Examining sample representativeness and data quality in the linked Next Steps survey and Student Loans Company administrative data

CLS working paper number 2024/1

By Charlotte Booth[1], Claire Crawford[2], Nasir Rajah[1], Richard J Silverwood[1], Morag Henderson[1]

[1] Centre for Longitudinal Studies, UCL Social Research Institute, 20 Bedford Way, London, WC1H 0AL
[2] Centre for Education Policy & Equalising Opportunities, UCL Faculty of Education and Society, 9-11 Endsleigh Gardens, London, WC1H 0EH

CENTRE FOR LONGITUDINAL STUDIES

Economic and Social Research Council

Contact the author
Charlotte Booth
UCL Centre for Longitudinal Studies
charlotte.booth@ucl.ac.uk

This working paper was first published in February 2024 by the UCL Centre for Longitudinal Studies.
UCL Social Research Institute
University College London
20 Bedford Way
London WC1H 0AL
www.cls.ucl.ac.uk

The UCL Centre for Longitudinal Studies (CLS) is an Economic and Social Research Council (ESRC) Resource Centre based at the UCL Social Research Institute, University College London. It manages four internationally-renowned cohort studies: the 1958 National Child Development Study, the 1970 British Cohort Study, Next Steps, and the Millennium Cohort Study. For more information, visit www.cls.ucl.ac.uk.

This document is available in alternative formats. Please contact the Centre for Longitudinal Studies.
Tel: +44 (0)20 7612 6875
Email: clsfeedback@ucl.ac.uk

Disclaimer

This working paper has not been subject to peer review.
CLS working papers often represent preliminary work and are circulated to encourage discussion. Citation of such a paper should account for its provisional character. A revised version may be available directly from the author.

Any opinions expressed here are those of the author(s) and not those of the UCL Centre for Longitudinal Studies (CLS), the UCL Social Research Institute, University College London, or the Economic and Social Research Council.

Find out more

Email: clsfeedback@ucl.ac.uk
Visit: cls.ucl.ac.uk
Follow: @CLScohorts

How to cite this paper

Booth, C., Crawford, C., Rajah, N., Silverwood, R., Henderson, M. (2024) *Examining sample representativeness and data quality in the linked Next Steps survey and Student Loans Company administrative data*. CLS Working Paper 2024/1. London: UCL Centre for Longitudinal Studies.

Abstract

Linked cohort and administrative data provide rich data resources with wide ranging research possibilities. Yet, it is important to understand the quality and representativeness of linked cohort and administrative data, in order to establish the reliability of estimates from these. At the Next Steps age 25 survey (2015/16) participants were asked for their consent to link their survey data to various administrative records. One of which was the Student Loans Company (SLC), who provided data for consenting participants on any student loan applications, payments, and repayments made across a fourteen-year period (from 2007-2021). We examined sample representativeness for the linked Next Steps-SLC data by comparing participant characteristics between those who did and did not consent to data linkage, for those who were and were not successfully linked, and relative to national population statistics where possible. Among age 25 respondents, certain groups were less likely to consent to data linkage, including those from ethnic minority groups and more disadvantaged backgrounds. In the limited instances where comparable national data was available, the linked sample was found to be reasonably representative of these wider populations. Data quality was examined by evaluating agreement between similar variables held across both sources, which on the whole, revealed a high level of agreement suggesting high data quality. Finally, a novel policy-relevant research question was investigated, to showcase some of the research possibilities of using this linked data. We conclude that future research should capitalise on this new linked data resource for investigating a range of outcomes among student loan borrowers but note that certain groups may be under-represented in the linked data, due to differential linkage consent rates.

Keywords

Administrative data; Cohort studies; Data linkage; Next Steps; Student Loans Company; Linkage quality; Representativeness.

Introduction

Over the past decade, there have been more examples of studies linking cohort and administrative data, with the aim of enhancing research possibilities (Calderwood & Lessof, 2009). Interest in using linked cohort and administrative data is growing due to the richness of information which augmented datasets can provide (Harron et al., 2017). For example, administrative data can be used to fill in gaps in information between typical waves in longitudinal cohort studies (Peycheva, Ploubidis, & Calderwood, 2021). Yet, it is necessary to establish the quality and sample representativeness of linked cohort and administrative data, to establish the likely reliability of estimates from these data resources.

Linked data may be subject to multiple levels of selection and may not be representative for a number of reasons, including: (1) selection into the survey, (2) selective sample attrition in longitudinal studies, (3) selective consent to data linkage, and (4) differential linkage error (Silverwood et al., 2024). For example, a previous study found that Next Steps cohort members were less likely to consent to administrative data linkage if they were either female or from an ethnic minority background (Peycheva et al., 2021). Differential linkage error can also occur either when there is a failure to match individuals to their records (i.e., missed matches) or when false links occur between unrelated records (i.e., false matches) (Silverwood et al., 2024).

Silverwood et al. (2024) have provided a framework for assessing linkage quality. The first approach is to compare linked data to a gold-standard dataset, where false matches and missed matches can be quantified and the sources of potential biases can be uncovered. This however is only possible where the true match status of the pair of records is known. Another approach is to compare linked data to external population estimates (Harron, Doidge, & Goldstein, 2020), or to survey populations known to be nationally representative (Silverwood et al., 2024), as long as the external benchmark is aligned to the surveys' target population.
To examine data quality of linked cohort and administrative data, one can compare similar variables held in both sources and evaluate their comparability. For instance, where income is measured in survey data and also held in administrative data, one can evaluate the within-individual agreement between variables. Discrepancies can arise due to differences in how the constructs are assessed in either data source and/or potential measurement error, which can be informative about the quality of information provided by one or both sources.

The current study

We examined sample representativeness and data quality in the recently linked Next Steps and Student Loans Company (SLC) data. Our objectives were to (i) evaluate sample representativeness across key participant characteristics of the linked sample with that of the wider study sample and, where possible, external population data, (ii) assess data quality by comparing similar variables held across both sources, and (iii) investigate a novel policy-relevant research question to showcase some of the research possibilities of these linked data.

Background and context

Next Steps is a longitudinal cohort study following a nationally representative sample of around 16,000 people born in England in 1989-90 (Calderwood et al., 2021). The study began in 2004, when participants were 14 years old, in Year 9 at school. Annual surveys were conducted by the Department for Education up until age 20 (2010/11). The study was then taken over by the Centre for Longitudinal Studies (CLS), who conducted data collection at age 25 (2015/16), with another wave currently underway at age 32 (2022/23).

The SLC is a non-profit government-owned organisation that has, since 1990-91, administered loans to students attending higher education (HE) colleges and universities in the United Kingdom (Bolton, 2019). Most of those entering HE do so at either age 18 (i.e., straight from school) or age 19 (i.e., after an additional year of college or a "gap year"). Consequently, Next Steps cohort members mostly accessed HE for the first time in the academic years 2008-09 or 2009-10.

In the academic years between 2006-07 and 2011-12, in England, HE students could apply for income-contingent loans to cover the cost of tuition fees (~£3,000) and a contribution towards their living costs (i.e., maintenance loans) of up to ~£6,000 in London and ~£4,500 outside of London. Means-tested maintenance grants were also available to those from lower income households (up to ~£3,000), covering part of the maintenance loan, which being a grant, would not need to be paid back.

Average student debt upon graduation for those at the time of the Next Steps cohort, was around £20,000 (Bolton, 2022). Interest rates for these loans, and all those pre-2012, were set in line with inflation and hence had a zero real interest rate (Bolton, 2022). Student loan repayments were income-contingent, set at 9% of income above a pre-defined threshold (see Appendix A for thresholds between 2012-2021), with a maximum term of 25 years after which they would be written off.

Data description and linkage

Next Steps survey

The Next Steps age 25 survey was conducted between August 2015 and September 2016,[3] with an achieved sample of 7,707 individuals. An important aspect of the survey was the request for administrative data linkage. Participants were asked for their consent to link their survey records to nine different administrative data sources from government departments and non-governmental bodies, including education, health, and economic records (Calderwood et al., 2021).

Among all respondents, 77% agreed to at least one type of data linkage (Calderwood et al., 2021). The consent rate for SLC linkage was slightly lower at 58% (Rihal, Gomes, & Henderson, 2021). This may reflect self-selection, as some participants may not have consented if they did not attend university or take out student loans, although our analysis suggests this is likely to explain only a small part of the difference. Alternatively, and more likely, it may reflect the lower consent rates observed for economic records more generally, e.g., consent rates were 57% for data held by HM Revenue and Customs, and 59% for data from the Department for Work and Pensions, compared to 70% for education data.

SLC data linkage

Data linkage was completed by the SLC in June 2021, using personal identifiers including first name, surname, sex, date of birth, address, and National Insurance Number (NINO) for those who supplied it. Matching was done on at least three personal identifiers and no fuzzy matching took place, which would allow for small differences in the matching of an individual identifier (e.g., in the case of misspelling). Further details about the linkage procedure can be found in the data linkage user guide (Rihal et al., 2021). Data can be accessed via the secure server on the UK Data Service and are registered under the study number (SN)8848 (UK Data Service, 2022).

A participant flowchart including the consent and linkage rate is shown in Figure 1. Of the total survey respondents, 4,501 consented to SLC data linkage (58%), and of those, 2,219 were successfully linked (49%). However, the number of participants who consented to SLC data linkage exceeded the number who reported ever going to university (*n* = 3,622). Therefore, many of those who consented were unlikely to have been eligible for student loans, making it difficult to evaluate the linkage rate based on the total number of consenters.

[3] As survey data collection crossed two tax years, a variable indicating which tax year participants completed the survey (1 = April 2015-March 2016; 2 = April-September 2016) was created in order to compare accurately with SLC data, which is mostly structured by tax year. However, exact survey date was unknown, therefore anyone who completed the survey between 1-5 of April may have been misclassified.

Therefore, we additionally calculated the consent and linkage rate conditional upon those who ever went to university (*n* = 3,622).[4] Among this potentially eligible sample, a similar proportion consented to data linkage (62%), suggesting that individuals who went to university were only slightly more likely to consent to SLC data linkage than respondents overall. However, a much higher proportion were successfully linked (80%), supporting the chosen eligibility criteria. This figure also reflects national estimates for the proportion of university students who took out student loans at the same time as Next Steps, which was 80% in 2008-09 and 83% in 2009-10 (Bolton, 2022), although we cannot be sure that this is the only reason for individuals to be omitted from the linked data. Indeed, as shown later, the data provides some indications that there may have been at least some missed matches.

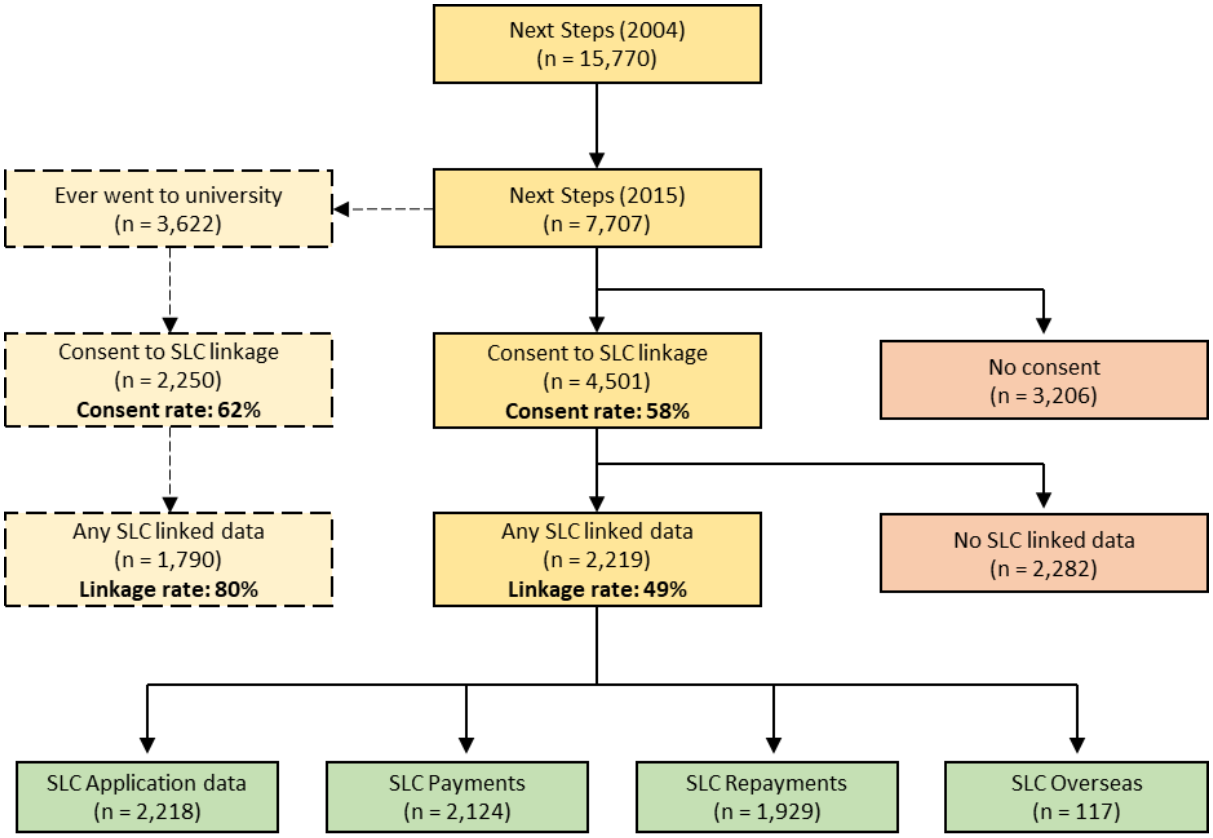

Figure 1. Participant flow-chart.

[4] In principle individuals could have taken out a loan to access higher education in a further education college – and hence might not be captured by the ever-attended university measure.

SLC datasets

Four datasets were provided by the SLC for participants who consented to data linkage and were successfully linked (*n* = 2,219). The datasets reflect: (i) Applications (*n* = 2,218), (ii) Payments (*n* = 2,124), (iii) Repayments (*n* = 1,929), and (iv) Overseas students (*n* = 117).[5]

SLC Application data

The SLC application dataset holds information on 2,218 individuals about student loan applications made across a 13-year period (from 2007 to 2020). Data are in long format structured by academic year. There are 23 variables, including the university and course applied to (reflecting the university and course from which they had accepted an offer), and the amount of loan requested in each year according to the different support types, e.g., tuition fee support (*n* = 2,114), maintenance support (*n* = 2,141), and means-tested maintenance support (*n* = 1,634).

SLC Payments data

The SLC payments dataset holds information on 2,124 individuals about loan payments made across a 14-year period (from 2007 to 2021). Data are in long format structured by tax year end[6] and support type. There are 95 fewer individuals in the payments than applications dataset,[7] which may reflect individuals who applied for loans but never took them out. Those who did receive payments, mostly received both full-time tuition fee (*n* = 1,964) and full-time maintenance support loans (*n* = 2,001). While a much smaller number of part-time tuition fee loans were taken (*n* = 77).

First payment received

We analysed the year that a first payment was received, shown as a cumulative probability plot in Figure 2.[8] The majority of first loan payments were received either in the academic year 2008-09 (54%), or 2009-10 (28%),[9] corresponding to when cohort members would be 18 or 19 years old. Further analyses revealed that most participants received student loan payments across four tax years (52%), which most likely corresponds to three academic years.[10]

---

[5] The overseas dataset holds information on 117 participants who moved overseas, including the date of departure, date of return, and country of residence, which was not analysed here.

[6] Tax year end 2007 refers to the period between 6th April 2006 and 5th April 2007.

[7] One individual appears in the payments data without a corresponding application record.

[8] One individual appears in the payments data but was recorded as receiving £0 in loans, which was recoded as missing for analysis.

[9] On the basis that most students start university in the first term of each academic year, we assume that payments first received in the tax year ending 2009 (5th April 2009) correspond to the academic year 2008-09, and that first payments received in the tax year ending 2010 correspond to the academic year 2009-10.

[10] Similarly, and on the basis that student loan payments are made termly, we assume that an individual studying for a three-year degree would receive loan payments across four tax years. For example, an individual starting university in September 2008 and graduating in July 2011 would receive

A substantial proportion (21%) received payments across five tax years (or four academic years). Smaller proportions received payments across one tax year (2%), two tax years (6%), three tax years (7%), six tax years (9%), seven tax years (3%), or more (1%).

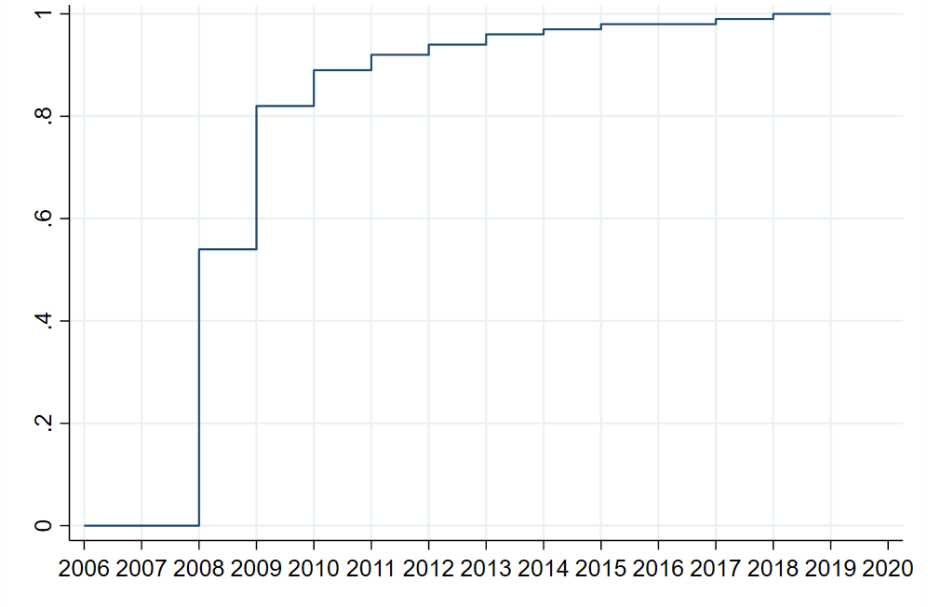

Figure 2. Cumulative probability plot of academic year beginning when first payment was received (*n* = 2,123).

SLC Repayments data

The SLC repayments dataset holds information on 1,929 individuals about loan repayments made across a 12-year period (from 2009 to 2021). Data are in long format structured by tax year end and repayment type, including: (i) PAYE, (ii) Self-assessment, (iii) Obligatory UK, (iv) Obligatory overseas, (v) Voluntary UK, (vi) Voluntary overseas, (vii) Refunds, and (viii) Overpayments. Before tax year end 2020, the most common repayment type was PAYE (*n* = 1,835), reflecting repayments taken automatically from employed persons pay, and Self-assessment (*n* = 199), reflecting repayments from self-employed persons via tax returns. However, after 2020, these repayment types were no longer observed in the data, instead recorded as 'Obligatory UK' only (*n* = 1,479). Therefore, it appears that it is not possible to discern repayments from employed or self-employed persons in the SLC data after 2020. In this paper, we only analysed the first four repayment types, i.e., any type of obligatory repayment, which should reflect 9% of income above the specified repayment threshold for that year (Appendix A).

loan payments in tax years 2008-09, 2009-10, 2010-11 and 2011-12, while an individual starting a four-year degree at the same time would additionally receive a loan payment in tax year 2012-13.

First repayment made

We analysed the number of years after receiving a final payment before participants made their first repayment, shown as a cumulative probability plot in Figure 3. Eleven percent of participants who had received a payment from the SLC had made no repayment by 2021. However, most of the sample made their first repayment within one year of receiving their last payment (56%), or within two years (70% - cumulative), indicating that most of the sample were earning above the repayment threshold soon after leaving university.

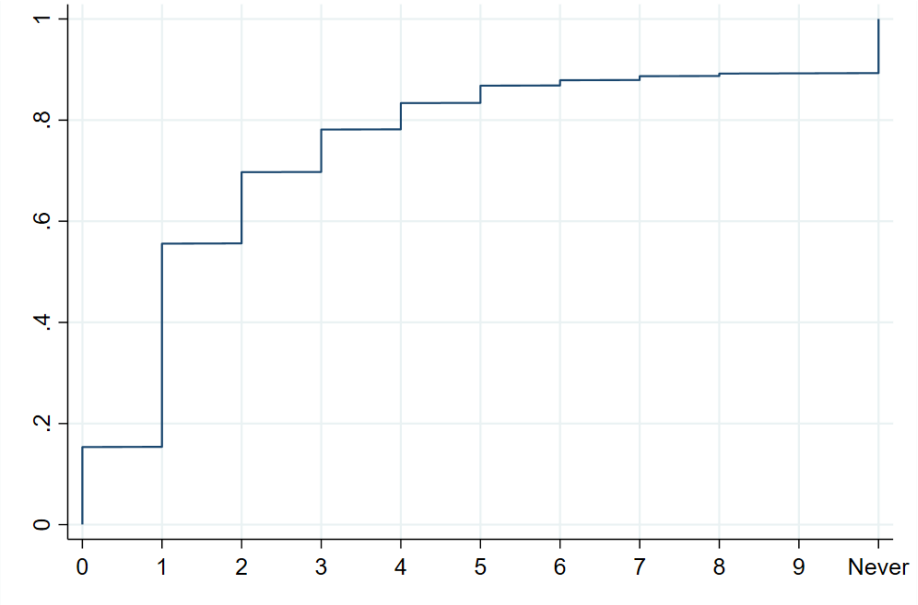

Figure 3. Cumulative probability plot of years after last payment before making first repayment (*n* = 2,123).

Amount of loan repaid

We calculated the total amount of loan repayments made within five years of receiving a final payment (*n* = 1,961). The average (mean) amount of loan repayments per participant was £1,962 (SD = £2,593), and the average total loan payments received was £20,264 (SD = £8,254). Therefore, the average proportion of loans repaid of the total after five years was 9.7%.

Sample representativeness

Assessing sample representativeness requires defining the underlying population of interest to which the sample is being compared, which depends on the research question of interest. In this section, we focus on comparing (a) individuals who consented vs. did not consent to data linkage and (b) consenters for whom the linkage was successful vs. unsuccessful, within (i) all respondents at the age 25 survey in Next Steps, and (ii) a potentially eligible sample of respondents who ever went to university.[11] Comparisons were also made to external population statistics of HE students at a similar point in time. Unfortunately, there are no published statistics documenting the characteristics of student loan borrowers, so we are unable to make direct comparisons to this population of interest.

Analyses used survey design and attrition weights from the age 25 survey to restore sample representativeness relative to the target population of Next Steps (pupils in England who were in Year 9 in 2004) (Calderwood et al., 2021). However, as these weights were designed for use on the full sample of respondents at age 25, and may not be as effective for specific subgroups, such as university attenders or the linked sample.

All age 25 respondents

We first compared background characteristics and educational attainment outcomes among the full sample of respondents at age 25 (*N* = 7,707), according to whether consent for SLC data linkage was given (*n* = 4,501) or not (*n* = 3,206) – see Table 1. As shown in Columns 2-5, consenters were more likely than non-consenters to have a parent with a degree (17% vs. 13%) and be from a White ethnic background (88% vs. 79%) compared to any ethnic minority background, apart from mixed ethnicity where consent rates did not differ. Consenters were slightly less likely to be female (48% vs. 52%), and less likely to have been eligible for free school meals at age 16 (14% vs. 17%). Consenters showed higher educational attainment, including being more likely to have achieved 5 or more GCSE grades A*- C (48% vs. 42%), to have reported ever attending university (39% vs. 33%), and also completing a university degree by age 25 (29% vs. 23%).

However, it is still the case that many of those who consented to SLC data linkage never reported attending university and were therefore unlikely to have been eligible for student loans. As shown in Columns 6-9, perhaps due to this ineligibility, differences in participant characteristics were observed between those who were (*n* = 2,219) and were not (*n* = 2,282) successfully linked. For example, those who were successfully linked were far more likely to have a parent with a university degree (27% vs. 10%) and more likely to be from an ethnic minority background (e.g., 3% vs. 1% Indian) compared to those who consented to data linkage but were not found in the SLC data. Therefore, as above, further analyses were conducted on a potentially eligible sample, who reported attending university.

[11] University attendance, as measured in the survey, was used to indicate eligibility for student loans. However, 19% (*n* = 429) of those who were linked to SLC data did not report ever attending university. Some of these (*n* = 60) were participants who appeared in the applications dataset without a corresponding record of loan payments, so may have applied to university but not gone. Others may have attended further education colleges that are not considered universities, or simply misreported university attendance in the survey (e.g., if they dropped out early).

Those who ever attended university

The same analyses were conducted conditional upon participants reporting ever attending university (*N* = 3,622) – see Table 2. The average characteristics of those who attended university (Column 1, Table 2) differed from the full sample of age 25 respondents (Column 1, Table 1). University attenders were more likely to be female, to have a parent with a university degree, and be from an ethnic minority background. They were less likely to have been eligible for free school meals, and perhaps unsurprisingly, showed higher educational attainment (e.g., 79% of university attenders achieved 5 or more A*-C grades at GCSE, compared to 46% of the full sample).

However, similar differences were observed between consenters (*n* = 2,250) and non-consenters (*n* = 1,372) among this eligible sample of university attenders. Consenters were more likely than non-consenters to have a parent with a degree (31% vs. 24%), and more likely to be from a White ethnic background (83% vs. 68%) compared to any ethnic minority, apart from mixed ethnicity. They also showed slightly higher educational attainment, including a higher rate of good GCSE grades (81% vs. 74%), and completion of a university degree by age 25 (67% vs. 59%).

Yet, in contrast to previous results, fewer differences in participant characteristics were observed between those who were (*n* = 1,790) and were not (*n* = 460) successfully linked (Columns 6-9, Table 2). Although, a lower proportion of females (49% vs. 62%) and a slightly higher proportion of Pakistani/Bangladeshi (3% vs. 2%) individuals were observed in the linked sample. Together, this suggests that, among university attenders who consented to data linkage, little additional bias was introduced by the possible failure to link participants to their administrative data.

Compared to national population statistics

Unfortunately, there are no published statistics documenting the characteristics of student loan borrowers, so we are unable to make direct comparisons to this population of interest. However, Britton et al. (2019) report on the gender split of a roughly 10% sample of student loan borrowers for a range of cohorts, including 2008-09, who they were able to successfully link to PAYE or Self-Assessment records from HMRC. They found that 57% of this sample of student loan borrowers were female (43% male), which is slightly higher than the proportion of females (50%) in the linked Next Steps-SLC sample (*N* = 2,219), perhaps due to the lower consent rate observed in females.

In addition, there are published summaries of the characteristics of students in HE more generally. For example, data from the Higher Education Statistics Agency (HESA) shows that among full-time first-degree UK-domiciled students attending UK HE institutions in 2008-09, 55% were female and 21% were from an ethnic minority background.[12] This compares to 50%

[12] Source (including some authors' calculations): https://www.hesa.ac.uk/data-and-analysis/publications/students-2008-09/introduction

female and 17% from an ethnic minority background among the linked Next Steps-SLC sample, which again, may be lower due to the lower consent rates observed in these groups.

However, HESA statistics also show that among 18/19-year-old first-degree UK-domiciled students attending UK HE institutions in 2008-09, just over 1% were attending part-time.[13] This is in line with the linked Next Steps-SLC sample, in which less than 1% applied for part-time loans in the same year. Similarly, HESA statistics show that 26% of full-time undergraduate students in 2008-09 attended a Russell Group university,[14] which was also estimated to be 26% in the Next Steps-SLC linked sample for the same year.

In summary, the characteristics of the linked sample who attended university (Column 7, Table 2) appear to be broadly in line with the characteristics of age 25 respondents who reported ever going to university (Column 1, Table 2), and indeed to wider populations for which we have been able to make comparisons. This provides some reassurance of the representativeness of the linked data, albeit on relatively limited dimensions. The main difference is in terms of the proportion of the sample who are female or from an ethnic minority background, which is lower in the linked sample than in the overall sample who reported ever going to university, driven primarily by the lower consent rate among females and ethnic minority individuals.

[13] Source: authors' calculations using https://www.hesa.ac.uk/data-and-analysis/publications/students-2008-09/introduction. These figures were calculated by taking the total number of UK domiciled first degree students (and the total number attending part-time) from Table D, calculating the number of 18- and 19-year-old first year UK domiciled students in total and attending part-time by combining these figures with the percentages reported in Table Ii, and then using these numbers to calculate the percentage of 18- and 19-year-old UK domiciled first year students attending part-time.
[14] Source (author calculated from Table 0): https://www.hesa.ac.uk/data-and-analysis/publications/students-2008-09

Table 1. Participant characteristics according to consent and data linkage on the full sample of survey respondents (*N* = 7,707).

	(1) Respondent	(2) No consent	(3) Consent	(4) RR	(5) 95% CI	(6) Not linked	(7) Linked	(8) RR	(9) 95% CI
Sex									
Male	3,426 (50.8%)	1,365 (48.3%)	2,061 (52.5%)	1.00		1,083 (54.3%)	978 (49.7%)	1.00	
Female	4,281 (49.2%)	1,841 (51.7%)	2,440 (47.5%)	0.94***	0.90-0.98	1,199 (45.7%)	1,241 (50.3%)	1.12**	1.03-1.22
Parent degree status									
No degree	6,269 (84.9%)	2,697 (87.5%)	3,572 (83.3%)	1.00		1,977 (89.8%)	1,595 (73.1%)	1.00	
Degree	1,381 (15.1%)	480 (12.5%)	901 (16.7%)	1.13***	1.08-1.19	293 (10.2%)	608 (26.9%)	1.84***	1.70-1.98
Ethnicity									
White	5,262 (84.6%)	1,915 (79.4%)	3,347 (88.0%)	1.00		1,843 (91.3%)	1,504 (82.7%)	1.00	
Mixed	368 (2.9%)	156 (3.1%)	212 (2.8%)	0.92	0.82-1.04	98 (2.4%)	114 (3.3%)	1.26*	1.03-1.54
Indian	518 (2.3%)	271 (3.1%)	247 (1.8%)	0.75***	0.67-0.84	79 (1.0%)	168 (3.2%)	1.84***	1.64-2.06
Pakistani & Bangladeshi	802 (3.7%)	466 (5.2%)	336 (2.7%)	0.70***	0.63-0.77	143 (2.1%)	193 (3.6%)	1.43***	1.25-1.63
Black & Black British	560 (4.0%)	298 (5.6%)	262 (3.0%)	0.71***	0.63-0.80	82 (1.7%)	180 (4.9%)	1.76***	1.55-1.99
Other	197 (2.5%)	100 (3.6%)	97 (1.8%)	0.70**	0.56-0.87	37 (1.4%)	60 (2.4%)	1.41*	1.07-1.87
Eligible FSM									
No	6,207 (84.6%)	2,506 (82.7%)	3,701 (86.0%)	1.00		1,837 (83.2%)	1,864 (90.4%)	1.00	
Yes	1,087 (15.4%)	553 (17.3%)	534 (14.0%)	0.90**	0.83-0.97	313 (16.8%)	221 (9.6%)	0.65***	0.55-0.78
Five GCSE A*- C									
No	3,373 (54.0%)	1,515 (57.9%)	1,858 (51.5%)	1.00		1,375 (69.6%)	483 (23.7%)	1.00	
Yes	4,241 (46.0%)	1,649 (42.1%)	2,592 (48.5%)	1.11***	1.06-1.15	866 (30.4%)	1,726 (76.3%)	3.41***	3.06-3.80
Ever been to university									
No	4,085 (63.7%)	1,834 (67.5%)	2,251 (61.2%)	1.00		1,822 (86.6%)	429 (21.3%)	1.00	
Yes	3,622 (36.3%)	1,372 (32.5%)	2,250 (38.8%)	1.11***	1.07-1.16	460 (13.4%)	1,790 (78.7%)	5.81***	5.21-6.49
University degree									
No	5,018 (73.1%)	2,227 (76.8%)	2,791 (70.7%)	1.00		1,914 (89.1%)	877 (41.7%)	1.00	
Yes	2,689 (26.9%)	979 (23.2%)	1,710 (29.3%)	1.13***	1.08-1.18	368 (10.9%)	1,342 (58.3%)	3.36***	3.11-3.64
N	7,707	3,206	4,501			2,282	2,219		

Note: Proportions (%) and Risk Ratios (RR) from modified Poisson regression models are weighted to account for sample design and sample attrition; FSM = Free School Meals; Column 3 reflects the total of columns 6 and 7; Significance level: ****p* < .001, ** *p* < .01, * *p* < .05.

Table 2. Participant characteristics according to consent and data linkage on those who attended university (*n* = 3,622).

	(1) University	(2) No consent	(3) Consent	(4) RR	(5) 95% CI	(6) Not linked	(7) Linked	(8) RR	(9) 95% CI
Sex									
Male	1,526 (46.9%)	567 (45.0%)	959 (48.0%)	1.00		160 (38.0%)	799 (50.6%)	1.00	
Female	2,096 (53.1%)	805 (55.0%)	1,291 (52.0%)	0.96	0.91-1.02	300 (62.0%)	991 (49.4%)	0.90***	0.86-0.94
Parent degree									
No degree	2,593 (71.6%)	1,043 (76.5%)	1,550 (69.1%)	1.00		314 (67.6%)	1,236 (69.4%)	1.00	
Degree	998 (28.4%)	314 (23.5%)	684 (30.9%)	1.13***	1.07-1.20	143 (32.4%)	541 (30.6%)	0.98	0.93-1.04
Ethnicity									
White	2,244 (78.1%)	684 (68.4%)	1,560 (83.3%)	1.00		346 (86.1%)	1,214 (82.6%)	1.00	
Mixed	172 (3.3%)	68 (3.5%)	104 (3.2%)	0.90	0.77-1.06	12 (2.3%)	92 (3.4%)	1.09	0.96-1.23
Indian	363 (4.2%)	188 (6.1%)	175 (3.2%)	0.71***	0.62-0.81	30 (2.3%)	145 (3.4%)	1.08*	1.01-1.16
Pakistani & Bangladeshi	377 (4.1%)	206 (5.9%)	171 (3.1%)	0.71***	0.63-0.82	26 (1.9%)	145 (3.4%)	1.11**	1.04-1.19
Black & Black British	350 (6.3%)	170 (9.4%)	180 (4.6%)	0.69***	0.60-0.79	31 (3.9%)	149 (4.8%)	1.05	0.97-1.14
Other	115 (4.0%)	56 (6.7%)	60 (2.6%)	0.60**	0.45-0.79	15 (3.4%)	45 (2.3%)	0.92	0.75-1.14
Eligible FSM									
No	3,000 (90.6%)	1,117 (88.7%)	1,883 (91.6%)	1.00		375 (93.6%)	1,508 (91.2%)	1.00	
Yes	372 (9.4%)	179 (11.3%)	193 (8.4%)	0.88*	0.77-0.99	29 (6.4%)	164 (8.8%)	1.06	0.98-1.15
Five GCSE A*- C									
No	738 (21.5%)	323 (25.9%)	415 (19.1%)	1.00		84 (16.7%)	331 (19.7%)	1.00	
Yes	2,864 (78.5%)	1,040 (74.1%)	1,824 (80.9%)	1.16***	1.07-1.26	373 (83.3%)	1,451 (80.3%)	0.96	0.91-1.01
University degree									
No first degree	1,295 (36.2%)	551 (41.1%)	744 (33.5%)	1.00		163 (35.5%)	581 (33.0%)	1.00	
First degree	2,327 (63.8%)	821 (58.9%)	1,506 (66.5%)	1.12***	1.06-1.20	297 (64.5%)	1,209 (67.0%)	1.02	0.97-1.08
N	3,622	1,372	2,250			460	1,790		

Note: Proportions (%) and Risk Ratios (RR) from modified Poisson regression models are weighted to account for sample design and sample attrition; FSM = Free School Meals; Column 3 reflects the total of columns 6 and 7; Significance level: ****p* < .001, ** *p* < .01, * *p* < .05.

Data quality comparisons

In this section, we compare similar variables held across both data sources (age 25 survey and SLC data), to explore the extent of individual-level agreement and to infer data quality. We do this for three separate variables: (i) whether the individual attended a Russell Group university, (ii) whether they reported ever taking a student loan and were found to have received an SLC payment, and (iii) their estimated annual income at the time of the survey.

Russell Group university

At the age 25 survey, those who reported being awarded a first degree (*n* = 2,689) were asked to report the awarding institution in a free text box. From this, CLS derived a variable to indicate whether it was a Russell Group university or another institution. However, a large amount of missing data was observed, with only half of those who attended university with complete data (*n* = 1,754), of which, 27% reported graduating from a Russell Group university.

Institution is also available in the SLC applications data for each year participants applied for a student loan (one row per year). We derived a variable indicating whether they ever applied for a student loan to attend a Russell Group (or other) institution between 2007 and 2016 (i.e., before the time of the survey). Data were available for 2,182 participants (a greater number than observed in the survey), of which, 23% were found to have applied for a loan to attend a Russell Group university between 2007 - 2016, which was slightly lower than the estimate observed specifically for the academic year 2008-09 (reported previously).

We then compared information (Table 3) for those with observed data in both sources (*n* = 910). Agreement was very high (97%), with 28% of this selected sub-sample applying to/graduating from a Russell Group university, and 69% applying to/graduating from a non-Russell Group institution. Some minor discrepancy was observed, with 2% applying to attend a Russell Group university (between 2007-2016) but reporting that they graduated from a non-Russell Group institution (in 2016), and a tiny proportion (< 1%) who reported graduating from a Russell Group university but were only found to have applied to attend a non-Russell group institution in the SLC data. These discrepancies could reflect individuals who attended one institution but later graduated from a different one, either due to dropping out of one, or perhaps completing courses at two institutions but only reporting one in the survey.

Table 3. Agreement between SLC and survey data – Russell Group university (*n* = 910).

	SLC data		
Survey data	No	Yes	Total
No	640 (68.8%)	22 (2.2%)	662 (71.0%)
Yes	8 (0.9%)	240 (28.0%)	248 (29.0%)
Total	648 (69.8%)	262 (30.2%)	910 (100.0%)

Note: Proportions (%) are weighted to account for survey design and survey attrition.

Received a student loan

At the age 25 survey, participants were asked whether they had ever taken a tuition fee or a maintenance support loan from the SLC (0 = 'No', 1 = 'Yes'). Among those who reported ever attending university (*n* = 3,622), 81% reported taking a tuition fee loan, and 80% reported taking a maintenance loan, which matches the overall proportion expected to have taken loans, as reported earlier from national data (Bolton, 2022).

For comparison, we created a variable from SLC payments data to indicate whether participants ever received a tuition fee or maintenance loan before the time of the survey. Among those who ever attended university and consented to data linkage (*n* = 2,250), 73% were found to have received a tuition fee loan and 75% were found to have received a maintenance loan, which is slightly lower than national figures, possibly due to differential consent rates.

To assess data quality, we compared information across sources (Table 4). Focusing on the combination of any loan taken, agreement across sources was high (83%), as 76% reported and were found to have received a loan in the SLC data, and 7% reported to have not and were not found in the SLC data (despite consenting to data linkage). Some discrepancy was observed, as a substantial proportion of participants who said they took out a student loan were not found in the SLC payments data (15%), which may reflect missed matches. A smaller proportion who reported never taking a student loan were found to have received a payment in the SLC data (2%), which could reflect either measurement error in the survey data, or false matches in the linked data (although this is not very likely, as the matching process was fairly stringent).

Table 4. Match between SLC and survey data for any loan taken/received – on an eligible sample of those who went to university and consented to data linkage (*n* = 2,250).

Any student loan taken			
	SLC data		
Survey data	No (0)	Yes (1)	Total
No (0)	165 (7.3%)	41 (1.8%)	206 (9.1%)
Yes (1)	332 (15.3%)	1,712 (75.6%)	2,044 (90.9%)
Total	497 (22.6%)	1,753 (77.4%)	2,250 (100.0%)

Note: Proportions (%) are weighted to account for survey design and survey attrition.

Estimated income

Using SLC repayments data, we derived a variable reflecting estimated gross annual income. Students are expected to repay 9% of their gross income above a certain threshold (Appendix A). Therefore, to estimate income at age 25, the total repayment amount in 2015-16 (excluding voluntary payments, refunds, and overpayments)[15] was divided by 0.09 and the repayment threshold (e.g., £17,335 in 2015-16) was added back to the estimate. Income could therefore only be estimated for those who earned above the threshold and made a

[15] Repayments from 2016-17 were used for participants who completed the survey after April 2016, because the survey crossed over two tax years.

repayment in that year (*n* = 1,415). Average (mean) estimated gross annual income in the SLC data was £26,561 (SD = £9,415) and the median was £24,302.

Income is widely regarded as difficult to measure in surveys, as it is prone to item non-response and measurement error (Angel, Disslbacher, Humer, & Schnetzer, 2019). At the age 25 survey, participants were asked to report their gross income (continuous) from their main job, by either week, month, or year,[16] and advised to consult their payslip to help them answer correctly. Of those who were employed[17] at the time of the survey (*n* = 5,739), missing income data was observed at a rate of 9%. The derived and deposited version of this variable is provided as a weekly estimate, which we multiplied by 52 to estimate gross annual income. Among a comparable sample of those who ever went to university and consented to data linkage (*n* = 1,820),[18] average estimated gross income (from main job) was £24,358 (SD = £11,156) and the median was £23,724.

To assess data quality, we compared gross annual income in the SLC data with gross annual income (from main job) in the survey data. The sample reflects only those with complete data in both (*n* = 1,285). Individuals in this sample are constrained (by virtue of data availability) to have income above the income threshold according to the SLC data, but we chose not to impose a similar constraint on income reported in the survey data, meaning that it was possible for individuals to report income below the threshold in the survey and still appear in the sample.

Despite this asymmetry, however, estimated income was only slightly higher on average in the SLC compared to the survey data (Table 5). Figure 4 shows the income distributions obtained from the two sources overlayed in a kernel density plot, highlighting the similarity of the two distributions from about £30,000 per year onwards. The dissimilarity at lower income levels is likely driven by the different sample inclusion criteria described earlier. Regarding the within-individual level agreement across sources, a pairwise correlation showed that the two estimates were significantly positively correlated ($r = .827$, $p < .001$) – see Figure 5 for a scatterplot.

Table 5. Estimated gross annual income (£) at age 25 – survey and SLC data (*n* = 1,285)

Variable	Mean	SD	25%	50%	75%
Survey estimate	26,226	10,102	19,852	24,960	30,155
SLC estimate	26,810	9,290	20,217	24,624	30,746

[16] While in principle loan repayments are due on all income above the threshold, in practice, as we saw above, the vast majority (90%) of obligatory repayments were made via PAYE (i.e., levied on employment income, also referred to as earnings) and the regulations further state that the threshold is applied separately for each job, meaning that an individual with a second job in which they earnt below the threshold would not be liable for repayments on their income from this second job. Therefore, in practice, only those reporting their income via self-assessment are likely to be making repayments on the basis of total income rather than earnings. Therefore, it makes sense to compare the income measure estimated from the SLC data with the measure of income from their main job in the survey.

[17] Those who were self-employed were asked to report their take home pay after taxes and costs (*n* = 473) which was not analysed, as it did not correspond to gross pay, which is used to calculate SLC repayments.

[18] Extreme values for self-reported income, greater than £100,000 per annum (*n* = 12), were excluded to reduce the large positive skew of data.

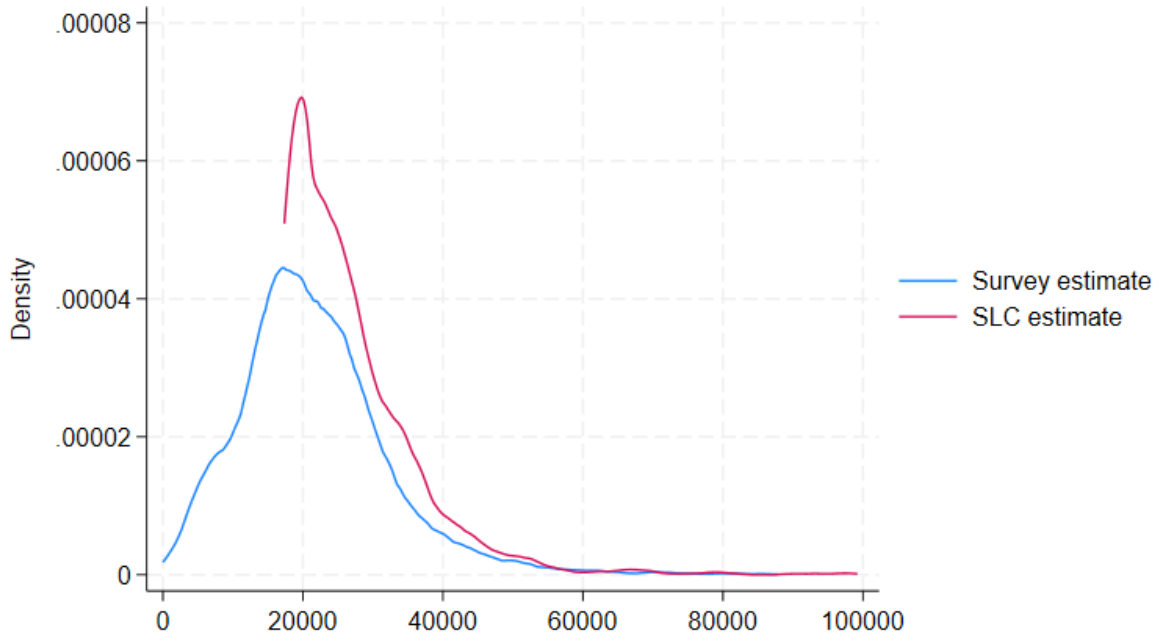

Figure 4. Overlayed kernel density plot for income estimates across sources (*n* = 1,285).

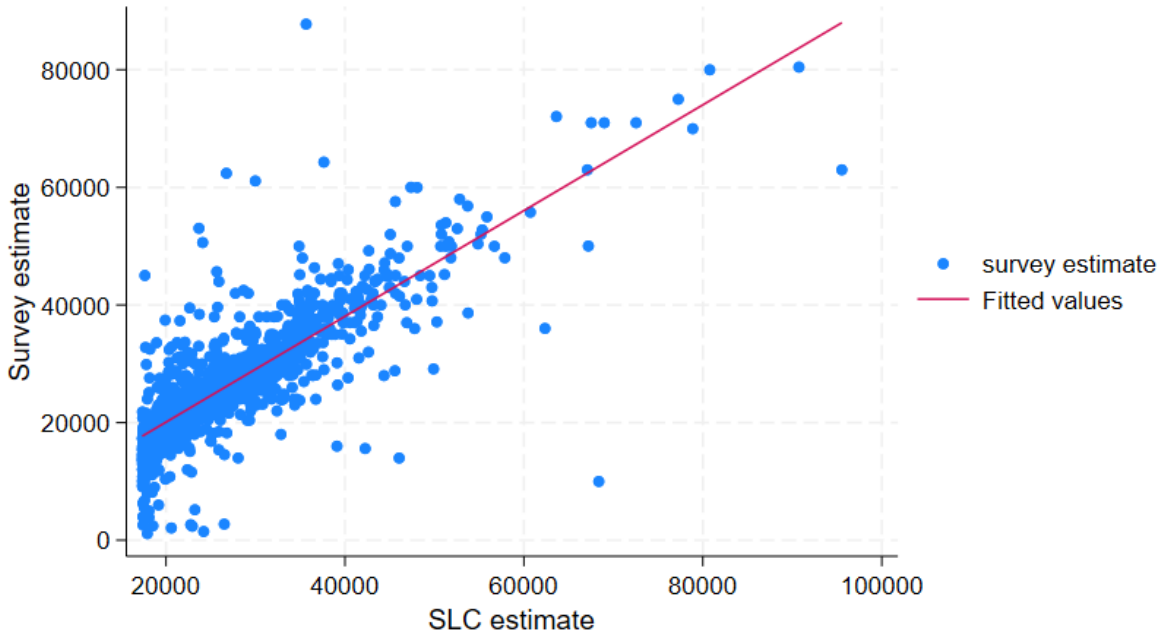

Figure 5. Scatterplot showing within-individual level agreement for income estimates across sources (*n* = 1,285).

Using the data for policy-relevant research

To illustrate some of the policy-relevant research questions that the linked Next Steps-SLC data could help address, we used the data to analyse which types of graduates[19] are likely to be affected by recent changes to the repayment terms for student loans in England (for an overview of the changes see Waltmann, 2022).

Policy context

As outlined above, students in England can apply for income-contingent loans to cover the full amount of their tuition fees, and a proportion of their living costs. Since the Next Steps cohort went to university, the cap on tuition fees in England has risen to £9,250 per year and means-tested maintenance grants have been abolished, replaced by higher maintenance loans, meaning that students now graduate from university with significantly higher debt than the Next Steps cohort.

The terms under which student loans are issued and must be repaid have also changed over time. While graduates must still repay 9% of their income above a threshold, that threshold has changed over time. The period over which graduates are liable for repayments has also been extended, and the way interest is calculated has also changed.

Students entering HE in 2022-23 would have expected to face a positive real interest rate of RPI +3% on their debt, and to make repayments of 9% of their income above a threshold of £29,860 (rising in line with average earnings growth) for a period of up to 30 years, after which any remaining outstanding debt would be written off. This set of loan conditions is referred to as Plan 2.[20]

The UK government recently announced changes to the student loans system, however, meaning that students entering HE in 2023-24 and onwards will now be subject to tougher loan repayment terms (Waltmann, 2022). While loans are now only subject to a 0% real interest rate (based on RPI), the repayment term has been extended to 40 years, and the income threshold above which repayments must be made has been reduced to £25,000 and frozen for the next three years. This reformed system is known as Plan 5.

Contribution

Previous analysis has focused on estimating the implications of these changes for the total cost to government of the student loans system, and for individuals with different lifetime earnings (Waltmann, 2022). This revealed that the reforms are “regressive”, as they are likely to disadvantage lower earning graduates more than higher earning graduates (Waltmann, 2022). More specifically, a higher proportion of lower earning graduates will be drawn in to making repayments in any given year, with these individuals also more likely to end up making repayments across the full repayment period. Meanwhile, higher earning

[19] We refer to individuals making student loan repayments as graduates, although not all those making repayments will have completed a university degree.
[20] Some reforms have been introduced for *Plan 2* borrowers, although we will not discuss those here.

graduates are likely to repay their student loans in full more quickly, resulting in an overall reduction in the amount they will repay, due in part to lower interest repayments.

Here, we use the Next Steps-SLC linked data to add to our understanding of the likely implications of this reform, focusing on the types of individuals who, in their 20s, are likely to be drawn in to making repayments under the new Plan 5 system, who would not have made any repayments under the previous system. We use the SLC data to identify individuals liable for repayments under the different student loan systems and use the rich information available from Next Steps to explore the background characteristics of these individuals, their education experiences, and the types of jobs they work in at age 25.

Approach

To undertake this analysis, we need to observe precise information about annual incomes, to identify individuals with income above the old (Plan 2) threshold, who would make repayments under either system, and those with income below the old threshold but above the new (Plan 5) threshold, who would be drawn in to making repayments as a result of the reforms. We are able to undertake this analysis using Next Steps-SLC data because the new Plan 5 repayment threshold is actually very similar (in real terms) to the repayment threshold faced by the Next Steps cohort (who were on what was described as Plan 1 repayment terms).

We restricted attention to those making a loan repayment in 2015-16, to ensure we captured the characteristics of individuals contemporaneously with survey data. Our sample of interest is therefore 1,385 individuals who took part in the age 25 survey, were linked to SLC data, and made a student loan repayment in 2015-16.[21] We split our sample into two groups: those with income above £20,665 (equivalent to £29,860, the Plan 2 threshold, in 2023 prices) (*n* = 929), who would make repayments under either system, and those with income between £25,000-£29,859 (in 2023 prices)[22] (*n* = 457), who would be drawn in to making repayments under Plan 5 reforms. This reflects an increase of 49% for graduates making repayments.

This approach implicitly assumes that the types of graduates who had incomes within the relevant range among the Next Steps cohort are similar to those who will enter HE this year. This is a difficult assumption to substantiate, but in terms of gender and ethnicity, comparing HESA statistics from 2008-09 to the most recent data (2021-22), the gender composition of undergraduate students attending UK HE institutions has not changed (i.e., 55% vs. 56%), but the proportion of ethnic minority students has increased (i.e., from 21% to 30%).[23] One further similarity is that the cohorts will both have graduated into uncertain labour markets following large macroeconomic shocks (i.e., the 2008 financial crisis in the case of Next Steps and the COVID-19 pandemic for today's graduates).

[21] This group were comparable to the full linked sample – i.e., 48% female and 17% ethnic minority.
[22] Prices were inflated from April 2015 to April 2023 using the Retail Price Index (RPI).
[23] Source: https://www.hesa.ac.uk/data-and-analysis/students/whos-in-he

Results

Increase in individual repayments

We first calculated how much more an average graduate would be likely to repay due to the lower repayment threshold. The average (median) income for the whole sample of Next Steps graduates who made repayments in 2015-16 was £33,991 (in 2023 prices). Therefore, an average graduate would be repaying £372 under the previous system (9% of the difference between £29,860 and £33,991) and £809 under the new system (9% of the difference between £25,000 and £33,991), reflecting an increase of £437 (+54%) per year, equivalent to around 3% of net annual income.[24]

Who is likely to be drawn in to making repayments?

Background characteristics

We estimate that those drawn in to making repayments would be more likely to be first in their family to attend HE (74% vs. 68%),[25] be from an ethnic minority background (22% vs. 15%) and have been eligible for free school meals (9% vs. 5%). In addition, they would be less likely to have attended private school (7% vs. 14%).

Table 1. Comparison of background characteristics between repayment groups.

	Pre-reform group (n = 929)	Drawn in group (n = 457)
Female sex	46.1%	50.3% *RR = 1.18*
First in family	67.6%	74.0% *RR = 1.36**
Ethnic minority	14.5%	21.7% *RR = 1.63****
Free school meals at age 16	4.9%	9.1% *RR = 1.96***
Private school ages 14-18	14.3%	6.7% *RR = 0.43***

Note: Survey design and attrition weights applied; RR = risk ratio; Significant at ***p < .001, **p < .01, *p <.05.

[24] Net income (£27,137) calculated by deducting income tax (20%) and national insurance contributions (12%).

[25] First in family was operationalised as not having a parent with a university degree.

Education experiences

We estimate that those drawn in to making repayments would be less likely to have attained five or more GCSE grades A*- C (76% vs. 90%), attended a Russell group university (14% vs. 35%), or studied a STEM (science, technology, engineering, or maths) subject (18% vs. 28%). They would be more likely to have studied other/combined subjects (16% vs. 9%).

Table 2. Comparison of education experiences between repayment groups.

	Pre-reform group (*n* = 929)	Drawn in group (*n* = 457)
Five GCSE (A*-C) age 16	90.2%	76.3% *RR = 0.35****
Russell group university	35.1%	13.9% *RR = 0.30****
OSSAH subject	42.5%	48.9% *RR = 1.19*
LEM subject	20.5%	17.3% *RR = 0.81*
STEM subject	27.7%	17.9% *RR = 0.57****
Other/combined subject	9.3%	16.0% *RR = 1.84***

Note: OSSAH = Other social science and humanities, LEM = law, economics and management, STEM = science, technology, engineering, and maths; Survey design and attrition weights applied; RR = risk ratio; Significant at ****p* < .001, ***p* < .01, **p* < .05.

Employment outcomes age 25

We estimate that those drawn in to making repayments would be more likely to work part-time (10% vs. 1%), be currently in education while also working (6% vs. 3%), be employed on a zero-hours (8% vs. 1%) or non-permanent contract (17% vs. 8%), and work in a semi-routine or routine occupation (19% vs. 3%). Conversely, they would be less likely to work in London (17% vs. 27%), work full-time (85% vs. 98%), work in the public sector (27% vs. 34%), or work in a higher, managerial, admin, or professional occupation (42% vs. 81%).

Table 3. Comparison of employment outcomes age 25 between repayment groups.

	Pre-reform group (*n* = 929)	Drawn in group (*n* = 457)
London	26.8%	16.7% *RR = 0.54***
Full-time job	97.6%	85.2% *RR = 0.14***
Part-time job	1.0%	10.1% *RR = 10.61***
Private sector job	58.0%	60.8% *RR = 1.11*
Public sector job	34.4%	26.8% *RR = 0.70***
Currently in education	3.2%	6.0% *RR = 1.90**
Zero hours contract	1.3%	7.6% *RR = 6.14****
Non-permanent contract	7.6%	16.7% *RR = 2.44****
Semi-routine & routine occupations	2.6%	19.1% *RR = 8.94****
Higher managerial, admin & professional	80.8%	42.2% *RR = 0.17****

Note: Survey design and attrition weights applied; RR = risk ratio; Significant at ****p* < .001, ***p* < .01, **p* < .05.

Summary of policy-relevant research

Our analyses suggest that lowering the income repayment threshold to £25,000 will result in a large proportion of graduates being drawn in to making repayments (double as many at age 25), who would not have been expected to make any repayments under the previous system. In general, those making repayments will also be expected to repay substantially more than before (additional 54% for an average earning graduate).

The recent student loan reforms were designed to benefit taxpayers, by increasing the proportion of the student loan bill that is eventually repaid, which is likely to be achieved by extracting higher repayments from lower and middle earning graduates (Waltmann, 2022). Our analyses further add to our understanding of this issue by showing which types of graduates will be drawn in to making repayments in their mid-20s as a result of these reforms.

Graduates drawn in to making repayments will be more likely to be from disadvantaged and ethnic minority backgrounds. They will have lower educational attainment and will be less likely to attend Russell Group universities. They will be more likely to work part-time and be currently in education while also working, and also more likely to have precarious working conditions, including being on zero-hours and non-permanent contracts and working in routine or semi-routine occupations. These results support previous findings that the recent student loan reforms are regressive (Waltmann, 2022), as they will target graduates from more disadvantaged backgrounds and those working in precarious conditions.

General Discussion

This paper has sought to highlight the opportunities and potential challenges presented by the recent linkage between the Next Steps survey and Student Loans Company administrative data.

Our investigation of sample representativeness revealed that among age 25 survey respondents, those from an ethnic minority or lower socioeconomic background (and females) were less likely to consent to data linkage. This led to an under-representation of these individuals in the linked Next Steps-SLC sample compared to those who reported ever having gone to university by age 25. In terms of other characteristics, however, the composition of the two samples is broadly similar, suggesting that the linked data are, on the whole, largely representative of this potential group of interest.

Due to a lack of available data on the population of student loan borrowers, it was not possible to discuss the representativeness of the linked sample to this specific underlying population of interest. However, certain characteristics of the linked sample, where it was possible to make comparisons, were found to be similar to the population of UK HE students, such as the proportion of students attending Russell Group universities. However, as above, a slightly lower proportion of female and ethnic minority students were represented in the linked sample, likely due to the lower consent rates. Yet, it should be noted that the survey design and attrition weights used for this analysis were designed for use on the full sample of age 25 respondents and may not be as effective when applied to the linked sample. Future research may consider constructing weights specifically for this sample, using information collected from cohort members at earlier sweeps (Calderwood et al., 2021).

Our data quality comparisons revealed high agreement across sources. The Russell Group variable showed the highest degree of overlap, with 97% agreement. This suggests that using the SLC data to ‘fill in’ missing information on HE institution from the Next Steps data might be possible, albeit only for those who took out student loans and were successfully linked to SLC data.

Agreement across sources was slightly lower for the variable indicating whether participants took out a student loan (83%). The largest discrepancy was observed for those who self-reported taking a loan but were not found in the SLC data (15%), which may have reflected missed matches. This could be related to the fact that not all participants provided a National Insurance Number (NINO), a unique identifier used to facilitate the linkage, where available. Future linked administrative data projects could therefore usefully aim to improve the linkage rate amongst consenting participants, either by attempting to better capture identifiers for matching, or applying ‘fuzzy matching’ techniques, which allow small differences in the matching of an individual identifier (e.g., in the case of misspelling), although this could lead to higher rates of false matches.

The income data also showed a high level of agreement, with similar distributions across the two sources, and a strong positive correlation. However, due to the way that data was collected and operationalised, estimated income from the SLC was only observed if greater than the repayment threshold (i.e., £17,335), which meant that the distributions were less aligned at lower income levels. Discrepancies could have been caused by the different reporting periods, as the survey estimate reflected income at one point in time, whereas the SLC estimate encompassed income across one full financial year. Therefore, those with variable income across the year would be likely to show discrepancies.

A benefit of linking survey and administrative data is the possibility to augment sensitive information or that which may be difficult to self-report accurately. However, using SLC data, income could only be estimated for those who took out student loans and earned above a certain threshold, which excluded those who did not attend university or take out student loans. Therefore, other sources of linked data, such as that from HM Revenue and Customs could provide more complete income data for the wider sample.

In summary, we found that sample representativeness and data quality in the linked Next Steps-SLC data were, on the whole, good, and provided useful additional information about Next Steps respondents, opening up a range of new research questions. It is worth noting, however, that ethnic minority groups and those from more disadvantaged backgrounds may be under-represented in the linked sample as a result of lower consent rates. We conclude, therefore, that future research should capitalise on this new linked data resource for investigating a range of outcomes among student loan borrowers, while acknowledging the potential issues around representation.

References

Angel, S., Disslbacher, F., Humer, S., & Schnetzer, M. (2019). What did you Really Earn Last Year?: Explaining Measurement Error in Survey Income Data. *Journal of the Royal Statistical Society Series A: Statistics in Society, 182*(4), 1411-1437. doi:10.1111/rssa.12463

Bolton, P. (2019). *Student Loan Statistics: Briefing Paper* [Commons Library Research Briefing 1079]. Retrieved from https://dera.ioe.ac.uk/id/eprint/34780/1/SN01079%20(1).pdf

Bolton, P. (2022). *Student Loan Statistics: Research Briefing* [Commons Library Research Briefing CBP01079]. Retrieved from https://researchbriefings.files.parliament.uk/documents/SN01079/SN01079.pdf

Calderwood, L., & Lessof, C. (2009). Enhancing longitudinal surveys by linking to administrative data. In P. Lynn (Ed.), *Methodology of longitudinal surveys* (pp. 55-72): John Wiley & Sons Ltd.

Calderwood, L., Peycheva, D., Henderson, M., Silverwood, R. J., Mostafa, T., & Rihal, S. (2021). *Next Steps: Sweep 8 - Age 25 User Guide (3rd edition)*. London: UCL Centre for Longitudinal Studies.

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

Harron, K., Doidge, J. C., & Goldstein, H. (2020). Assessing data linkage quality in cohort studies. *Annals of Human Biology, 47*(2), 218-226. Retrieved from https://doi.org/10.1080/03014460.2020.1742379

Peycheva, D., Ploubidis, G. B., & Calderwood, L. (2021). Determinants of consent to administrative records linkage in longitudinal surveys: evidence from Next Steps. *Advances in Longitudinal Survey Methodology*, 151-180. Retrieved from https://doi.org/10.1002/9781119376965.ch7

Rihal, S., Gomes, D., & Henderson, M. (2021). *Next Steps: Linked Student Loans Company administrative datasets User Guide (2nd edition)*. London: UCL Centre for Longitudinal Studies.

Silverwood, R. J., Rajah, N., Calderwood, L., De Stavola, B. L., Harron, K., & Ploubidis, G. B. (2024). Examining the quality and population representativeness of linked survey and administrative data: guidance and illustration using linked 1958 National Child Development Study and Hospital Episode Statistics data. *International Journal of Population Data Science, 9*(1).

Waltmann, B. (2022). Student loans reform is a leap into the unknown. Retrieved from https://ifs.org.uk/sites/default/files/output_url_files/BN341-Student-loans-reform-is-a-leap-into-the-unknown.pdf

University College London, Institute of Education, Centre for Longitudinal Studies (2022). Next Steps: Linked Administrative Datasets (Student Loans Company Records), 2007-2021: Secure Access [data collection]. UK Data Service. SN:8848, DOI: http://doi.org/10.5255/UKDA-SN-88-48-1

APPENDIX A

Table 1 - Student Loans Company repayment thresholds by year for Plan 1 borrowers.

Tax Year	**Threshold**
Pre- 2012	£15,000
2012-13	£15,795
2013-14	£16,365
2014-15	£16,910
2015-16	£17,335
2016-17	£17,495
2017-18	£17,775
2018-19	£18,330
2019-20	£18,935
2020-21	£19,390

Note: Students repay 9% of their income above the threshold.