard errors for the persistence coefficient $\beta_1(\tau)$, where τ is the quantile level.