National Audit Office

Report
by the Comptroller
and Auditor General

Department for Business, Innovation & Skills

Student loan repayments

Technical paper

DECEMBER 2013

Our vision is to help the nation spend wisely.

Our public audit perspective helps Parliament hold government to account and improve public services.

The National Audit Office scrutinises public spending for Parliament and is independent of government. The Comptroller and Auditor General (C&AG), Amyas Morse, is an Officer of the House of Commons and leads the NAO, which employs some 860 staff. The C&AG certifies the accounts of all government departments and many other public sector bodies. He has statutory authority to examine and report to Parliament on whether departments and the bodies they fund have used their resources efficiently, effectively, and with economy. Our studies evaluate the value for money of public spending, nationally and locally. Our recommendations and reports on good practice help government improve public services, and our work led to audited savings of almost £1.2 billion in 2012.

Contents

The National Audit Office study team consisted of:

Martin Malinowski and Diana Tlupova, under the direction of Peter Gray.

This report can be found on the National Audit Office website at www.nao.org.uk

For further information about the National Audit Office please contact:

National Audit Office
Press Office
157-197 Buckingham Palace Road
Victoria
London
SW1W 9SP

Tel: 020 7798 7400

Email: enquiries@nao.gsi.gov.uk

Introduction

1 This technical paper accompanies the publication of the National Audit Office's value for money report, *Student Loan Repayments*, published in November 2013. The report assesses whether the approach for collecting student loans, as adopted by the Department for Business, Innovation & Skills (BIS), Student Loans Company (SLC) and HM Revenue & Customs (HMRC), is value for money and whether these organisations are ready for new challenges.

2 Forecasting of future loan repayments is an essential tool for BIS' financial planning. The scale and projected growth of student loans following the Government's 2010 changes to higher education funding mean that accurately forecasting repayments into the future will be important to manage its exposure to repayment risk. A necessary prerequisite for robust forecasting is the use of high-quality modelling assumptions which make optimal use of available information on trends and associations in borrower earnings.

3 The main aim of this paper is to understand how good BIS's forecasting of future loan repayments is for the purposes of decision-making and what more could be done to improve the forecasts.

4 This paper includes:

- Our review of the current BIS model for forecasting student loan repayments (Part One).
- A discussion of the data and methodology we used to estimate the propensity of borrowers to repay based on subject studied and higher education institution attended and the main results (Part Two).

5 Our analysis was reviewed internally and submitted to external experts who provided comments. We have reflected these comments in this iteration of the paper.

Part One

HERO model review

1.1 This Part examines BIS's approach to forecasting student loan repayments. It covers the following:

- BIS's model for forecasting repayments;
- analysis comparing forecast to actual repayments; and
- our assessment of BIS' forecasting, and how BIS is changing its approach.

1.2 BIS uses modelling to estimate the total cost of providing loans over a 25-30 year period. Initially, BIS used the Student Loan Repayment Model (SLRM), developed in the early 1990s. However, in 2010 BIS concluded that this model was not fit for the purpose of providing valuations which could aid in selling the loan book and commissioned an alternative from their consultants, Deloitte. The result of this work was the HERO model, which was implemented in June 2011.

Structure of the HERO model

HERO model inputs

1.3 The HERO model is an Excel-based micro-simulation model which aims to forecast incomes and the associated repayments of both current and future student loan borrowers. It holds data on demographic and behavioural characteristics of students in order to predict their borrowing behaviour and estimate their repayment of student loans. The model relates only to income-contingent repayment (ICR) loans for English domiciled students studying in the UK, and EU-domiciled students studying in English Higher Education Institutions. By simulating the behaviour of individual borrowers, the model attempts to forecast repayment cash flows over 30 years into the future for over 3.8 million borrowers. A simplified map of the model is presented in Figure 1.

1.4 The key part of the model is the module which forecasts borrower earnings. This module uses Student Loan Company data on existing borrowers to generate a representative selection of borrower profiles and loan amounts. Different assumptions are then used to model earnings paths for individual borrowers and how this translates into loan repayments.

Macroeconomic assumptions

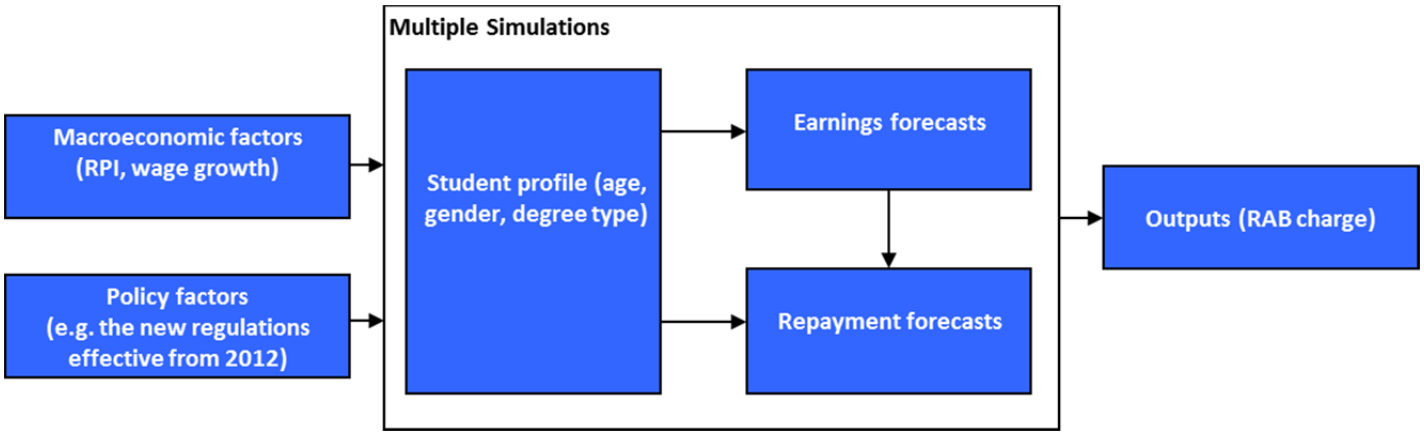
1.5 Macroeconomic assumptions are the external factors that affect the model outputs. External factors, such as economic growth and inflation, are important in estimating future repayments. The latest version of HERO model (March 2013) relies on a set of macroeconomic forecasts to predict borrower incomes in future years. These include:

- Nominal Earnings Growth (NEG) - used to inflate all wages from their 2009 values to their future-year equivalents
- Retail Price Index (RPI) - used in the calculation of student loan interest as per the policy rules applying to the respective cohort
- The Bank of England base rate (BR) - used in the calculation of student loan interest, as per the policy rules applying to the respective cohort

Medium-term forecasts are taken from the Office of Budget Responsibility's (OBR) quarterly forecasts, while long-term data are drawn from the OBR assumptions about long-term productivity growth and the inflation target.

Figure 1

A simplified description of the HERO model



Notes

1. The model performs simulations of income and loan repayments profiles for a sample of borrowers. The results of these sample calculations are then aggregated and scaled up for the whole portfolio of income-contingent repayment loans.
2. The RAB charge is the portion of loans that BIS does not expect to be repaid.

Source: NAO analysis of HERO model and supporting information

Future earnings forecasts

1.6 The model uses a combination of data from: the Labour Force Survey (LFS), British Household Panel Survey (BHPS) and SLC data on existing borrowers to forecast earnings profiles of borrowers it depicts.

1.7 The methodology for forecasting future incomes of borrowers consists of several steps. First, BHPS data covering the period 1991 to 2008 is used to establish an income percentile transition matrix which assigns probabilities to transitions between income percentiles based on their observed frequency in the survey evidence. This matrix is then used together with a random number generator to generate an earnings profile for each depicted individual.

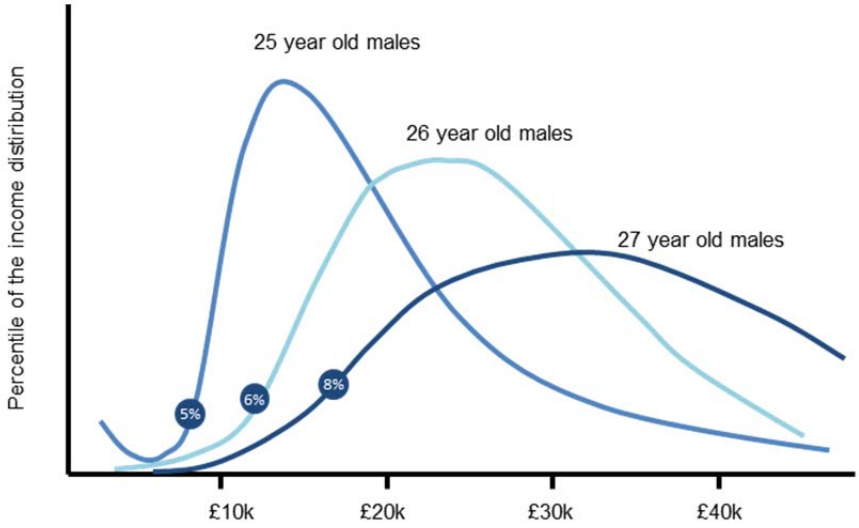
1.8 Inputs relating to borrower characteristics (Gender, Age, Loan Size, Type of Degree, Earnings), are drawn from SLC data on the actual borrower population. This data is also used to simulate the composition of future year cohorts for which there is not yet data available.

1.9 The income percentile transitions which jointly constitute an earnings profile are generated using a first-order Markov process, or in other words a random process where a value in the next period depends only on the value in the current period and not on the values in preceding periods. The model assumes that the income of a student loan borrower in the next year is determined based on his/her income in the current year, their age, gender and level of educational attainment implied by their degree.

1.10 Percentile information on borrowers is transformed into a salary figure by mapping the percentile information from the transition matrices onto actual graduate earnings distributions derived using the LFS data for 2001- 2009. The earnings information converts income percentile information into actual earnings information based on the borrower's characteristics (age, gender and educational attainment). Figure 2 shows an example of a borrower's income path.

Figure 2

Example of borrower's income trajectory (male, 3 working years from age 25)



Notes

1. Distributions are for illustrative purposes only.

Source: NAO analysis of HERO model and supporting information

1.11 For a given forecast year, the income corresponding to the forecast income percentile is adjusted to reflect the compounded impact of wage growth up to that date. The adjustment is performed by applying outturn nominal wage growth as published by the ONS, and nominal wage growth as forecasted by the OBR in its medium-term Budget 2013 forecasts. Over the longer-term (from 2023/24 onwards), nominal wages are assumed to grow at the rate of 4.4 per cent per annum, as per OBR forecasts.

HERO model outputs

1.12 Using the input assumptions, HERO model generates several outputs, such as number of borrowers, loan balances, earnings and repayment analysis and proportion of the initial loan value that will never be repaid. The latter, known as the Resource Accounting and Budgeting (RAB) charge.

1.13 The RAB charge represents the cost to the government in a given year of issuing loans to future borrowers. The RAB charge is calculated as the face value of loans made in any one year less the discounted or present value of future repayments.

$$RAB\ Charge = \frac{Face\ Value\ of\ Loans - Discounted\ Value\ of\ Repayments}{Face\ Value\ of\ Loans} * 100\%$$

1.14 The main factors in determining the RAB charge are the size of the loans issued, the interest rate charged on the loans and the earnings of borrowers in the future, which together with the repayment terms will determine the rate of repayment. Non-repayment of loans may occur due to the interest-rate subsidy, low earnings, or debt write-off in the case of permanent disability or death.

Changes to the modelling approach

1.15 In 2012 BIS concluded that there was potential to improve the model's approach to simulating borrower earnings paths by linking forecast earnings to several years of past earnings, rather than just the previous year's earnings (as is the case in the HERO model). BIS aims to have developed an improved forecasting model, called the Stochastic Earnings Pathways (STEP) model, by spring 2014. STEP will be a stochastic wage model, and no longer uses the transition matrix approach described above in paragraph 1.7. The upgrade is also slated to include improved assumptions around unemployment and the treatment of male and female borrowers.

Accuracy of the HERO model

Forecasting the repayments

1.16 Forecasting of future repayments is an essential tool for BIS' financial planning. It is important that BIS understands and accurately projects repayments into the future, given the scale and growth of student loans. Forecasting has limitations, however, as projections are based on complex and often uncertain assumptions

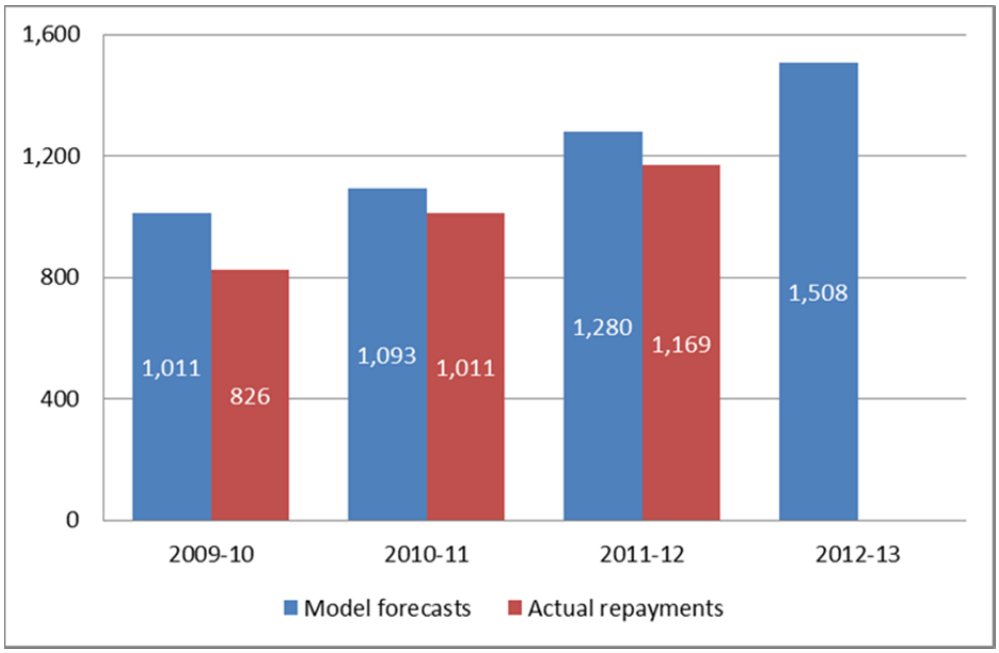
1.17 As student loans have become more complex, BIS has found it increasingly difficult to forecast repayments. In particular, the model has over-forecasted repayments for older cohorts, as it was not fully accounting for the fact that as some borrowers repay in full, the repayment characteristics of the remaining borrowers become steadily less similar to the borrower population represented in the model.

1.18 Complex simulation models based on uncertain assumptions, such as HERO, should be tested for accuracy against actual data. BIS does not regularly compare actual repayments to those forecasted by the model and does not analyse the reasons for any variance. In 2011 BIS analysed the difference between forecast repayments and actual amounts collected, and found that by 2009 the gap had grown to 17 per cent, approximately £150 million in 2008-09.

1.19 Forecasting has improved since the HERO model was introduced, but BIS still consistently over-forecasts the repayments. We conducted our own comparison analysis of the difference between the model projections and actual repayments from 2009-10 onwards. Forecasts were approximately 7 to 9 per cent higher than actual repayments in 2010-11 and 2011-12 (Figure 3).

Figure 3

Comparison of model forecasts to actual repayment data (£m)



Notes

1. The data exclude voluntary early repayments, which are more volatile and difficult to forecast, and only include repayments based on earnings.
2. Actual repayments data is taken from SLC statistical first releases, 26 June 2012 and 25 June 2013. Actual data for 2012-13 is not yet available, as HMRC confirmation takes place later due to processes in the tax system.
3. Forecasts from data provided by BIS. Forecasts were made using the HERO model developed by Deloitte in 2011, excluding 2009-10, which was produced using the Student Loan Repayment Model.

Source: National Audit Office analysis of BIS and SLC data

1.20 Other organisations have also either examined the accuracy of BIS forecasts or estimated their own RAB charge. The Institute for Fiscal Studies concluded that the HERO model overestimates annual earnings at the top of the income distribution compared to their profiles of lifetime earnings, leading to a lower RAB charge.[1] The Institute for Public Policy Research has estimated a RAB charge of 39 per cent which is higher than BIS's projections of 35 per cent.[2] The Higher Education Policy Institute has estimated that if the RAB charge rises above 47 per cent, then the current Higher Education funding policy would be more expensive than the one it has replaced.[3]

Robustness of assumptions

Macroeconomic forecasts

1.21 To test the reasonableness of the macroeconomic forecasts used as inputs to the HERO model, we examined the effect of using forecasts provided by other organisations other than the OBR. Re-performing modelling using macroeconomic assumptions from the Bank of England (BoE) and the International Monetary Fund (IMF) resulted in a slightly lower RAB charge and higher discounted value of loan repayments than BIS's own estimate **(**Figure 4). Although this demonstrates that BIS is using the most conservative set of macroeconomic assumptions available, all forecasting organisations have had to significantly revise their growth forecasts on the basis of the UK's persistently low growth in recent years.

1.22 This may yet prove to be the case with the current set of OBR forecasts, and Figure 4 provides four alternative modelled scenarios to the OBR forecast - one in which real income growth occurs at 50 per cent of the rate assumed by the OBR from 2013/14, and a 'worst-case' scenario in which real income growth is set to zero from 2013/14 onwards. Model forecasts using medium-term macroeconomic assumptions provided by the BoE and IMF are also provided.

[1] L Dearden, H Chowdry, , and G Wyness'Government proposals for higher education would squeeze high earners less and cost the taxpayer more *', Institute for Fiscal Studies,* November 2010, available atwww.ifs.org.uk/publications/5354, accessed 11 November 2013.

[2] Institute for Education Policy Research, *'A critical path. Securing the future of higher education in England'*, June 2013.

[3] J Thompson and B Bekhradnia, *'The government's proposals for higher education funding and student finance – an analysis'*, November 2010.

Figure 4

Effect of different macroeconomic assumptions on the RAB charge

Source	RAB charge						(2012/13 present value)
	2005/06	2006/07	2007/08	2008/09	2009/10	2010/11	
Office for Budget Responsibility (central case)	27.9%	31.2%	33.1%	34.7%	35.3%	36.4%	£28.1
Office for Budget Responsibility (50% real income growth)	28.3%	31.7%	33.7%	35.4%	36.1%	37.2%	£27.8
Office for Budget Responsibility (0% real income growth)	29.8%	33.5%	35.7%	37.7%	38.5%	39.9%	£26.8
Bank of England	28.1%	31.2%	33.0%	34.5%	35.0%	35.9%	£28.2
International Monetary Fund	25.9%	28.9%	30.6%	32.1%	32.6%	33.4%	£29.2

Notes

1. Simulation of 100,000 student earnings profiles carried out using the 2013 HERO model where existing medium-term assumptions have been substituted for those available for another forecasting organisation in each case.
2. The estimated RAB charge represents the cost to government of issuing student loans in a given year in terms of the likely impairment as a proportion of the loan's face value.
3. The loan book value is the present value of future repayments.

Sources: Office for Budget Responsibility, March 2013 Economy Supplementary Tables; IMF World Economic Outlook, 2013; Bank of England Inflation Report, May 2013; NAO analysis using the HERO model

Assumptions about borrower earnings

1.23 Assumptions used in the HERO model to forecast borrower earnings and earnings growth may lead to overestimates of their true levels. The assumptions about earnings dynamics are based on wage data from former students which comes from the period 1991 to 2008. The model assumes that the historically inferred probabilities of moving from one income percentile to the next will apply to future periods. However, available evidence suggests that the salaries enjoyed by newer graduates[4] may have stagnated, and that growth in the salaries of graduates is increasingly being shared unevenly:

- Analysis of the LFS shows that average pay for graduates observes a decreasing trend – for instance for non-manual professions in 1993, only 6.3 per cent of the 'lower than average pay' bracket was made up by graduates.[5] In 2008, this percentage had more than doubled to 15.6 per cent. Brynin argues that this quantitative analysis is proof that an increasing percentage of graduates enter jobs which are not clearly 'graduate' according to the old definitions of having high pay and high upwards earnings mobility.
- Evidence on graduate earnings growth indicates that earnings growth is higher for higher earners, and varies depending on subject studied or university attended. Green and Zhu have found, using the Labour Force Survey for 1997-2004, that there is increasing dispersion in the returns to graduate education in Britain.[6]
- Other researchers have analysed the effect on earnings of students who graduate in a recession. This has particular relevance to borrowers entering the English labour market after the 2008 recession was underway. For instance, one study of Canadian men graduating 1976 to 1995 found that an average-sized recession (a 5 per cent increase in unemployment), is associated with an average initial loss in earnings of around 9 per cent.[7] The strength of this association halved after five years but was found to persist for as long as ten years.

[4] Note that not all graduates will be borrowers, and similarly not all borrowers will have graduated. BIS's modelling includes an adjustment for its estimate of the proportion of borrowers who will withdraw from their course.

[5] M Brynin, 'Individual Choice and Risk: The Case of Higher Education', *Sociology*, vol. 47 issue 2, April 2013, pp. 284-300

[6] F Green and Y Zhu, 'Overqualification, job dissatisfaction, and increasing dispersion in the returns to graduate education', *Oxford Economic Papers*, vol. 62 issue 4, 2010, pp. 740-763

[7] P Oreopoulos, T Wachter and A Heisz, 'The Short- and Long-Term Career Effects of Graduating in a Recession', *American Economic Journal: Applied Economics*, vol. 4 issue 1, January 2012, pp. 1-29

1.24 There is therefore evidence that using assumptions from historic data to represent the earnings dynamics of future cohorts may introduce inaccuracy to forecasts. Future estimates of the starting salary for borrowers with no SLC repayment data (i.e. all those taking loans out after 2012-13 or where there is missing data) may be biased as the income percentiles of these borrowers and the income distribution are randomly based on the 'simulated borrower grouping' which is based on the historic BHPS and LFS survey data - not post-recession data. And income transitions may be incorrect as they do not reflect the latest evidence on the dispersion of returns within the cohort of graduates.

1.25 The HERO model, in its forecasting of future earnings, takes into account various borrower characteristics, such as age, gender and degree type. However, it does not factor in data on higher education institution attended or subject studied. We recognise that including this additional data would make matrix-based modelling extremely complex. However, BIS is now developing a stochastic wage model and, depending on how its approach develops, there are likely to be benefits in making better use of borrower data. Our analysis in Part Two indicates that there is a statistically significant correlation between the type of institutions attended and subjects studied with the income of borrowers, even when prior years of earnings are controlled for. BIS should consider how this information would affect the accuracy of its repayment forecasts.

1.26 The HERO model and its forthcoming replacement do not explicitly attempt to capture changes in the subject or provider breakdown of student cohorts which may occur in future. In as much as these developments are likely to change the income dynamics of the cohorts which are represented within the modelling, omitting them from the design of the model could introduce inaccuracy to the forecasting of loan repayments. The implications could be more profound for the cohorts who have entered the labour market after the start of the 2008 recession, with a likely fall in income persisting over several years.

Part Two

Multivariate data analysis

2.1 In Part One of this paper we highlighted that certain borrower characteristics which are not used by the HERO model - such as subject studied and university attended - could potentially influence borrower earnings, and therefore repayment. This section outlines the analysis we employed to test this assertion.

2.2 Forecasting future repayments is complex, and refining the model is a continuous process. BIS is aiming to develop an improved model, and this paper suggests ways that it might be refined further. Our aim was to consider the impact of subject or university on the propensity of borrowers to repay, and to explore the potential benefits if BIS were to factor these into its modelling.

2.3 We would caution against the use of our analysis to support arguments in favour of expanding the intake of courses or HEI groups shown to have an association with higher borrower earnings. We have not made any assessment of causality, or controlled for borrowers' circumstances (for example, prior attainment or social background), as we were not attempting to assess the added value of attending certain universities or studying particular subjects.

2.4 Differential admission requirements for entry to these courses and their providers may introduce a selection bias, wherein it is difficult to determine whether borrowers who go on to earn a high salary do so because of characteristics present prior to their first degree (which secured them entry), or because of the skills their course of study has subsequently equipped them with. Indeed, research commissioned on behalf of the Department does not suggest that there are large differences in wage returns across broad types of HEI, when controlling for family background.[8]

Data

2.5 We carried out our analysis using data from the Student Loan Company's databases which holds information on current and previous Income-Contingent Repayment (ICR) loan borrowers. We restricted our analysis to the 2,638,451 unique borrower profiles from cohorts 2000 - 2012. This data was extracted on 30th of April 2012.

[8] I Walker and Y Zhu, 'The impact of university degrees on the lifecycle of earnings: some further analysis', BIS Research Paper No. 112, August 2013.

2.6 The extract taken from the Student Loan Company's databases is a cross-sectional dataset which records various borrower characteristics. We merged this dataset with data provided by HMRC on borrower earnings records, and created a single longitudinal dataset which we used to perform our analysis. The characteristics we used in our analysis were:

- **Gender -** the borrower's gender
- **Date of birth -** the borrower's date of birth
- **Cohort -** the year in which the Statutory Repayment Due Date (SRDD) of the loan falls - normally the April following the course end date.
- **Qualification type -** the type of degree which attracted funding (e.g. BA Hons)
- **Higher Education Institution -** the name of the institution attended by the borrower
- **Degree subject grouping -** as defined by the Higher Education and Skills Agency (HESA)'s Joint Academic Coding System (JACS), version 2.0.
- **Repayment status -** the status of the borrower as recorded by the SLC at 30/04/2012 (e.g. "Repaying through HMRC (PAYE and SA)")
- **Income by tax year for student borrowers** - Nominal earnings from cohorts 2000-2010 derived from employer and self-assessment submissions to HMRC, which provides this data to the Student Loans Company. This data was then transformed to its 2009 equivalent level using the ONS Average Earnings Index.

Methodology

2.7 Our analysis aims to explore whether including extra variables relating to borrower characteristics offers the potential to improve the predictive fit of the HERO model. Our literature review uncovered some evidence that graduate earnings are affected by subject of study[9] and the institution at which study takes place.[10] It is therefore plausible that income-contingent loan repayments would also be affected by these factors.

[9] G Conlon and P Patrignani, '*The returns to higher education qualifications',* BIS Research Paper Number 45, June 2011.

[10] A Chevalier and G Conlon, '*Does it pay to attend a prestigious university?'*, Centre for the Economics of Education, London School of Economics, May 2003.

2.8 We focused our analysis on the ability of the data to forecast borrower incomes, as this is the core part of the HERO model. In particular, we were interested in testing whether borrower information such as subject of first degree and HE institution attended have an impact on two features, both of which directly influence the propensity of borrowers to repay, and the level of repayments made:

- The amount earned over time; and
- Income dynamics - i.e. the transition of borrower earnings from one income percentile to the next.

2.9 In order to tackle these questions, we first recoded SLC data categories into new groupings for subject type, HEI type and repayment status, so as to be able to present findings in a more easily understandable format. Figure 5 shows our categorisation of repayment status, and how the SLC definitions map onto this.

Figure 5

Reconciliation of SLC and NAO categorisations of repayment status

NAO Categorisation	Frequency	SLC Status Categories
No earnings information or in arrears	374,621	Overseas borrowers in arrears or not paying; borrowers with no match to HMRC data, or no current employment record; untraced or unclassified borrowers
Not repaying	903,529	Borrowers not earning enough to repay (unemployed, on incapacity benefit, economically inactive, or under the earnings threshold); borrowers for whom the Student Loans Company is awaiting their next tax return
Repaying on time	1,006,635	Borrowers repaying their loan on time through Pay As You Earn, Self-Assessment, the Repayment of Teachers' Loans Scheme, the Prevent Overpayment Scheme, or directly as overseas borrowers
Fully repaid or cancelled	353,666	Borrowers who have fully repaid or had their loans cancelled
Total	2,638,451	

Notes

1. The numbers of borrowers shown above are based on the data extracted in April 2012. This differs from the March 2013 data presented in the NAO report Student Loan Repayments.

2.10 Figure 6 sets out our subject definitions, and how they correspond to the top-level JACS code.

Figure 6

Reconciliation of JACS and NAO subject categorisations

NAO categorisation	Number of borrowers	JACS categories
Maths and Computer Sciences	181,736	G - Maths and Computer Sciences
Engineering	98,468	H - Engineering
Science & Technology	390,618	K - Architecture, Building ; C - Biological Sciences; J - Technologies; D - Veterinary Sciences, Agriculture and related subjects; F - Physical Sciences
Business & Administration	273,504	N - Business & Administration studies
Languages	142,887	T - Eastern, Asian , African & American languages; R - European languages & literature; Q - Linguistics, Classics & related subjects
Social Studies	414,217	V - Historical & Philosophical studies; P - Mass Communication & Documentation; L - Social Studies
Education	253,021	X - Education
Art & Design	309,815	W - Creative Arts & Design
Law	122,020	M - Law
Medicine	198,798	A - Medicine & Dentistry; B - Subjects allied to Medicine
Not recorded / Other	253,367	No code provided
Total	2,638,451	

2.11 In order to assess the level of education attained by borrowers, we also created a new variable which assigned a category to the borrower based on whether their first degree was for a sub degree or not (Figure 7).

Figure 7

Highest Level of education attained

NAO Categorisation	Number of borrowers	Description
Sub	187,794	One year diplomas and foundation courses
First or Higher	2,450,657	Undergraduate 3yr honours degrees or above
Total	2,638,451	

2.12 SLC data contained borrowers attending 825 HE institutions. We aggregated these HEIs into seven groups based on institutional membership information available on the websites of the respective groups in 2013 (Figure 8).

Figure 8

University Groups used in our analysis

Group	Number in sample	Web-site for more information
Russell Group	568,293	http://www.russellgroup.ac.uk/our-universities.aspx
1994 Group	148,240	http://1994group.co.uk/universities.php
University Alliance	664,728	http://www.unialliance.ac.uk/member/
Million+	372,675	http://www.millionplus.ac.uk/who-we-are/our-affiliates/
GuildHE	154,162	http://guildhe.ac.uk/members
Other Large HEIs	518,091	Non-affiliated HEIs (> 2,000 borrower profiles in SLC database)
Other Small HEIs	212,262	Non-affiliated HEIs (< 2,000 borrower profiles in SLC database)
Total	2,638,451	

Notes

1. The size and membership of the groups change over time, and so may now differ from their composition when many of the loans were taken out.
2. On 8 November 2013, 1994 Group announced that they were ceasing to exist as an affiliation.

2.13 In our results, we present descriptive statistics for repayment status on April 30 2012; first by HEI group and then by subject category of the first degree of study. This information excludes the 2011 and 2012 cohorts, as repayment status data for these cohorts was incomplete at the time an extract was taken from the SLC's systems.

2.14 In addition to descriptive statistics, we carried out multivariate regression analyses using a panel data approach. This involved a Generalised Least Squares (GLS) regression using similar regressors to those which BIS is using in its forthcoming updated Stochastic Earning Pathways (STEP) model. The purpose of using regression analysis is to assess whether there is a statistically significant relationship between borrower characteristics such as subject of study or HE provider which persists once other factors in the regression have been controlled for.

2.15 To quantify the likely degree of benefits in terms of forecasting accuracy from including any significant regressors in the forecasting model, we also performed a one-year-ahead forecast using regression coefficients and inputs from prior years. This was achieved by estimating regression coefficients for two models:

- The 'Base Case' - a specification of regression similar to that proposed by BIS for its forthcoming STEP model (five lags of earnings, age, gender and educational attainment included)
- The Augmented Model - the Base Case augmented with extra regressors (HE provider affiliation, and Subject type).

2.16 For forecasting, we used regressions performed on a sample of observations which did not include those from the forecast year. We then used the estimated regression coefficients to fit values to the forecast year.

2.17 To assess the accuracy of both models' predictions, we compared the total of the fitted values with the total of the SLC-held actual earnings figures corresponding to these fitted values. To ensure an appropriate basis for comparison, only observations with enough data-points to generate a forecast were used to compare models. This ensured that the comparison between model specifications was based on an identical number of estimated incomes. This comparison is intended to give an idea of the effect and magnitude of the factors influencing forecasting using regression analysis.

2.18 Finally, we performed illustrative analysis of two sample cohorts to demonstrate how subject type and HEI group can affect income dynamics over periods of more than one year. We compared the evolution of borrower income percentiles among borrowers from the 2005 cohort which shared the same age and gender, but which had studied at different universities or in different subjects. This is a similar conceptualisation of borrower earnings dynamics which is employed in the HERO model itself.

Main results

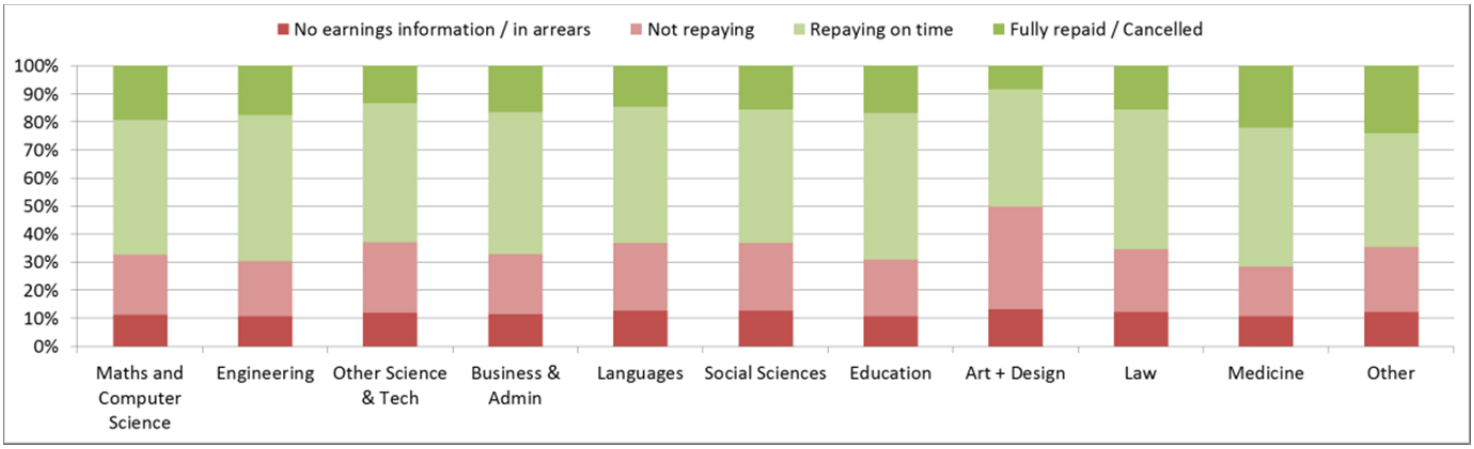
Descriptive analysis

2.19 Figure 9 shows overall repayment status on 30th April 2012 by first degree subject for the cohorts 2000 to 2010. It shows a somewhat similar propensity to be in repayment for all subjects with the exception of borrowers studying Art and Design - who display a visibly lower propensity to be in repayment.

2.20 Figure 10 overleaf illustrates overall repayment status on 30th April 2012 by HEI group for the cohorts 2000-2010. It suggests that borrowers who studied at the Russell Group and 1994 Group of universities display higher than average propensity to be in repayment compared to borrowers from other university groups - where there is little variation in the breakdown, on aggregate.

Figure 9

Repayment status by course area



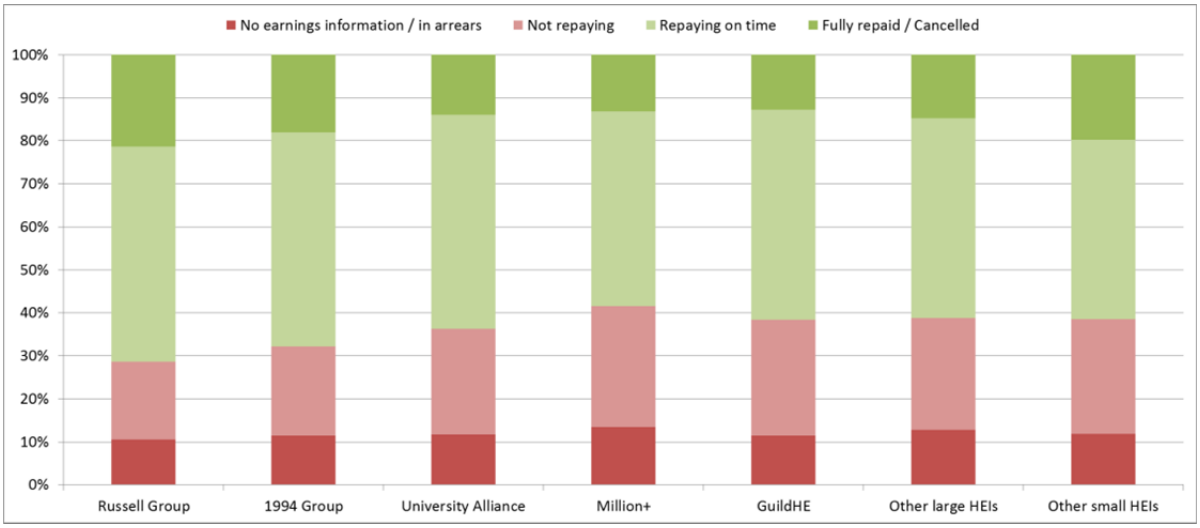
Notes

1. We have not explored the causation of these trends, or the prior attainment or circumstances of the borrowers. The figures should therefore not be viewed as an analysis of the added value of studying certain subjects.
2. Repayment status on 30/04/2012, cohorts 2000 - 2010 only.
3. Fully repaid and Cancelled share a grouping as they are outside the scope of the SLC's collection strategy.

Source: NAO analysis of SLC data

Figure 10

Repayment status by HEI Affiliation



Notes

1. We have not explored the causation of these trends, or the prior attainment or circumstances of the borrowers. The figures should therefore not be viewed as an analysis of the added value of attending certain universities.
2. Repayment status on 30/04/2012, cohorts 2000 - 2010 only
3. Fully repaid and Cancelled share a grouping as they are outside the scope of the SLC's collection strategy

Source: NAO analysis of SLC data

2.21 Figure 11 overleaf contains the tabulated regression coefficients of the Random-effects GLS regression containing the same variables as BIS's 'STEP' model, but with additional variables for course subject and HE provider affiliation. To mitigate for multicollinearity in the categorical variables used as regression coefficients, two categories have been omitted from the regression: 'Social Sciences' from the subject category, and 'University Alliance' from the HE provider category. These omissions form the base category against which the regression coefficients should be interpreted - i.e. a female borrower, who originally studied a Social Sciences Sub Degree at a University Alliance University.

2.22 The regression coefficients from Figure 11 should be interpreted with respect to the regression coefficient for the constant term, which represents the regression's current-year estimate of income, expressed in 2009 levels. For instance, the regression estimates that a borrower with the same characteristics as the base category save for having studied at a Russell Group university would earn on average £2,080 more per year than a borrower from the base category. Similarly, based on the regression coefficients, we would expect a borrower who studied an Art and Design-related subject to be earning £1,200 per year less, on average.

Figure 11

Random Effects[7] GLS Regression of borrower characteristics on current-year earnings[1]

Variable	Regression coefficient	Standard Error[2]	95% confidence interval (low)	95% confidence interval (high)
Earnings (t-1)	0.43	0.106**	0.22	0.64
Earnings (t-2)	0.16	0.0478**	0.06	0.25
Earnings (t-3)	0.13	0.0345**	0.06	0.20
Earnings (t-4)	0.09	0.0278**	0.04	0.14
Earnings (t-5)	0.03	0.0103**	0.01	0.05
Age	-103	10.9**	-124	-81
Maths + Computer Science[3]	500	135**	236	764
Engineering[3]	-264	106*	-471	-56
Science & Technology[3]	-212	45.7**	-302	-123
Business & Admin[3]	604	127**	356	852
Languages[3]	-542	71.5**	-682	-402
Education[3]	617	59.0**	501	732
Art & Design[3]	-1,200	147**	-1,489	-913
Law[3]	1,380	253**	884	1,880
Medicine[3]	-176	58.7**	-291	-61
Other[3]	40.2	51.1	-60	140
Russell Group[4]	2,080	300**	1,490	2,670
1994 Group[4]	1,190	177**	841	1,540
Million[4]	-396	69.0**	-531	-261
GuildHE[4]	-144	55.0**	-252	-36
Other Large HEIs[4]	175	35.0**	107	244
Other Small HEIs[4]	-407	85.6**	-574	-239

Gender[5]	1,310	135**	1,050	1,560
Educational Attainment[6]	1,100	147**	812	1,390
Constant	7,010	1,090**	4,860	9,150

Notes

1. Based on 655,686 observations.
2. Asterisks denote statistical significance (* = 95% and ** = 99% confidence leve)l
3. Subject dummies (omitted category = Social Sciences).
4. HEI group dummies (omitted category = University Alliance).
5. Gender dummy (1=Male).
6. Educational attainment dummy (1=First or higher, 0=Sub).
7. Results of Breusch-Pagan LM test for random effects P<0.0000 - reject null hypothesis of no random effects.
8. We have not explored the causation of these trends, or the prior attainment or circumstances of the borrowers. The figures should therefore not be viewed as an analysis of the added value of attending certain universities or studying certain subjects.

Source: NAO Analysis of SLC data

2.23 Figure 12 overleaf depicts the results of our forecasting exercise. We report total earnings for our out-of-sample one-year-ahead prediction fitted using the regression coefficients for each model. The difference versus the corresponding SLC actual earnings data is expressed under each estimate in brackets, as a percentage of the actual. For illustrative purposes, we also provide one-year forecasts for two models which include only one year of lagged earnings instead of five.

Figure 12

One-year-ahead out-of-forecast income predictions (£m, 2009 levels)

Source:	2006	2007	2008	2009	2010
Actuals	£7,849	£12,403	£15,769	£16,120	£22,323
One-lag model	£6,560 (-19.7%)	£10,131 (-22.4%)	£13,220 (-19.3%)	£13,728 (-17.4%)	£18,607 (-16.6%)
Augmented one-lag model	£6,605 (-18.8%)	£10,206 (-21.5%)	£13,313 (-18.5%)	£13,839 (-16.5%)	£18,714 (-16.2%)
Actuals			£2,731	£4,845	£7,376
Five-lag model			£2,562 (-6.6%)	£4,627 (-4.7%)	£6,746 (-8.5%)
Augmented five-lag model			£2,569 (-6.3%)	£4,636 (-4.5%)	£6,761 (-8.3%)

Notes

1. Forecast error is given as a percentage of the corresponding actuals.
2. Regressors include: lagged earnings, age, gender, educational attainment, subject of study and HE provider affiliation.
3. One-lag and Five-lag model are identical in terms of regressors except for the number of lags of earnings included.
4. Augmented model contains dummies for HEI group and subject of study.

Source: NAO Analysis of SLC Data

Income Dynamics

2.24 Using observed earnings dynamics from the British Household Panel Survey, the HERO model calculates transitions made by borrowers from the income percentile in one year to the income percentile in the next. There are only four input variables used in this assessment - prior year income percentile, gender, age, and level of educational attainment.

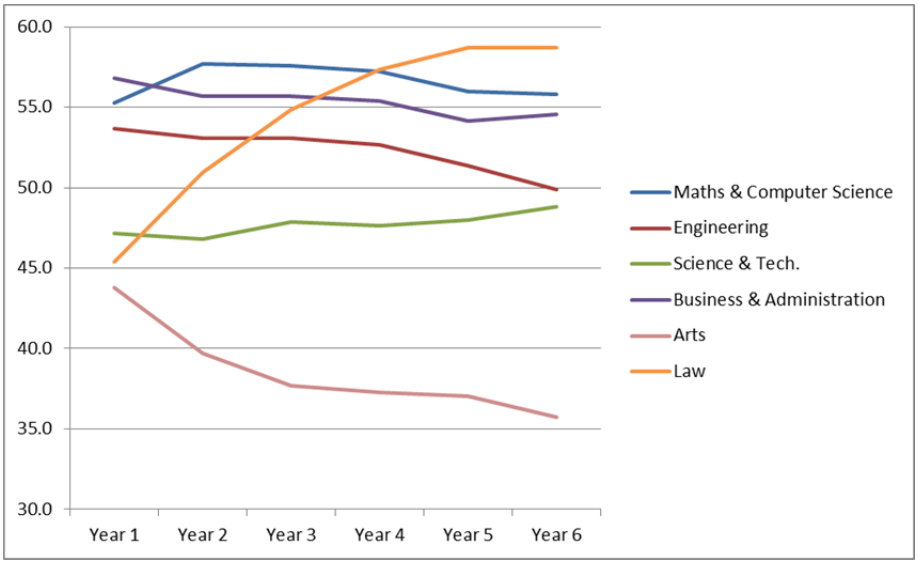
2.25 We have already seen in Figure 11 that the association between Subject and HEI affiliation on current year earnings is statistically and economically significant. The relatively modest improvements in one-year-ahead forecasts from Figure 12 notwithstanding, it is possible that the association may persist over multiple forecast years - thereby potentially introducing inaccuracy to forecasts.

2.26 A full analysis of the dynamic effects of Subject type and HEI type on borrower earnings is beyond the scope of this paper. However, we provide some evidence that modelling income transition solely using the HERO variables is unrealistic for some categories of borrower. We estimated income percentiles over six years for two categories of borrower in our sample which would be represented within the HERO model - 22 year old male borrower enrolled in a first degree, and 22 year old female borrower enrolled in a first degree. **Figure 13** sets out what the average percentile of the borrowers in the income distribution, by degree subject.

2.27 Figure 13 shows that over time, big differences emerge in average position in the income distribution when comparing borrowers by subject grouping. For instance, 22 year old male first degree borrowers enrolled in Law and Art + Design start in relatively similar percentiles in their SRDD year. However, within five years, earnings data suggests these two groups of borrowers are on average at opposite ends of the earnings spectrum for 27 year old male first degree borrowers.

Figure 13

Average income percentile by subject type, male first degree borrowers, 22 years old in year 1[1]



Notes

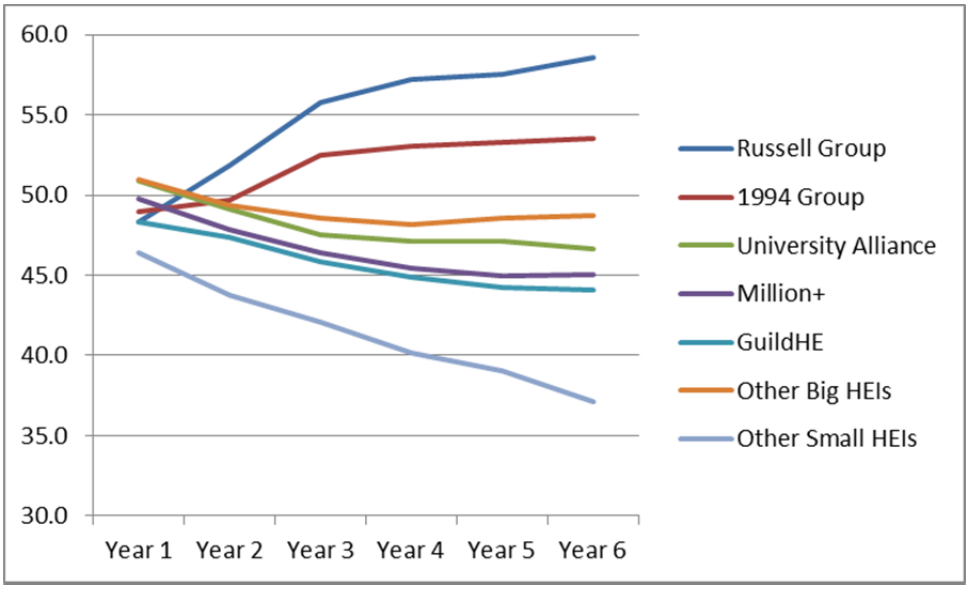
1. We have not explored the causation of these trends, or the prior attainment or circumstances of the borrowers. The figures should therefore not be viewed as an analysis of the added value of studying certain subjects.
2. Percentiles are calculated by year, using all profiles which had non-missing earnings data.
3. Some series have been omitted for clarity (Languages, Social Studies, Education, Medicine and Other).
4. Sample size = 199,372.

Source: NAO analysis of SLC data

2.28 A similar observation can be made regarding the earnings dynamics of Russell Group and GuildHE borrowers in the earnings distribution for 22 year old female first degree borrowers (**Figure 14**). GuildHE borrowers on average occupy a higher percentile on the income distribution compared with Russell Group borrowers. Five years later, the average percentile rank of Russell Group borrowers in the income distribution for 27 year old female first degree borrowers is thirteen percentage points higher than the average percentile rank for GuildHE borrowers.

Figure 14

Average income percentile by HEI group, female first degree borrowers, 22 years old in Year 1[1]



Notes

1. We have not explored the causation of these trends, or the prior attainment or circumstances of the borrowers. The figures should therefore not be viewed as an analysis of the added value of attending certain universities.
2. Percentiles are calculated by year, for all profiles which had non-missing earnings data.
3. Sample size = 286,300.

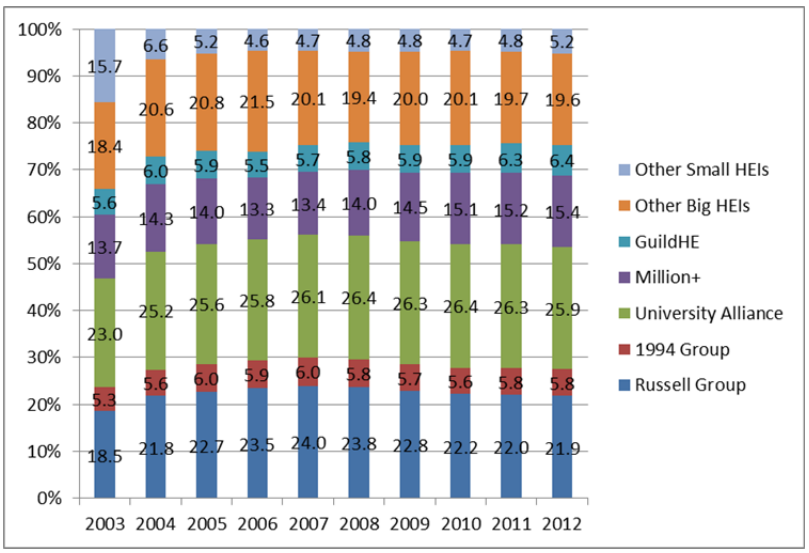
Source: NAO analysis of SLC data

2.29 Estimating the income transition effects for students of the same age and/or sex risks inaccuracy if changes in subject or HE provider composition in these groups over time is not accounted for. As the HERO model has been calibrated based on a particular breakdown of HE provider and subject types, there is a risk of inaccuracy being introduced if the share of individual groups with divergent characteristics increases in future years, and the model is not recalibrated to reflect this, or modified to incorporate the effect of these variables.

2.30 **Figure 15** shows that the share of borrowers accounted for by each university group has stayed relatively constant over time. In the last five years, there has however been a slight trend for the Russell Group universities to lose share and for the GuildHE and other smaller HEIs to increase it.

Figure 15

HEI groups' percentage share of borrower cohorts over time, 2003-2012

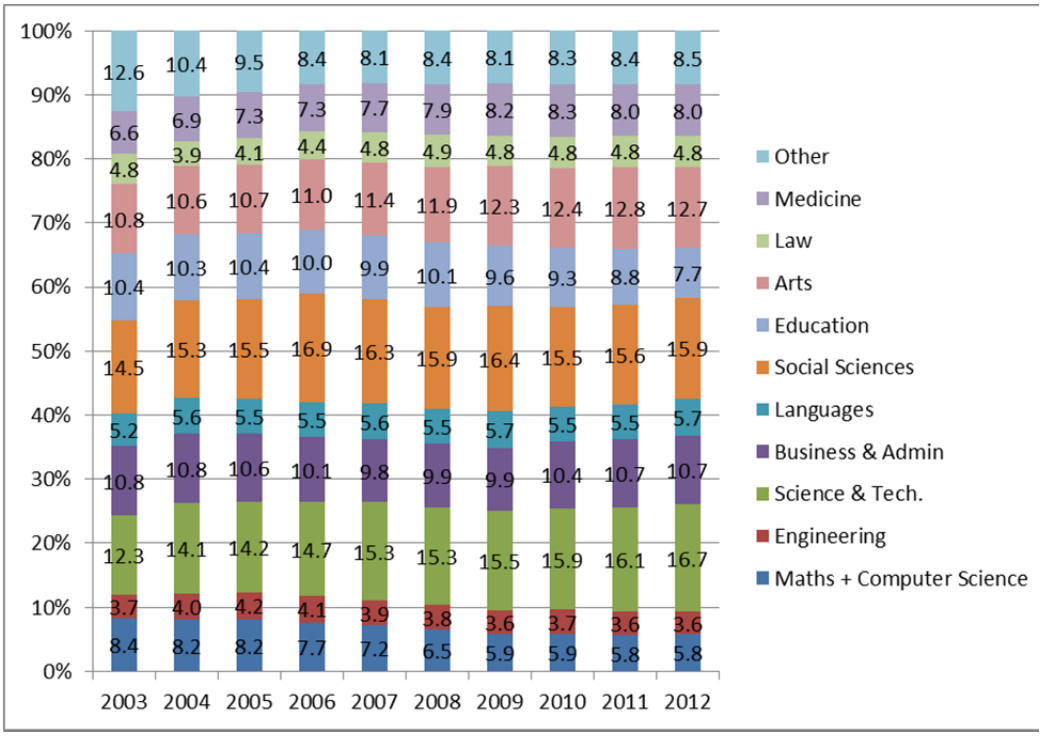


Source: NAO analysis of SLC data

2.31 Figure 16 shows that there have not recently been any large shifts in the composition of subject area amongst borrower cohorts, though there is a slight declining trend in Maths and Education, while Science and Technology subjects have shown steady increases in recent years.

Figure 16

Percentage share of subject area in borrower cohorts, 2003-2012



Source: NAO Analysis of SLC data

Discussion

2.32 Our analysis has led us to two conclusions. Firstly, it is clear that adding more than one lagged year of earnings as an explanatory variable substantially improves the predictive fit of our model (Figure 12). BIS proposes that its new STEP model will use up to five lags of earnings to simulate borrower earnings. Our analysis indicates that this could help BIS forecast more accurately than with the HERO model.

2.33 Secondly, we conclude based on our regression analysis that subject type and HE provider affiliation are associated with current-year earnings in a statistically significant way, once other factors such as lagged earnings are accounted for (Figure 11). As set out in Figure 12, the effect of including these variables on one-year-ahead forecast quality is, however, relatively modest. Our one-year-ahead forecast analysis suggests that adding these new variables to the underlying regression results in an improvement to forecast accuracy of approximately 0.2 to 0.3 per cent, or £7-10 million, when considering the model using five lags of earnings.

2.34 This is a small figure in the context of the £826 million of repayments received in the 2009-10 financial year, but it is plausible that this improvement in the forecast error underestimates the true extent of potential benefits to incorporating the extra data into the modelling. Firstly, Figures 15 and 16 show that there have been relatively minor changes in the composition of cohorts. This may cease to be the case in future as the increased role of student choice and the effects of new funding arrangements cause popular HE providers and subjects to increase their share.

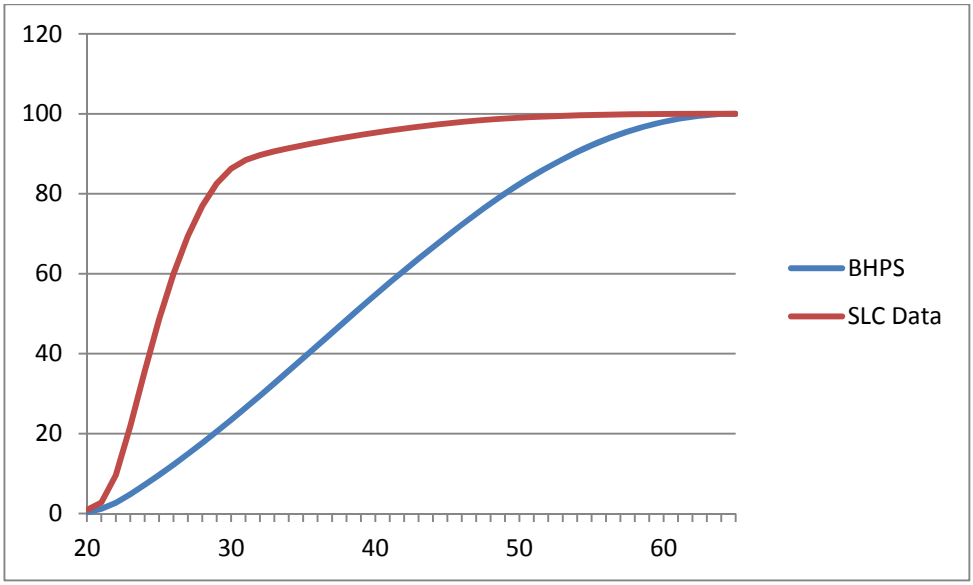
2.35 It is also worth noting that the improvement in forecast error only represents the improvements from a single year. Given that the size of the loan book is projected to increase from £46 billion to £200 billion by 2042, the improvement in forecast accuracy could well be larger, and increase over time.

2.36 Up to this point, BIS has mainly drawn on the BHPS, rather than the SLC's data, in building its forecast models. There are some good reasons for this - our dataset covers the period 2000-2010, whilst the BHPS covers 1991-2009, giving it greater representation of older graduates (**Figure 17**).

2.37 Given its unbalanced coverage of the UK population, we do not envisage that a modelling approach solely relying on the SLC data would be more accurate than one which uses the BHPS. The BHPS's usefulness in the proposed specification of the STEP model is however limited by data scarcity for younger borrowers - the BHPS has notably fewer graduates in the 20-25 year old age range than in almost any other range it covers.

Figure 17

Cumulative percentage of the BHPS and SLC graduate populations by age



Source: NAO analysis of the BHPS and SLC data

2.38 This has implications for BIS's approach which go beyond the impact on forecast accuracy for the income of 20-25 borrowers alone. Regression coefficients estimated from small sample populations may be biased. Moreover, borrowers in this age range will not have five lags of earnings to use for predicting their wages. This will increase the value of introducing variables such as Subject type and HEI affiliation, as Figure 12 suggests that these factors have greater impact on forecasting accuracy when fewer than five lags are present in the wage regression. This is an important consideration, as the one-year-ahead income forecast used by the model for new borrower cohorts will itself be used as an input to forecast future earnings in the forecast - potentially compounding any error introduced in the initial forecast.

2.39 It would therefore seem that the richness of the SLC dataset presents opportunities to increase the accuracy of the STEP model's forecasts, by fully reflecting the heterogeneity of this population in the regression coefficients used to fit earnings values. BIS should therefore consider whether there is a case for using the more detailed borrower data held by the SLC to improve the accuracy of its modelling.

Design and Production by
NAO Communications
DP Ref: 10307-002

National Audit Office