MINISTRY OF EDUCATION
Te Tāhuhu o te Mātauranga

Statistical methods and results for *Was it worth it?*

New Zealand Government

Beyond tertiary study

This report forms part of a series called Beyond tertiary study.
Other topics covered by the series include how graduates' earnings change over time, labour market outcomes, education and economic growth, and qualifications and income.

Author
Ralf Engler, Senior Research Analyst
Email: ralf.engler@minedu.govt.nz
Telephone: 04-463 7039

Acknowledgements
The author gratefully acknowledges comments provided by Roger Smyth, Zaneta Park, Warren Smart, David Earle, Sarah Crichton and Isabelle (Izi) Sin.

All views expressed in this report, and any remaining errors or omissions, remain the responsibility of the author.

Published by
Tertiary Sector Performance Analysis
Strategy and System Performance
MINISTRY OF EDUCATION

This report is available from the Ministry of Education's Education Counts website:
www.educationcounts.govt.nz.

June 2014

ISBN (web) 978-0-478-42299-3

Statistical methods and results for *Was it worth it?*

1 INTRODUCTION

This document describes the technical and methodological details used in the report *Was it worth it? Do low-income New Zealand student loan borrowers increase their income after studying for a tertiary qualification?* (Engler 2014).

Three statistical techniques were used. **Individual growth models** were used to look at how income changes before and after studying for a tertiary qualification. These are regression models which track income for *individual* people over time, contingent on a number of controlling factors. They are also known as random coefficient models, multilevel models, mixed models and hierarchical linear models in the literature.

Multiple correspondence analysis was used to look at the association between the subjects students studied in the tertiary qualifications they completed, and the demographic characteristics of those students. It is used to transform multivariate categorical data to a low-dimensional graphical representation. This method is also known by other names, including optimal scaling and reciprocal averaging.

Quantile regression was used to derive the post-completion distribution of income, controlling for the same factors as we used in the individual growth models. Unlike ordinary least squares regression, which models the relationship between one or more covariates and the conditional mean of the response variable, quantile regression extends the regression model to use conditional quantiles of the response variable, such as the median, or some other percentile.

This document contains no conclusions or interpretations of the data. These are in the original report. But this report does contain results tables from the regression models, which knowledgeable readers can interpret. This information is used to provide confidence in the results of the analyses. Future research into this type of question can build on and extend the work that we have done here.

This report begins by describing the data used in the various analyses, and then considers each of the three statistical methods in turn. An appendix contains a variety of statistical tables.

All statistical procedures were carried out at Statistics New Zealand premises, in their secure 'DataLab' (see next section for more details). The software used was SAS Enterprise Guide Version 4.3, although Base SAS was used to run the statistical procedures, not Enterprise Guide.

2 THE INTEGRATED DATA INFRASTRUCTURE

This study used the data in Statistics New Zealand’s Integrated Data Infrastructure (IDI). The IDI consists of a number of datasets about a person’s tertiary enrolments and completions, information on people’s incomes, whether they were on a benefit, and what type of benefit, details about a person’s student loan, and the dates when a person either left or entered the country. These disparate datasets are linked by StatisticsNZ and confidentialised, so that a longitudinal picture of a person’s tertiary education and employment can be constructed.

There are no names or other personal identifying information in the data.

All processing of the data was performed on StatisticsNZ premises, in a secure area known as the ‘DataLab’. All results from our analysis was first checked by StatisticsNZ staff to ensure confidentiality of the output. In addition, both the original and this report were checked by StatisticsNZ staff to ensure the results were appropriately confidentialised.

The following is the disclaimer required by StatisticsNZ for any work published using IDI data.

The results in the tables in this report are not official statistics, they have been created for research purposes from the Integrated Data Infrastructure (IDI) managed by Statistics New Zealand.

The opinions, findings, recommendations and conclusions expressed in these tables and accompanying report are those of the authors not Statistics NZ.

Access to the anonymised data used in this study was provided by Statistics NZ in accordance with security and confidentiality provisions of the Statistics Act 1975. Only people authorised by the Statistics Act 1975 are allowed to see data about a particular person, household, business or organisation and the results in this report have been confidentialised to protect these groups from identification.

Careful consideration has been given to the privacy, security and confidentiality issues associated with using administrative and survey data in the IDI. Further detail can be found in the Privacy impact assessment for the Integrated Data Infrastructure available from www.stats.govt.nz.

The results are based in part on tax data supplied by Inland Revenue to Statistics NZ under the Tax Administration Act 1994. This tax data must be used only for statistical purposes, and no individual information may be published or disclosed in any other form, or provided to Inland Revenue for administrative or regulatory purposes.

Any person who has had access to the unit-record data has certified that they have been shown, have read, and have understood section 81 of the Tax Administration Act 1994, which relates to secrecy. Any discussion of data limitations or weaknesses is in the context of using the IDI for statistical purposes, and is not related to the data's ability to support Inland Revenue's core operational requirements.

3 THE STUDY POPULATIONS

We started with the entire New Zealand population present in the Integrated Data Infrastructure dataset, from the years 1999 to 2008; this latter year was the latest data for people's incomes from Inland Revenue at the time of the original study.[1] We then selected anyone who spent more than 9 months in any one of those years in New Zealand, to provide us with a New Zealand resident population. We excluded anyone with an income more than $100,000 in 2008 dollars in any of the years. This served to remove a number of outliers, earning more than a million dollars in some years. Finally we excluded people who had zero income, from all sources, in each of the 10 years of interest.

The study population is made up of people who never studied at tertiary level during the ten year study period, and those who had studied, some of whom completed their qualifications. This enabled us to determine pre-study, during study, post-study and post-completion incomes, and the changes in income associated with these events.

One of the factors we considered was age. Since age varies with time, we chose birth year as the variable to account for differences in people's age. We could then follow a birth year cohort across time, and compare their income with people born in a different year, across the same time period. There are a number of different birth years present in the original population, so we made the decision to run separate regression models for different birth year cohorts. However, single birth year cohorts are not that large in New Zealand, especially considering those who completed tertiary qualifications, so to ensure we had sufficient numbers of people to model robustly, we combined three birth year cohorts for each of the three 'age groups' used in the study. These will be described in the next section. The base study population was therefore all those people born in the nine birth years of interest, fulfilling the other selection criteria outlined above.

From this base study population we selected sub-sets for each of the three analyses. For the individual growth models, we chose those people whose income in 1999 was less than the 2008 student loan repayment threshold. This was done to answer the question of how income changed after completing a tertiary qualification *for people with low incomes*. Clearly, 'low income' is a relative term, and any number of definitions could have been used. For the purposes of this study however, in which we were interested in whether people earned sufficient income to be obliged to start repaying their student loans, it made sense to start with people whose incomes were below this threshold, and then to see if their income rose above this threshold after studying for a tertiary qualification.

This definition of 'low income' is not without its problems. While the study population for the individual growth models had an income below the 2008 threshold in 1999, on average, that income was rising. So it is possible that someone's income was above the repayment threshold in later years, independent of whether they studied for and completed a tertiary qualification. However, more stringent definitions run the risk of excluding people with rising incomes, thereby artificially lowering the average income for a group, and biasing any conclusions. On the other hand, to impose no restrictions on the starting income meant that the average income (for most groups) was already above the repayment threshold.

In practice, starting with people whose income was below the 2008 repayment threshold in 1999 was a reasonable compromise. On average, this group's average income was below the repayment threshold up to the time of completing a qualification. The only group whose pre-study incomes came close to the repayment threshold were women in their thirties or forties

[1] Income is from wages, salaries, self-employed income, benefits, paid parental leave and accident compensation payments. It excludes unearned income such as rents, dividends and interest payments.

who were in receipt of a benefit. In addition, the analysis using quantile regression provides estimates of pre-study income for people at different points in the post-completion income distribution, which again allows us to see how people with low-incomes fare after completing a tertiary qualification.

For the multiple correspondence analysis, considering the association between the subjects studied and students' demographic factors, we chose from the base study population those people who had an income in 1999 below the 2008 repayment threshold, and who had completed a tertiary qualification. We deliberately excluded those people who had studied but did not complete, in case there were systematic differences between completers and non-completers in the subjects of their qualifications.

For the analysis using quantile regression, where we look at the distribution of post-completion income, we also start with the sub-set of the base study population of all those people who completed a tertiary qualification. However, for this analysis we include people with any income (up to $100,000 per annum), so we could consider people with graduate incomes at the 25th, 50th and 75th percentiles for a particular combination of demographic factors. There was also a further requirement for this analysis, namely that to be included, there had to be a record in the data for one year prior to tertiary study—to obtain the pre-study income—and one year after completing a tertiary qualification—to obtain the post-completion income. The main difference between this analysis, and the individual growth modelling, is that the quantile regression only considers two points in time, pre-study and post-completion. Hence, it gives us no information about the dynamics of how income changes over time. However, the individual growth models do provide this information.

4 THE STUDY VARIABLES

In this section we describe all of the factors that are used in any of the statistical analyses we used in our study. Not all of them are used in each analysis.

We start with our dependent variable, pre-tax income.

Pre-tax Income

The data for this study is primarily circumscribed by the years of income data we had access to at the time of the analysis. This data, from Inland Revenue, contains the annual pre-tax personal income of people paying income tax in New Zealand for the ten years from 1999 to 2008 inclusive. Income includes wages, salaries, self-employed income, benefits, paid parental leave and accident compensation payments. It excludes unearned income such as rents, dividends and interest payments.

All incomes are converted to 2008 dollars using New Zealand's labour cost index[2] to facilitate comparisons with the 2008 student loan repayment threshold. For the analysis, incomes were divided by 1,000, and regression results are reported with this transformation.[3] In the text however, we report the un-transformed dollar amounts.

The Inland Revenue (IR) income data is reported as net of non-taxable deductions and self-employment expenses. Some people have larger deductions and expenses than their original income, so their taxable income is negative. In these cases, the fact that someone has been able to declare a loss in a particular year is quite clear—they have a negative taxable income. In other cases, a person may still have been able to reduce their taxable income through expenses and deductions, but their final taxable income is still positive, so the extent to which the taxable income is net of deductions won't be obvious. Since we can't distinguish between these two situations, we have not excluded anyone with negative incomes in any one year.

There are also some people in the IR data with very large annual incomes, in the millions of dollars. Typically this level of income is seen for one or two years, with either 'average' or zero incomes in other years. Including these high-income individuals in our study populations significantly increases the variation in our data, which in turn reduces the power of our statistical tests. Again, there were relatively few people with this level of income.

Finally, there were some people who had zero income in every year of the study data. These would include people totally dependent on others, but it also includes people who have died. At the time this study was done, it is not possible to reliably determine who has died using the StatisticsNZ IDI data. While some people with zero income in each year did take out a student loan, on balance it was felt that it was better to exclude this subset of the population. It is clear that people without income do not have paying jobs and are not eligible for benefits, so by themselves are not able to repay a student loan, although any loan might be repaid by a partner or spouse. Note that the study populations do contain people who might have had a zero income in any one of the years in the study period, just not in every year.

Studies using regression to model income typically use the natural logarithm of income as the dependent variable, particularly where the focus of interest is in the percentage increase in income for people with income at different levels. The log transformation also has the advantage of standardising variances, a requirement of most regression models.

[2] This index was chosen to be consistent with that used by Mahoney and colleagues (2013). The indexes used to do the conversions come from Statistics New Zealand, http://www.stats.govt.nz/infoshare/. All salary and wage rates were used to determine the index.

[3] This was done so the means, variances and standard errors in the models were smaller numbers. It has no other effect on the results or the conclusions.

However, interpreting log-transformed incomes is not straightforward. Converting the log of income back to dollars also does not help, because the resultant amount is the geometric mean, not the arithmetic mean (Sokal and Rohlf 1981). This makes comparing an estimated income for a particular set of factors back to a threshold value problematic, because geometric means are generally lower than arithmetic means. In effect, using a log transformation, and then reporting the back-transformed value, will under-estimate someone's income, so comparisons to the student loan repayment threshold will be unintuitive.

However, using untransformed income in the regressions leads to biases in the standard errors of the estimated regression coefficients (from which income is calculated), which in turn affects statistical significance tests. However, the estimates themselves—the average income for a particular set of factors in the regression model—are unbiased (Hayes and Cai 2007). Unbiased standard errors can be calculated using a technique called bootstrapping. This is the approach we used in this study for the regression on income trajectories. The bootstrapping technique is described in the section where it was used.

New Zealand based population

The focus of this study is on people living in New Zealand. This is because the rules regarding the repayment of student loans are different for people staying in New Zealand after they complete their study, compared to those who move overseas.[4]

We used border-crossing data from New Zealand's immigration service to determine the movement of people into and out of New Zealand. We only selected people who were in New Zealand for more than nine months in every year between 1999 and 2008. This seemed a reasonable amount of time for people to have earned an income in New Zealand. This definition is also in line with recent work done by the Ministry of Business, Innovation and Employment (Papadopoulos 2012) and the Australian Bureau of Statistics (2010). This effectively excludes people who take an extended overseas holiday, or who gain work experience overseas.

Time

Time is measured in years, starting in 1999. Year is centred on 1999, which means that in the model, 1999 is represented by 0, 2000 by 1, etc.

There is also a separate measure of time which is used after a person completes a qualification. The counter for this measure of time is zero for the year the qualification is completed, and then increments in lock step with the year variable in the following years. Using these two time variables enables us to track rates of change in income before and after a qualification is completed, and to separate the background rate of change in income from the extra premium in annual increments afforded by completing a qualification (Singer and Willet 2003).

For the analysis of changes in income on completing a qualification for people with different starting incomes, two time points are used. One is the year before enrolment in tertiary study, and the second time period is one year after completing the qualification. Where a qualification completion event was seen after the last year of enrolment for that qualification, we adjusted the completion year to be the last year of enrolment.

Gender

The individual datasets that make up the IDI often record the gender of the same individual. If the gender is recorded in error in one of the datasets, this will become apparent when Statistics New Zealand builds the IDI. We take advantage of the data cleaning done by Statistics New Zealand and use the gender determined to be the correct one during the IDI build process. This also applies to someone's birth date, from which we derive their birth year.

[4] The rules for paying back a student loan in New Zealand can be found at http://www.studylink.govt.nz/finishing-study/paying-back-your-loan/index.html.

Age and birth year

Income varies significantly with age. But it also varies over historical time for people of the same age. That is to say, a person who was 30 years of age in 1950 will not necessarily have the same income as someone who was 30 years of age in 1980, all else being equal. To account for this, we used birth year cohorts to control for age. Any one age group in our study consists of all people born in the same year. Unfortunately, there were too few people in any one birth year cohort to model robustly, particularly with the number of factors we use in the models, or when considering only people who have completed a particular tertiary qualification. So we have combined three birth year cohorts for each of our three age groups. There were differences in income between the three birth year cohorts *within* an age group category, but the differences were less than those *between* the different age group categories, so we did not include birth year as a separate variable in the models. While combining three birth year cohorts adds variation to the data which we do not account for in the models, this is mitigated by the larger sample sizes we are able to use. In practice, the standard errors for the regression coefficients were acceptably small.

In our initial analysis we actually used a fourth age group category, people in their fifties, but because the number of people in this age group was quite low for some combinations of factors, we did not formally report the results for this group in our original report.

The age group category labels and their constituent birth year cohorts are:

- people in their twenties, comprising people born between 1978 and 1980
- people in their thirties, comprising people born between 1968 and 1970
- people in their forties, comprising people born between 1958 and 1960
- people in their fifties, comprising people born between 1948 and 1950.

For simplicity, the age group categories are labelled as 20, 30, 40 and 50 years of age in 1999, or we use the phrasing, people in their twenties, thirties, forties or fifties.

By using birth year cohorts across the same period of historical time we implicitly control for the impact of economic and labour market conditions on income—all people were subject to exactly the same conditions between 1999 and 2008.

We should point out that by the end of the study period in 2008, those born in 1979 will be 29 years of age. These people cannot be compared to people born in 1969 when they are 30 years of age in 1999. That is, it is not appropriate to look at the results as one longitudinal study from those aged 20 to those aged 59, concatenating the ten year histories across birth year cohorts. This is because a person 29 years of age in 2008 will be experiencing quite different economic and labour market conditions to someone who was 30 in 1999, apart from the fact that their upbringing, education, income expectations, the types of employment available, the types of jobs they work in etc, would also have been different for these two birth year cohorts.

Benefit status

Beneficiary status was derived from records kept by the Ministry of Social Development. This data indicated the period of time, in months, a person was a beneficiary, what type of benefit was being received, and the amount of the benefit. The benefit types in the data were:

- Domestic purposes
- Unemployment
- Sickness
- Invalid
- Widow
- Emergency and hardship

- Independent Youth
- Orphan and unsupported child.

The domestic purposes and unemployment benefit types were the most common.

Being on a benefit can be modelled as a binary variable (a person is either on a benefit or not), or as months per year on a benefit. We tried months on a benefit, as a continuous variable in the regression model, and while it produced useful results for people who were on a benefit for 3 to 9 months in a year, the results were less realistic at the extremes, particularly for people 12 months on a benefit in any one year. To make the models work using months on a benefit, we would have needed to treat each month value as a category, but this would have made the models much more complex than they already were. Therefore out final models use benefit status as a binary variable; someone was considered to be on a benefit if they were in receipt of any benefit for one or more months in a year.

We also experimented with including the different benefit types as separate categories in the models. However, as Figure 5 in the original report shows, the type of benefit is correlated with gender, so only one of these factors can be included in a model without introducing co-linearity problems. We chose to use gender, so interpreting the effect of being a beneficiary on the income of a person of a particular gender or age, the reader will need to be mindful of the most likely type of benefit a person might be receiving.

The effect of these decisions on how to include benefit status in our modelling is that the factor ON BENEFIT in the models captures the *average* time spent on a benefit, for the *average* benefit type. The average time spent on a benefit varies with the type of benefit, and the average benefit type will vary with gender. For this reason, in our main report we have viewed benefit status as a proxy for ‘access to the labour market’.

Tertiary study

Whether and when a person was enrolled in a programme of tertiary study, and whether and when a qualification was completed, was derived from Ministry of Education data. The level of qualification enrolled in was also recorded, using the following categories; certificates at levels 1 to 3, certificates at level 4, diplomas and bachelors degrees.

People can enrol in more than one level of study, and these enrolments can be concurrent. We therefore noted the highest level of study in any one year, and the highest level studied during the years 1999 to 2008.

In some cases, the year a qualification is completed is the same as the last year of enrolment for a particular qualification. In other cases, the year of completion is the year *after* the last year of enrolment. In these latter cases, the completion year is adjusted to be the last year of enrolment.

Student loans

We used Inland Revenue data to determine who had taken out a student loan. Loan amounts of $20 or less are regarded as zero. A person is regarded as having a student loan if their outstanding loan balance was more than $20 in a particular year. We defined two variables for student loan status; one varied from year to year if a person had a student loan, while the other was true if a person ever had a student loan.

Field of study

We used the New Zealand Standard Classification of Education (or NZSCED) to classify the subject of a person’s qualification into various fields of study. [5] NZSCED has three levels of

[5] For the details of the New Zealand Standard Classification of Education, refer to: http://www.educationcounts.govt.nz/data-services/collecting-information/code_sets/new_zealand_standard_classification_of_education_nzsced

classification – broad, narrow and detailed field of study, although we only considered the broad and narrow fields. We used the NZSCED classification of the completed qualification as determined by the tertiary provider.

In our analysis, we used the field of study to explore the association between the subjects of completed qualifications and graduates' demographic factors, particularly gender. We were looking for broad patterns, so using the field of study as determined by the tertiary provider was adequate for the task. In later studies, it would be useful to consider other definitions of field of study, particularly those derived from the particular courses that make up a qualification.

5 HOW INCOME CHANGES FOR LOW EARNERS AFTER STUDYING FOR A TERTIARY QUALIFICATION

5.1 What was the statistical technique

We used individual growth models, considering changes in annual income against a range of predictors. These models are multi-level, considering change *within* people over time, for characteristics that change for any one individual person, and *between* people, for characteristics that differ between people. We closely followed the techniques outlined in the book *Applied Longitudinal Data Analysis* (Singer and Willett 2003), a framework for investigating change over time.

5.2 Where are the results reported

The results of this analysis are reported in section 3 of the original report.

5.3 The study populations used

Since this part of the analysis focussed on low income earners, people were selected if their income in 1999 (in 2008 dollars) was less than the 2008 student loan repayment threshold, which was $18,148. People with exactly zero income in each year of the study period were excluded, but a person could have zero income in any one year.

Each age group and each level of study were analysed separately, giving us a total of 16 different study populations.

Note that in the original report, only the results for the youngest three age groups were reported. There were too few people in their fifties completing higher-level qualifications for us to be as confident in the results. We discuss the results for those in their fifties in general terms in Section 5 in the original report.

The total number of people in each of the 16 study populations is given below.

Those not in study, or whose highest level of tertiary study was:	Age in 1999			
	20	30	40	50
Certificate at levels 1-3	61,041	46,302	43,842	35,307
Certificate at level 4	61,197	46,764	44,070	35,436
Diploma	64,713	48,519	45,336	35,856
Bachelors	72,954	50,301	46,371	36,078

Source: Statistics NZ Integrated Data Infrastructure, Ministry of Education interpretation.

All numbers are randomly rounded to base 3.

For any one age group, there is much overlap between the study populations across the highest level of tertiary study categories. For example, for young people with low incomes in their twenties, the 61,041 in the certificate at levels 1 to 3 study population includes people who never studied for any tertiary qualification at all, as well as those whose highest level of study was for a level 1 to 3 certificate. The next study population in that age group, 61,197 in size, again contains all those people who never studied for any tertiary qualification, as well as those whose highest level of study was for a level 4 certificate. While that group of people who never studied at tertiary level are included in all study populations within an age group category, anyone who completed a qualification at a particular level of study is not included in any other level of study population. And obviously there are no overlaps across age groups—people can't be more than one age at any one time!

Tables 7 to 10 in the Appendix show the sample size breakdowns of the various categories in the models. Sample sizes for those in their fifties were not considered large enough to produce robust results. The regression and other results are provided in this technical table for this age group, but in the original report, the results for this age group category are only described on general terms

5.4 Description of the method

We have used a multi-level model for change to analyse the data (Singer and Willett 2003). Specifically, we have modelled individual annual taxable income against time in years, controlling for various factors. Time is modelled non-linearly using a quadratic function, and the relationship between income and time may also be discontinuous, depending on whether or not a person studied for and then completed a tertiary qualification. The rate of change in income over time can also be discontinuous, with separate rates possible for the years before and after completing a qualification, or with changes in a person's benefit status, or if they take on or pay off a student loan.

The regression models we used contain fixed and random effects. Fixed effects are those which are constrained to be equal for a particular factor. In this study, we use fixed effects for gender, being on a benefit or not, being enrolled in a tertiary qualification or not, and completing a tertiary qualification or not. In essence we are stipulating that the effect of gender on income, for example, is the same for everyone with the same gender. This does not mean all men or all women have the same income. Rather, for all other factors being equal, being a woman results in the same difference in income, compared to a man, for every woman in a particular study population. The reader should note all regression models make this assumption. Random effects on the other hand are factors whose effects can vary between individuals. Random effects in our models are the pre-study income in 1999, the rate of change in income prior to completing a qualification, the rate of change in income after completing a tertiary qualification, and the step change in income on completing a qualification. This combination of fixed and random effects means these models are sometimes called 'mixed model regressions'.

The estimation method used to fit the multi-level model for change to the data was full maximum likelihood. This method was deemed on balance to be the best method for this analysis. We point out that this is not the default method used by the SAS software system. A full discussion of the estimation methods and their advantages and disadvantages can be found in Singer and Willett (2003, pp 85–92).

The regression model

A multi-level model for change can be written in two parts. The first part, known as the level-1 or the individual growth model, specifies *within*-individual change over time. In our case, it models how a person's income changes with time as a function of the factors that might vary with time for that individual. In the second part of the model, known as level-2 model, we specify *between*-individual differences.

Because the level-1 model describes within-individual change, the factors included in this part of the model must be able to vary across time for any one person. For example, a person may be enrolled in study in one year, but not in another year. A person may be in receipt of a benefit in one year, but not in the next year. Time in years, of course, varies with time. In our study, the only two variables which do *not* vary with time for any one person are gender and birth year, and birth year is controlled for by running separate models for each age-group category. All the time-varying variables other than time in years, and time since completing a qualification, are binary, yes-no indicators.

Whether a person has completed a tertiary qualification is also not constant through time, but once a qualification is completed, it can't be *un*-completed. In our study, we tracked all of the qualifications a person completed during the years 1999 to 2008, and used the highest level ever completed during the study period for that person.

The level-1 model for this study is:

$$
\begin{aligned}
&(\mathrm{INCOME}_{ij} \div 1000) \\
&\quad = \alpha_{0i} + \alpha_{1i}(\mathrm{ON\ BENEFIT}_{\mathrm{ij}}) + \alpha_{2i}(\mathrm{ENROLLED}_{\mathrm{ij}}) + \alpha_{3i}(\mathrm{HAS\ LOAN}_{ij}) + \alpha_{4i}(\mathrm{ON\ BENEFIT}_{ij} \times \mathrm{HAS\ LOAN}_{ij}) + \alpha_{5i}(\mathrm{COMPLETED}_{\mathrm{ij}}) \\
&\quad + \alpha_{6i}(\mathrm{COMPLETED}_{ij} \times \mathrm{ON\ BENEFIT}_{ij}) + \alpha_{7i}(\mathrm{COMPLETED}_{ij} \times \mathrm{HAS\ LOAN}_{ij}) + \alpha_{8i}(\mathrm{YEAR}_{ij} - 1999) + \alpha_{9i}(\mathrm{YEAR}_{ij} - 1999)^2 \\
&\quad + \alpha_{10i}\left((YEAR_{ij} - 1999) \times \mathrm{STUDIED}_{ij}\right) + \alpha_{11i}\left((YEAR_{ij} - 1999) \times \mathrm{ON\ BENEFIT}_{ij}\right) + \alpha_{12i}\left((YEAR_{ij} - 1999) \times \mathrm{HAS\ LOAN}_{ij}\right) \\
&\quad + \alpha_{13i}\left((YEAR_{ij} - 1999) \times \mathrm{STUDIED}_{ij} \times \mathrm{ON\ BENEFIT}_{ij}\right) + \alpha_{14i}\left((YEAR_{ij} - 1999) \times \mathrm{STUDIED}_{ij} \times \mathrm{HAS\ LOAN}_{\mathrm{ij}}\right) \\
&\quad + \alpha_{15i}\left((YEAR_{ij} - 1999) \times \mathrm{ON\ BENEFIT}_{ij} \times \mathrm{HAS\ LOAN}_{\mathrm{ij}}\right) + \alpha_{16i}\left((YEAR_{ij} - 1999) \times \mathrm{STUDIED}_{ij} \times \mathrm{ON\ BENEFIT}_{ij} \times \mathrm{HAS\ LOAN}_{\mathrm{ij}}\right) \\
&\quad + \alpha_{17i}(\mathrm{YEARS\ AFTER\ COMPLETING}_{ij}) + \alpha_{18i}(\mathrm{YEARS\ AFTER\ COMPLETING}_{\mathrm{ij}})^2 \\
&\quad + \alpha_{19i}(\mathrm{YEARS\ AFTER\ COMPLETING}_{ij} \times \mathrm{ON\ BENEFIT}_{ij}) + \alpha_{20i}(\mathrm{YEARS\ AFTER\ COMPLETING}_{ij} \times \mathrm{HAS\ LOAN}_{ij}) \\
&\quad + \alpha_{21i}(\mathrm{YEARS\ AFTER\ COMPLETING}_{ij} \times \mathrm{ON\ BENEFIT}_{ij} \times \mathrm{HAS\ LOAN}_{ij}) + \varepsilon_{ij}
\end{aligned}
$$

where the income (in thousands of dollars) for person *i* at time *j* is a quadratic function of time (in years, centred on 1999, and years after completing a qualification) and a range of other factors.

The term α_{0i} is the initial income in 1999, for a person *i* who is not studying, has not completed a qualification, who does not have a student loan, and is not on a benefit. The α terms 1 to 7 modify this pre-study income for people on a benefit (α_{1i}), who are enrolled in some tertiary study (α_{2i}), who have a student loan (α_{3i}), have completed a qualification (α_{5i}), together with their interactions, ON BENEFIT X HAS LOAN, which represents those people who are on a benefit *and* have a student loan (α_{4i}), COMPLETED X ON BENEFIT, which represents those who are on a benefit *and* completed a qualification (α_{6i}), and COMPLETED X HAS LOAN, which covers those who have a student loan *and* completed (α_{7i}). To illustrate how these work together to produce a particular income, the initial income (in 1999) of a person *i* not on a benefit, who is studying at tertiary level with a student loan (but who has not yet completed) is $(\alpha_{0i} + \alpha_{2i} + \alpha_{3i})$.

The remaining α terms 8 to 21 represent the rate of change in income over the next 9 years; terms 8 to 16 cover the time prior to completing a qualification, and terms 17 to 21 apply if a person completes a qualification. For example, for a person not on a benefit but who had a student loan while they studied, the annual rate of change in income is $(\alpha_{8i} + \alpha_{9i} + \alpha_{10i} + \alpha_{12i} + \alpha_{14i})$ for person *i* who has not completed their qualification, while it is $(\alpha_{8i} + \alpha_{9i} + \alpha_{10i} + \alpha_{12i} + \alpha_{14i} + \alpha_{17i} + \alpha_{18i} + \alpha_{20i})$ for person *i* if they *did* complete a qualification. In other words, the annual rate of change in income can be different for people who did and did not complete a qualification, and it can be different for the same person before and after completing a qualification. The quadratic term on time means that the overall rate of change in income is curvilinear. In general, income goes up over time, but the rate at which it increases slows down over time.

The model also allows a person's income to make a step change when they complete a qualification. For example, for a person not on a benefit and who didn't take out a student loan, their pre-study income is α_{0i}, and it is $(\alpha_{0i} + \alpha_{5i})$ after they complete a qualification. Of course if the results of the regression model show the term α_{5i} to be zero, then we can say there is no change in a person's income on completing a qualification for the particular combination of other factors in the model, and similarly for the other terms in the model.

It might not be clear from the foregoing as to whether the year in which a qualification is completed is included in the model—it's not; there is no YEAR x COMPLETED term in the model. This means that the coefficient for COMPLETED captures the average change in income on completing a qualification across all years, from which we can calculate a separate average value for men and women. While we lose information about the change in income across time for completing a particular level of qualification for a particular age group, we get a more robust estimate of this change integrated over the entire ten year period, because there are too few completion events in any one year to model robustly (see Tables 7 to 10 in the Appendix). But it will contribute to some of the variation we see in the change in income on completing (section 5.8 in this report). Section 5 in the original report discusses the economic and labour market conditions prevailing during the study period. These show that using an average change in income on completion is not unreasonable given how economic conditions changed over this period. We should also say the model does include terms which capture changes in income through time, including for those who completed a qualification.

The ε_{ij} terms in the model specification are the level-1 residuals, which is that portion of the income for person *i* in year *j* not explained by the other factors in the level-1 model. These will be the result of situations where income is associated with factors other than those in our level-1 model, including those which are particular to an individual person, or, where there are errors in the recording of income.

The level-2 model considers how these individual trajectories of income vary between men and women. The factor FEMALE is 1 for women and 0 for men. This means that the gamma terms in the level-2 model equal the alpha terms in the level-1 model for men. The level-2 model is:

$$\alpha_{0i} = \gamma_{00} + \gamma_{05}(\mathrm{FEMALE}_i) + \delta_{0i}$$

$$\alpha_{1i} = \gamma_{01} + \gamma_{06}(\mathrm{FEMALE}_i)$$

$$\alpha_{2i} = \gamma_{02} + \gamma_{07}(\mathrm{FEMALE}_i)$$

$$\alpha_{3i} = \gamma_{03} + \gamma_{08}(\mathrm{FEMALE}_i)$$

$$\alpha_{4i} = \gamma_{04} + \gamma_{09}(\mathrm{FEMALE}_i)$$

$$\alpha_{5i} = \gamma_{10} + \gamma_{13}(\mathrm{FEMALE}_i) + \delta_{1i}$$

$$\alpha_{6i} = \gamma_{11}$$

$$\alpha_{7i} = \gamma_{12} + \gamma_{14}(\mathrm{FEMALE}_i)$$

$$\alpha_{8i} = \gamma_{15} + \gamma_{24}(\mathrm{FEMALE}_i) + \delta_{2i}$$

$$\alpha_{9i} = \gamma_{16}$$

$$\alpha_{10i} = \gamma_{17} + \gamma_{25}(\mathrm{FEMALE}_i)$$

$$\alpha_{11i} = \gamma_{18} + \gamma_{26}(\mathrm{FEMALE}_i)$$

$$\alpha_{12i} = \gamma_{19} + \gamma_{27}(\mathrm{FEMALE}_i)$$

$$\alpha_{13i} = \gamma_{20} + \gamma_{28}(\mathrm{FEMALE}_i)$$

$$\alpha_{14i} = \gamma_{21} + \gamma_{29}(\mathrm{FEMALE}_i)$$

$\alpha_{15i} = \gamma_{22} + \gamma_{30}(\text{FEMALE}_i)$

$\alpha_{16i} = \gamma_{23} + \gamma_{31}(\text{FEMALE}_i)$

$\alpha_{17i} = \gamma_{32} + \gamma_{37}(\text{FEMALE}_i) + \delta_{3i}$

$\alpha_{18i} = \gamma_{33}$

$\alpha_{19i} = \gamma_{34}$

$\alpha_{20i} = \gamma_{35} + \gamma_{38}(\text{FEMALE}_i)$

$\alpha_{21i} = \gamma_{36}$

Specifically:

- Pre-study income (α_{0i}), the linear component of the rate of change in income prior to completing a qualification (α_{8i}), the change in income on completing a qualification (α_{5i}), and the linear component of the rate of change in income after completing a qualification (α_{17i}) are random coefficients, indicated by the inclusion of level-2, or between person, residual terms ($\delta_{.i}$). These are the terms that allow individuals' income to vary in these characteristics.

- We are also assuming that gender has no effect on the quadratic part of the rate of change in income either before or after completing a qualification. There is therefore no female term for coefficients α_{9i} and α_{18i}. These coefficients will therefore account for the average curvilinear effect across men and women.

- The two third-order interaction terms between gender and being on a benefit and completing, and between gender and time since completing and being on a benefit are also excluded as these were found not to be significant in nearly all models. The second-order interaction terms are included. This also applies to the interaction between time since completing, being on a benefit and having a student loan. Essentially we are saying that there is no difference between men and women for these categories. These correspond to coefficients α_{6i}, α_{19i} and α_{21i} respectively.

Both level-1 and level-2 residual terms are assumed to have normal distributions. We relax this assumption below. These assumptions can be written as:

$$\varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2)$$

$$\begin{bmatrix} \delta_{0i} \\ \delta_{1i} \\ \delta_{2i} \\ \delta_{3i} \end{bmatrix} \sim N \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_0^2 & \sigma_{01} & \sigma_{02} & \sigma_{03} \\ \sigma_{10} & \sigma_1^2 & \sigma_{12} & \sigma_{13} \\ \sigma_{20} & \sigma_{21} & \sigma_2^2 & \sigma_{23} \\ \sigma_{30} & \sigma_{31} & \sigma_{32} & \sigma_3^2 \end{bmatrix}$$

These functions indicate that the within-person residuals (ε_{ij}) are normally distributed with a mean of zero, and variance of σ_ε^2. The between-person residuals ($\delta_{0i}, \delta_{1i}, \delta_{2i}, \delta_{3i}$) are also normally distributed, with mean of zero, with variances of σ_0^2, σ_1^2, σ_2^2 and σ_3^2 respectively, and with co-variances between these variance terms. The level-2 error co-variance matrix is diagonally symmetric, such that σ_{01} is the same as σ_{10}, and similarly for the other co-variance terms. The complete set of variances and co-variances, at both level-1 and level-2 are known collectively as the model's variance components. These variance components are estimated by the modelling process, and provide important information about the relationship between the random effect factors.

The covariances of the level-2 residuals (σ_{01} to σ_{23}) have an important interpretation; they quantify the co-variance between each of the random effects in the model. An easier way to interpret these co-variances is to express them as a correlation coefficient, which is done by dividing them by the square root of their associated variance components (Singer and Willett 2003). For example, the correlation between pre-study income and the rate of change in income prior to completing is calculated as:

$$\rho_{01} = \frac{\sigma_{01}}{\sqrt{\sigma_0^2 \sigma_1^2}} \qquad \text{(Equation 1)}$$

Finally, one of the decisions that need to be made in a regression with random effects is to specify the structure of the error co-variance matrix. Various structure types are available, with varying levels of symmetry and constraints on the inter-relationship between the error co-variance parameters. The alternative to these symmetric error co-variance structures is the 'unstructured' type, which imposes no constraints on the error co-variance parameters. In the unstructured case, the error co-variance parameters are solely determined by the data, rather than having a particular relationship imposed on them. It is beyond the scope of this report to describe these error co-variance matrix structures in detail, but technical details can be found in Singer and Willett (2003, p 256). We tested a number of these structured error co-variance matrices to see if they produced models with a better fit to the data, as determined by goodness-of-fit statistics, than the model using an unstructured error co-variance matrix. We found that none of the alternative error co-variance matrices produced a model with a better fit to the data than the unstructured error co-variance matrix.

Table 1 in the Appendix maps the model parameters as described in this section to an example regression output. Tables 2 to 5 show the full regression outputs for each model.

5.5 Assumptions of the method

This type of regression has all of the assumptions of ordinary least squares regression, including normality of residuals, and homogeneity of variances.

Unfortunately, the latter assumption does not hold in our case. This is because the distribution of income between people diverges with time—their income trajectories can follow quite different paths, as we have seen. The usual method to standardise the variances is to take the log of the incomes.

Heterogeneous variances which arise from using untransformed income in the regression models lead to biases in the standard errors of the estimated regression coefficients, which in turn affects statistical significance tests. However, the estimates themselves—the average income for a particular set of factors in the regression model—are unbiased (Hayes and Cai 2007). However, unbiased standard errors can be calculated using a technique called bootstrapping.

Tests also showed that the residuals from the models were not always normally distributed, another assumption of regression analysis. This is also likely to result from the fact that the distribution of incomes was not consistent across time. Again, using bootstrapping to estimate standard errors relieves us from having to assume both normality of residuals, and homogeneity of variances.

5.6 Bootstrapping

The bootstrap procedure allows us to make statistical inferences about the regression coefficients without making the strong distributional assumptions normally required for regression analyses (Mooney and Duval 1993). The bootstrap approach is to treat the sample of individuals in our study populations as if they are the population to which we want to infer our results. From this population we draw many repeated random samples, *with replacement*, of the same size as the original study population. We perform the full regression analysis on each of these random samples, and consider the distribution of the estimates of the various regression coefficients. The mean of these coefficients calculated over all of the random samples is an estimate of the true coefficient, and the standard deviation of these means is an estimate of the standard error of the coefficient. When the sample results for any one coefficient are ranked, the $\left(\frac{n}{2}\right)^{th}$ and $\left(100-\frac{n}{2}\right)^{th}$ percentiles can be used as estimates of the $(100-n)^{th}$ per cent confidence limit. For example, when n is 5, the values at the 2.5th and 97.5th percentiles are the 95 per cent confidence limits. If this range does not include 0, then we can be 95 per cent certain the regression coefficient is different from zero.

In our analysis we used 500 random samples in the bootstrap procedure to give us an empirical distribution of each regression coefficient. That is, for each study population of birth year and level of study, we took 500 random samples with replacement from the study population in question, and performed the regression on each of these samples. We captured the regression output for each of these 500 analyses, and then, for each of the regression coefficients, we estimated the mean and standard deviation, and we ranked each of the 500 results. For our study, this process took about 72 hours to run.[6]

With 500 values in the distribution of each regression coefficient, we do not have exactly the percentiles that correspond to the usually reported 99.9, 99 and 95 per cent confidence limits. Indeed, with just 500 samples we cannot obtain the 99.9 per cent values as we would have had to use 2,000 random samples, but this would have quadrupled the time taken to run the bootstrapping procedure. We could have reported confidence limits for those values for which we had exact point estimates, namely the 95.2, 99.2 and 99.6 per cent limits. But to be consistent with almost all other published statistical reporting of confidence limits, we used simple linear interpolation between the 0.4th and 0.6th, and 2.4th and 2.6th percentiles (and their corresponding upper percentiles) in our samples to obtain the 99 and 95 per cent confidence limits respectively for each coefficient, and we used the first (0.2th) and last (99.8th) percentiles out of the 500 to obtain 99.6 percent confidence limits.

Sampling with replacement means that any one random sample from a population may exclude some people, and include some others more than once. It is through this method that the regressions performed on the samples differ slightly from the original study population. If this method produces much the same values for the regression coefficients as the original study population, with little dispersion around the true value of each estimate of the regression coefficients across the 500 iterations of the regression, then the standard errors will be small, and the confidence intervals will be narrow. If on the other hand there is much variation in the study population, for example in the change in income on completing a qualification, then the coefficients calculated across the 500 random samples of the bootstrap process will show greater variation as some individuals are left out of any on particular random sample, and others are included more than once. The average of these samples will still be a good estimate of the actual coefficient, but the standard error will be larger, and the confidence intervals will be wider, reflecting the wider empirical distribution of the calculated regression coefficients.

[6] The analysis at StatisticsNZ was performed on a 4-core computer with the CPU running at approximately 3 Gigahertz, with memory of 3.5Gigabytes. In other words, this was a reasonably fast computer. Most of the bootstrap processing was done over a weekend, so there would be few or no other users of the machine.

One of the considerations in bootstrapping regression models is whether to randomly sample the regression residuals,[7] or randomly sample individuals in the study population. Consensus appears to be that for social analysis, where there is little or no control over the explanatory variables, sampling individuals is more appropriate (Mooney and Duval 1993).

A further consideration in longitudinal data is what to randomly sample; individuals at a single point in time, or all the time points for a particular individual. In our case, since we are interested in *individual changes in income across time*, the natural unit of sampling is an individual with all of their annual incomes for the 10 years of data. Sampling this way preserves the dependence structure within individual time series (Wei, Pere, Koenker and He 2006).

5.7 Reciprocal causation

One of the problems with time-varying predictors in individual growth models, and indeed in many other regression analyses, is that of reciprocal causation, or endogeneity (Singer and Willett 2003). It can be likened to the familiar 'chicken and egg' problem; if X is associated with Y, can we conclude that X *causes* Y, or is it possible that Y *causes* X? We present a short discussion of this topic as a guide to readers who might wish to infer causal relationships between income and the other factors in our models.

In our study, we are not actually trying to determine what factors cause changes in income. Certainly, our study indicates which factors are associated with change in income, but we use those factors to explore differences in income trajectories, and compare income levels with the student loan repayment threshold. It is clear from our results that significant changes in income often occur after completing a qualification. It is this temporal delay—the income changes *after* completing—which gives us some confidence that there is a cause and effect relationship here, with successful education outcomes resulting in better personal income. But this is not the point of our study. In some respects it does not matter what causes someone's income to change, as long as it rises high enough after taking out a student loan for them to start to repay their loan. Taking out a student loan is commensurate with tertiary study, and for many people the end of that study results in them gaining a qualification, which coincides with the start of their repayment obligations.

However, there are many other factors which affect income which we haven't, and mostly cannot, include in our models. These include factors such as skill, motivation, prior experience, prior qualifications, and such like. A person's income will be a function both of their education, and these other factors. More motivated people are more likely to be better educated, and may be more likely to attempt qualifications at higher levels. The higher return to completing a bachelors degree is likely to be a function of completing the qualification *and* because of these other factors. So we would urge caution for those readers who would use our study to postulate cause and effect relationships between the factors in our models and income. And because these results are averages, they will not apply to *anyone* who completes a qualification at a particular level of study.

The influence of being on a benefit on income is also not clear, particularly because of the way in which we have modelled benefit status. In our study, we have not differentiated between the types of benefit, or on the duration of time someone is in receipt of a benefit in any one year (we discuss the reasons for this in Section 4 in this report). Being on a benefit clearly does affect income, being a proxy for engagement with the labour market, but it might be that, say, someone losing their job, which changes their income, makes it necessary for them to apply for

[7] The method works as follows. A random sample is taken from the original study population. A regression is performed, and the residuals are calculated. A resample of these residuals is then drawn randomly with replacement. Then these randomised residuals are added to the original sample of fitted response variables. These bootstrapped responses are then regressed again, to give estimates of the regression coefficients.

the unemployment benefit. The benefit status changes because income changed, mediated through the loss of employment. Again, we stress we are not making conclusions about the factors that affect someone's income. In our study, the direction of causation is immaterial; what is important is that people's income trajectory is different if they *are* in receipt of a benefit, and this affects their obligations to start repaying their student loan. But it would be wrong to conclude that being on a benefit causes lower incomes. More likely, the factors that result in lower levels of engagement with the labour market affect someone's income, and results in welfare dependence for a period of time.

5.8 Unexplained variation in the regression models

As we mentioned in Section 5.4 in this report, using random effects in our multi-level regression provides estimates of what are known as variance components. These can be used to calculate the extent of the correlations between the random effects, and to gauge how well the model parameters 'explain' individual income trajectories over time. In the original report we presented a table showing the correlation coefficients between the random effects (Section 3, Table 2). In this section here we will discuss unexplained variation in the models.

Estimated variance and co-variance components are somewhat more difficult to interpret as their numeric values have little absolute meaning, and there are no graphical aides to assist in their interpretation (Singer and Willet 2003).

Table 6 in the Appendix in this report shows the breakdown of the unexplained variation in income in the regression models. This variation is what is left after we control for age, gender, enrolment in study, level of study, time, benefit status, student loan status and completing a qualification. Firstly, it can be seen that the total variance unexplained is generally smaller for lower-level qualifications, and the unexplained variance is highest for those in their forties. This is probably because there is a wider range of step increases in income for higher-level qualifications, and incomes become more diverse as people get older. After a certain age however, incomes start to converge again. This might come about because incomes reach maximum levels for occupations or positions, people retire and so their income is determined by superannuation, or as incomes become limited by sickness and disability.

In addition, it can be seen that most unexplained variation occurs in the step change in income on completing a qualification, and in general this component comprises a larger share of the total variance for those completing higher-level qualifications. The smallest unexplained variation is in the rates of increase in income, both before studying and after completing a qualification. Finally, we can see there is about equal proportions of unexplained variation in income before studying, and in the changes in income for any one person through time.

5.9 Parsimonious models

Typically, a researcher will use the most parsimonious model for their analysis. A parsimonious model is one which uses the minimum number of factors to explain the maximum amount of variation in the data. As can be seen in Table 2 in the Appendix, the coefficient for the FEMALE x STUDIED x BENEFIT interaction for the pre-study rate of change in income is not significantly different from zero for young people with low incomes who studied for a certificate at levels 1 to 3. In a parsimonious model this term would be omitted and a new model refitted to the data. However, because we do separate analyses for each of the age group categories for each of the qualification levels, it makes sense to use the same model specification for each of these analyses. While this particular coefficient here is not significant for this group, this term is significantly different from zero for other age groups or qualification levels. We have omitted those higher-order interactions that weren't significant across all models, which we described in Section 5.4.

5.10 Systematic differences in people who study at different levels

One of the problems with studies of this type is that people are not randomly assigned to the different levels of study. That is to say, we didn't select a group of people, and then randomly assigned some of them to study certificates and levels 1 to 3, and some to diplomas, and some to bachelors degrees. Had we been able to do this, differences in graduate incomes could be attributed to the different levels of qualifications completed, all else being equal. This problem is not unique to this study; all analyses using historical administrative data have to deal with this problem.

Instead, we consider people who have elected to study at a particular level, and model their income trajectories. There is a risk that the people who were motivated to study at certificate level are fundamentally different to those who were motivated to study at bachelors level. This difference might be in their level of motivation, because of the longer time it takes to complete a higher-level qualification, and perhaps in the confidence they have in their academic ability, since it is generally regarded that study at higher levels is more difficult. A person choosing lower-level study may also not have the resources (time and energy, even if money is borrowed) to commit to study at higher levels. Or a particular qualification might not be offered where a person lives. So while there are greater returns to those who complete higher-level qualifications, study at these levels may not be a choice that everyone can make. So we need to be mindful of recommending study at particular levels to try and increase the chances that people have post-study incomes that oblige them to start to repay their student loans.

There are statistical techniques which ostensibly can overcome these types of problem. They are known as fixed effects regression models (Allison 2005).[8] These models are able to control for all possible characteristics of individuals in a study—even without measuring them—as long as they do not change over time. We did experiment with these models, since their ability to control for all individual characteristic makes them intuitively appealing. However, we didn't use them in the final study because these models don't produce an intercept term. We won't go into the mathematical reasons why, but because these models only look at within-person variation, and not between-person variation, the regression coefficients calculated by these models tell us how much change there is, on average, for completing a particular level of qualification, for example, but the models can't be used to determine what the *total* income is after competing. So it makes it difficult to determine if the total income has reached a particular threshold, which was the main point of our study.

Fixed effects regression models are not without their problems (Nickel 1981, Clark et al 2010). In our study, the between person variation was between 64 and 84 per cent, depending on the model. This variation is ignored in fixed effects models. In addition, if the unobserved factors we are trying to implicitly control for *do* vary with time, then a fixed effects regression model also won't give us the answers we want. For example, it might be that motivation and confidence do vary with time. In fact, someone's confidence might even increase if they successfully complete a tertiary qualification, which might affect their chances of gaining employment, or gaining higher paid employment.

5.11 Producing the graphs of income trajectories

The output from the individual growth models are estimates of the values of the coefficients in the level-2 model equation—the gamma terms ($\gamma_{..}$) in our model equations in Section 5.4. In our

[8] Random effects models also include factors which are termed 'fixed effects'. This is unfortunately confusing terminology.

case, these estimates are dollar values (in thousands of dollars). They are provided in Tables 2 to 5 in the Appendix. Apart from the estimate of the initial value, the other coefficients are the marginal values, or net effect, of the particular factor under consideration. For example, for people in their twenties with low incomes, who completed a certificate at levels 1 to 3 (Table 2 in the Appendix), the estimate of the annual income for a man, not on a benefit, with no student loan, who is not studying and has not completed a qualification, is \$3,040 (γ_{00}). The marginal value of completing a qualification, net of all other factors, is \$6,340 (γ_{10}). It is the amount he can expect his income to *increase*; it is not his total income at the time of completing.

In our report we have not focussed on the marginal incomes associated with the particular factors in our models, although they are of interest. What we did in our report was to calculate the total estimated incomes across time for prototypical persons (Singer and Willett 2003). Prototypical persons are those with particular combinations of characteristics, for which we plot results. We will illustrate this with an example.

Consider Figure 1 in our original report. It shows the income trajectories for prototypical people with low incomes in their twenties. Again let's consider those who may have studied for certificates at levels 1 to 3.

The regression output is shown in Table 2 in the Appendix. Each coefficient in the model equation can be replaced with the estimate of the coefficient, to produce the final model equation. Note that we have combined both level-1 and level-2 models to produce this final model equation. For this example, the (partial) model equation is:

$$
\begin{aligned}
&(\text{INCOME}_{ij} \div 1000) \\
&\quad = 3.04 - (0.16 \times \text{FEMALE}_i) + (4.04 \times \text{ON BENEFIT}_{\text{ij}}) + \left(3.37 \times (\text{FEMALE}_i \times \text{ON BENEFIT}_{ij})\right) - (0.34 \times \text{ENROLLED}_{\text{ij}}) \\
&\quad + \left(0.27 \times (\text{FEMALE}_i \times \text{ENROLLED}_{ij})\right) + (6.35 \times \text{HAS LOAN}_{ij}) - \left(0.62 \times (\text{FEMALE}_i \times \text{HAS LOAN}_{ij})\right) \\
&\quad - \left(5.62 \times (\text{ON BENEFIT}_{ij} \times \text{HAS LOAN}_{ij})\right) - \left(0.29 \times (\text{FEMALE}_i \times \text{ON BENEFIT}_{ij} \times \text{HAS LOAN}_{ij})\right) + \cdots
\end{aligned}
$$

Each factor can take on the values of 0 or 1. In the study population, any one individual *i* will be either male or female (0 or 1 for the FEMALE factor) for all time periods, but the other factors can vary with time *j*.

To facilitate the calculation of annual income values, we constructed tables in a spreadsheet, in which we defined our prototypical persons and their characteristics. For example, for men in their twenties, for those not on a benefit, with a student loan, who completed a qualification in 2003, the prototypical definition is:

Year	Year-1999	Studied	Completed	Post-completion year	Has loan	On benefit	Female	Estimated annual income (\$ 000)
1999	0	0	0	0	1	0	0	9.39
2000	1	0	0	0	1	0	0	11.06
2001	2	0	0	0	1	0	0	12.59
2002	3	0	0	0	1	0	0	13.97
2003	4	1	1	0	1	0	0	21.98
2004	5	1	1	1	1	0	0	23.90
2005	6	1	1	2	1	0	0	25.48
2006	7	1	1	3	1	0	0	26.72
2007	8	1	1	4	1	0	0	27.61
2008	9	1	1	5	1	0	0	28.16

This prototypical man's income is plotted as the solid black line in Figure 1 in the original report. It shows the income for a young man with low income over four years, prior to any tertiary study, and then shows how that income changes once he has completed a certificate at levels 1 to 3. He is never in receipt of any benefit income.

We have set the student loan status to true (1) in each year because we find that people in their twenties and thirties who *do* go on to take out a student loan have higher initial incomes prior to any study. For people in their forties (and fifties) there is no significant difference in pre-study incomes between those who do or do not go on to take out a student loan. If we don't adjust for this, the change in income on completing a qualification is the sum of the change due to completing, *and* the change up to the higher base income level of someone who goes on to take out a student loan. By plotting the pre-study income for those who go on to take out a student loan, the step increase in income on completing is then just that due to the completion event.

As can be seen, we ignore the period of enrolment (the coefficient is set to zero). We do this for the same reasons as described above. People's income usually drops while they study—they are either not at work if they study full-time, or they may work part-time, or reduce their hours of work. Once a person stops studying, completes a qualification and goes back to work, their income will increase both due to the fact they now work longer hours, *and* because they now have a higher level of qualification. We include ENROLLED in the regression model to capture the marginal dollar amount that is a function of being enrolled or not. But we can then ignore it in our prototypical trajectories, so the income trajectories only show the change in income due to the completion event.

For a prototypical *woman* with low income, not on a benefit and who took out a student loan, we would set the values in the Female column to 1 in each year, leaving all other values the same. In the model equations, all those terms with FEMALE would then be included in the equation, giving us the annual income for prototypical women with these sets of characteristics.

We will present one further example. The figures in the report show the income trajectories for people who have studied, but did not gain a qualification. The definition for a prototypical man on low income is:

Year	Year-1999	Studied	Completed	Post-completion year	Has loan	On benefit	Female	Estimated annual income ($ 000)
1999	0	0	0	0	1	0	0	9.39
2000	1	0	0	0	1	0	0	11.06
2001	2	0	0	0	1	0	0	12.59
2002	3	0	0	0	1	0	0	13.97
2003	4	1	0	0	1	0	0	17.10
2004	5	1	0	0	1	0	0	18.66
2005	6	1	0	0	1	0	0	20.07
2006	7	1	0	0	1	0	0	21.34
2007	8	1	0	0	1	0	0	22.46
2008	9	1	0	0	1	0	0	23.43

As can be seen, the factors indicating this person had studied are still set after the end of the study enrolment, and they also have a student loan, but the factor for completion, and years after completing, remain at zero. The annual incomes are the same as the earlier example up to year 3, but then differ because one prototypical person completed a certificate, and the other did not. The non-completion income trajectory is shown in Figure 1 in the report as a black dashed line.

6 ASSOCIATION BETWEEN STUDENTS AND THE SUBJECTS OF THE QUALIFICATIONS THEY COMPLETED

6.1 What was the statistical technique

In the report we consider whether there were systematic patterns in the subjects of the qualifications completed by the study populations. We did this to get a better understanding of the quite large differences in income between men and women, particularly those in their twenties, after completing qualifications at the same level. We used a multiple correspondence analysis to do this.

A technical description of multiple correspondence analysis can be found in Hoffman and Franke (1986) and a less technical description is given by Bendixen (2003). We will illustrate the steps of this analysis for a single study population, as the technique is not widely used in education research in New Zealand.

6.2 Where are the results reported

The results of this analysis are reported on page 16 in Section 3 of the original report.

6.3 The study populations used

For this analysis, we considered only people who had *completed* a qualification. However, to increase sample sizes, we relaxed the requirement that a person did not study at a higher level than the one they completed, as we did for the individual growth modelling.

The study populations were therefore all those people who had completed a qualification during the years 1999 to 2008, who earned less than the student loan repayment threshold in 1999, who did or did not have a student loan. They may also have been enrolled at levels of study other than the level they completed. Only the highest level of qualification completed was used in the analysis.

Separate analyses were performed for each of the three birth year cohorts, for each of the study levels, certificates at levels 1 to 3, certificates at level 4, diplomas and bachelors degrees.

We only included fields of study where there were a specified minimum number of people completing. We did this to reduce the number of fields of study in the analysis to make it easier to view the graphical output, especially when we were dealing with narrow fields of study, but it also meant we excluded those fields of study which relatively few students completed. For the broad fields of study, we only included those with at least 50 students. For the narrow fields of study the following rules were used:

- certificates at levels 1 to 3 with 200 or more students completing
- certificates at level 4 and diplomas with 75 or more students; and
- bachelors degrees with 100 or more students.

The explanatory variables included were gender, birth year and benefit status. Since we are not dealing with longitudinal data here, we had to define 'being on a benefit' differently from the individual growth modelling, where the benefit status could vary from year to year. For the multiple correspondence analysis we regarded someone as being on a benefit if they had received benefit income for 5 or more years over the 10 year study period.

6.4 Description of the method

Multiple correspondence analysis is used to find a low-dimensional graphical representation of the rows and columns of a multi-way contingency table. The contingency table consists of all the variables in the analysis crossed with themselves. An example will probably best illustrate this concept.

This contingency table is known as a Burt Table in correspondence analysis. It is a two-way table showing the frequencies of every possible combination of variables in the input data set. In the example below, for broad fields of study, for people who completed certificates at levels 1 to 3, we can see there were 8,529 people in their twenties who completed one of these qualifications during the study period. Of these 8,529 graduates, 5,061 were female, and 444 completed a certificate in Agriculture, environmental and related studies. We can also see that those in their twenties cannot also be in their thirties or forties, and that if they completed a certificate in Agriculture, they could not also complete a certificate in Engineering and related technologies (since we are only counting a single qualification completion for each person). This example table does not show all of the fields of study or other factors used in the actual analysis.

	20-29 years of age	30-39 years of age	40-49 years of age	Female	Male	Agriculture, environmental and related studies	Engineering and related technologies
20-29 years of age	8,529	0	0	5,061	3,471	444	711
30-39 years of age	0	5,109	0	3,519	1,593	333	309
40-49 years of age	0	0	3,993	2,712	1,284	339	222
Female	5,061	3,519	2,715	11,292	0	354	240
Male	3,468	1,590	1,281	0	6,342	762	1,002
Agriculture, environmental and related studies	444	333	342	354	762	1,116	0
Engineering and related technologies	711	309	222	240	1,002	0	1,242

The aim of multiple correspondence analysis is to find dimensions against which the variables in the Burt Table can be plotted, but fewer than the number of input variables. The strength of the association with any one dimension is indicated by a variable called Inertia. The next table shows the inertia values for this example. We used the Greenacre adjustment to the inertia values to account for the fact that the unadjusted inertia values provide an overly pessimistic indicate of fit (SAS 2003). Inertia is analogous to variance in principal component analysis.

We see from the output that there are two main dimensions (rows), which account for about 67 per cent of the variation in the data. Note also that only 71 per cent of the total variation is able to be explained.

Principal inertia	Adjusted inertia	Per cent	Cumulative per cent
0.36623	0.02402	55.11	55.11
0.30464	0.00531	12.18	67.28
0.28024	0.00163	3.73	71.01
0.26047	0.00019	0.45	71.46
Total	0.03115	71.46	

The next table shows the dimension coordinates for each of the variables in this example.

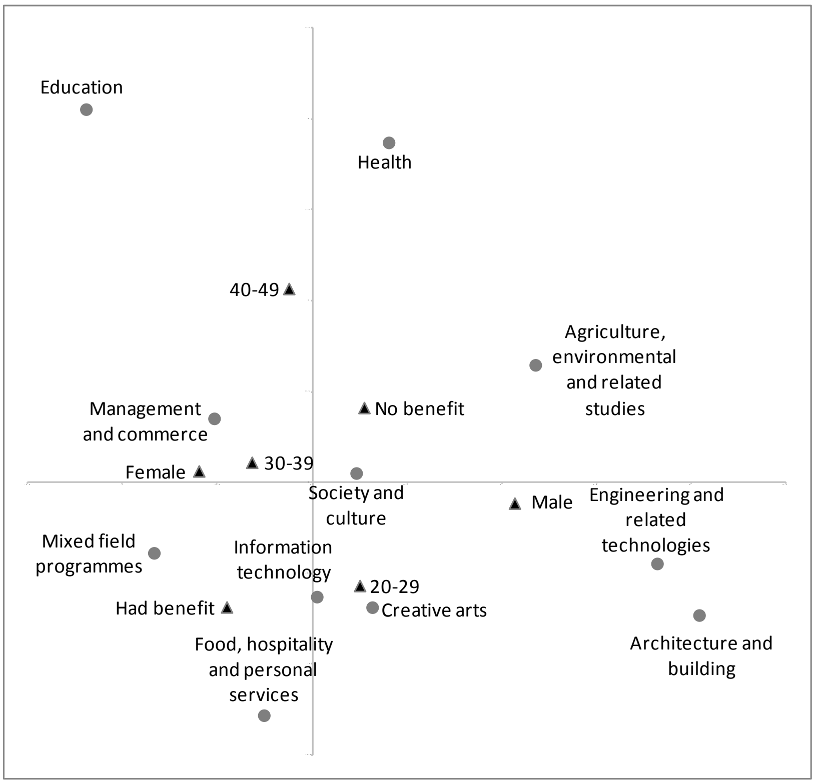
Variable	Dimension 1	Dimension 2
20-29	0.25	-0.57
30-39	-0.32	0.11
40-49	-0.12	1.07
Female	-0.60	0.06
Male	1.07	-0.11
Had benefit	-0.45	-0.68
No benefit	0.27	0.41
Agriculture, environmental and related studies	1.18	0.65
Architecture and building	2.04	-0.73
Creative arts	0.32	-0.69
Education	-1.19	2.05
Engineering and related technologies	1.82	-0.45
Food, hospitality and personal services	-0.26	-1.28
Health	0.40	1.87
Information technology	0.02	-0.63
Management and commerce	-0.52	0.35
Mixed field programmes	-0.84	-0.39
Society and culture	0.23	0.05

We can then plot these coordinates in two dimensional space. The first figure below shows the results when using broad fields of study, while the second shows the results when using narrow fields of study. Both tell the same story about the association between gender, age and the subject of the qualification completed, although using the broad field results in less clutter and therefore presents a clearer picture of the associations.

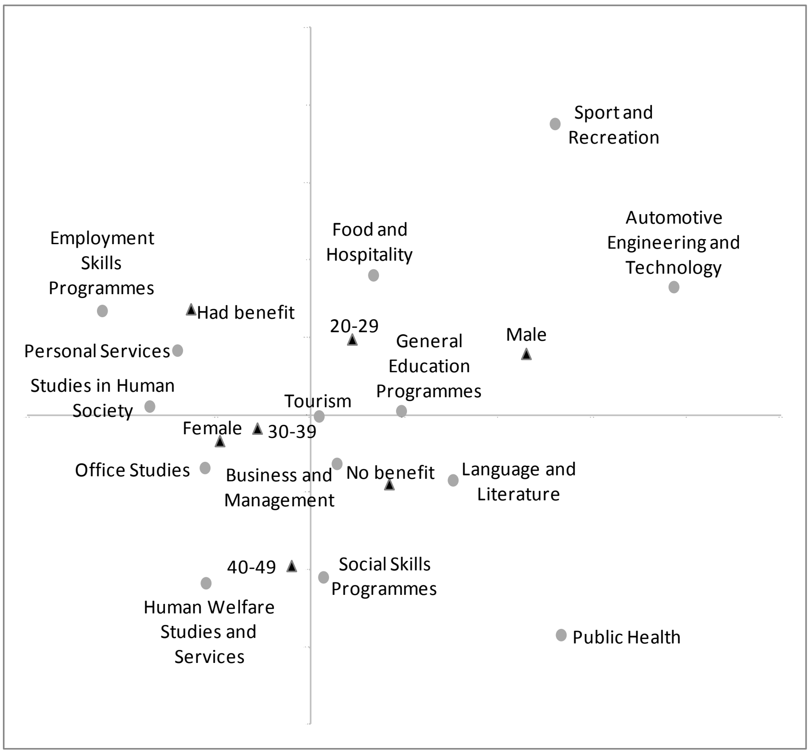
In summary, dimension 1 seems to be related to gender, with males on the right and females on the left. Dimension 2 appears to be related to age, with younger people higher on the axis, and older people lower. Benefit status also appears to be related to dimension 1, suggesting that, overall, women are more strongly associated with being on a benefit than men, and is somewhat more strongly associated with younger women than with older women, at least for this population of people who completed certificates at levels 1 to 3.

The grouping of the subjects along the horizontal axis is clearly evident, even when considering narrow fields of study. This indicates there is a clear association between gender and qualification subject, as described in the report. For lower-level qualifications, there is also an association with age.

Graphical representation of multiple correspondence analysis for broad fields of study for people completing a certificate at levels 1 to 3

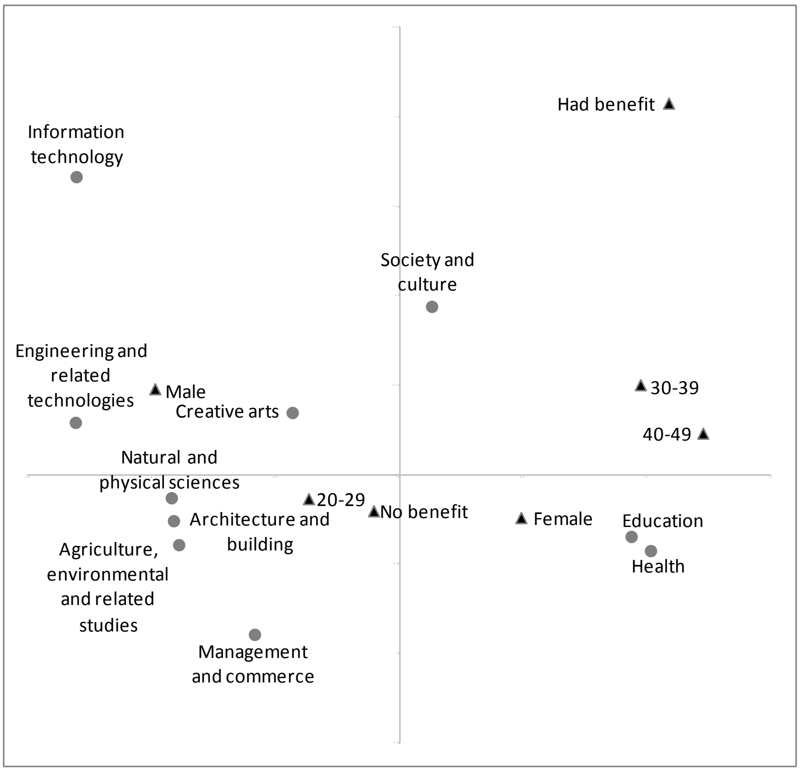


Graphical representation of multiple correspondence analysis for narrow fields of study for people completing a certificate at levels 1 to 3



The next figure shows the same results but for people who completed a bachelors. These results are for broad fields of study.

Graphical representation of multiple correspondence analysis for broad fields of study for people completing a bachelors degree



Dimension 1 in this case also seems to be related to gender, but unlike the results for people completing certificates at levels 1 to 3, these results show no other clear pattern for age or benefit status along either of the dimensions. This is supported by the results, which show that dimension 1 accounts for 76 per cent of the variance in the data, with dimension 2 accounting for 4 per cent. There appears to be a separation of qualification subject by gender, although the subjects themselves are different for males and females compared to certificates at levels 1 and 3. There appears to be little association between being on a benefit and studying for a bachelors degree. Overall, there is less association between gender and age and the subject of the bachelors degree completed, when compared to the associations seen for people completing certificates at levels 1 to 3.

The Appendix includes the dimension coordinates for each of the multiple correspondence analyses we discussed in the original report (Tables 11 and 12 in the Appendix).

7 POST-COMPLETION INCOME DISTRIBUTIONS

7.1 What was the statistical technique

For our analysis of how income changes on completing a qualification for people in different income quartiles, we used a technique called quantile regression. The results of quantile regressions are incomes at specific percentiles, such as the median, for a particular sub-group of the study population.

The purpose of doing this was to determine the graduate income distribution for different groups completing different levels of qualification. Would everyone in the post-completion income distribution be above the repayment threshold, or only the higher income groups?

A gentle introduction to quantile regression can be found in Cade and Noon (2003).

7.2 Where are the results reported

The results of this analysis are reported in Section 4 of the original report.

7.3 The study populations used

For this analysis we used graduates for whom we had an income measure one year before they were enrolled in tertiary study, and another one year after they graduated. This meant we were able to compare their pre-study income with their income after completing, controlling for the level of pre-study income and other demographic factors.

7.4 Description of the method

Whereas ordinary least squares regression results in estimates of the conditional *mean* of the response variable given certain values of predictor variables, quantile regression aims to estimate either the conditional *median* or other percentiles of the response variable (Cade and Noon 2003). And unlike ordinary regression, which assumes homogeneity of variances and normally distributed residuals, quantile regression makes no distributional assumptions, and is robust against outliers and unequal variances.

While a quantile regression can calculate estimated income across the entire span of percentiles, we have chosen to use the 25th, 50th and 75th percentiles. If more percentiles are included in the modelling, the standard errors of the regression coefficients are larger than when using just a few.

Confidence intervals were estimated using re-sampling, which uses bootstrapping to arrive at the probability distributions on which the intervals are based. This is done by the statistical software package.

We proceeded as follows. We modelled graduate income, in 2008 dollars, against the following factors and their interactions: whether the income was for the pre-study or post-completion time period, gender, whether the person was in receipt of benefit income before studying, and whether the person ever had a student loan during this period. All interactions were included, including the fourth-order interaction between all four factors. Models were run for each combination of birth year cohort and level of qualification completed. Pre-study income was included as a continuous variable.

The interaction terms allowed us to determine the combined effect of being female, being on a benefit, and having taken out a loan.

The results of the regression are presented in the original report in Tables 3 to 6.

The following table is the output from the quantile regression for people in their twenties who completed a certificate at levels 1 to 3, conditional on the post-compeltion income being at the 25th percentile. The estimates and confidence limits are in thousands of dollars.

	Parameter Estimates						
Parameter	DF	Estimate	Standard Error	90% Confidence Limits		*t* value	Prob
Intercept	1	19.01	1.06	17.26	20.75	17.9	<.0001
completed	1	2.36	1.67	-0.38	5.11	1.4	0.1562
female	1	-16.19	1.37	-18.44	-13.95	-11.9	<.0001
completed*female	1	-3.00	2.13	-6.51	0.51	-1.4	0.16
benefit	1	-8.40	1.22	-10.40	-6.40	-6.9	<.0001
completed*on benefit	1	1.46	1.84	-1.56	4.49	0.8	0.4265
female*on benefit	1	19.70	1.50	17.23	22.17	13.1	<.0001
completed*female*on benefit	1	-0.51	2.17	-4.09	3.07	-0.2	0.8143
had loan	1	-1.17	1.38	-3.43	1.10	-0.9	0.3974
completed*had loan	1	0.95	2.16	-2.60	4.50	0.4	0.6604
female*had loan	1	7.17	1.88	4.08	10.25	3.8	0.0001
completed*female*had loan	1	2.66	2.97	-2.23	7.55	0.9	0.3708
on benefit*had loan	1	-0.15	1.50	-2.61	2.32	-0.1	0.9213
completed*on benefit*had loan	1	-1.70	2.31	-5.50	2.09	-0.7	0.4603
female*on benefit*had loan	1	-7.86	1.95	-11.07	-4.64	-4.0	<.0001
completed*female*on benefit*had loan	1	0.32	3.10	-4.78	5.41	0.1	0.9184

The intercept value, $19,010, is the pre-study income for those at the 25th percentile of post-completion income for men in their twenties not on a benefit and who didn't have a student loan. For these men, their income was $2,360 higher one year after completing a certificate at levels 1 to 3. The analysis indicates that this amount is not significantly different from zero (p=0.1562), so we conclude the income for this group isn't significantly higher after completing one of these qualifications, at least in the first year after they completed.

For women in their twenties who were not on a benefit and who did not have a student loan, whose graduate income was at the 25th percentile, the pre-study income is $2,810. This is arrived at by adding the coefficient for women (-16.19) to the intercept (the total is calculated on the unrounded coefficients). The output tells us that the difference in income between these men and women is significant. However, we are interested in whether a woman who completes one of these qualifications sees a significant change in her income, not whether men and women have different incomes. The income after completing for a woman in her twenties who is not on a benefit and who didn't have a student loan is $2,180. This is arrived at by adding the coefficients for the intercept, the factors COMPLETED and FEMALE and the interaction term COMPLETED* FEMALE. It is clearly not significantly higher because it is smaller! But we cannot determine from the default regression output whether this is significantly different from zero.

To calculate the confidence limits for each of the differences in income between the pre-study and post-completion incomes, we used a pooled estimate of the standard error to calculate statistical significance.[9] For example, the table below shows data from the regression. We specified that estimates of income be reported with their standard errors for every combination of gender, loan status and benefit status. We then calculated the difference between the pre-study and post-completion incomes for each of these combinations, and calculated a *t*-value as the quotient of the difference in income divided by the pooled standard errors making up the difference. The significance indicators are determined from critical values of Student's *t*-distribution. If the difference is significantly different from zero, we can report that the post-completion income is significantly higher (or at least different) than the pre-study income. The following table shows an example output from these calculations, for graduates in their twenties who completed a certificate at levels 1 to 3, for people with graduates incomes at the 25th percentile with the particular combination of factors.

			Pre-study predicted income	Standard error	Post-completion predicted income	Standard error	Difference in incomes	*t* value	
Male	No loan	No benefit	19.01	0.95	21.37	1.49	2.36	1.34	ns
		On benefit	10.61	0.42	14.43	0.48	3.83	5.99	***
	Had loan	No benefit	17.84	0.83	21.15	0.97	3.31	2.61	**
		On benefit	9.29	0.18	12.37	0.45	3.07	6.31	***
Female	No loan	No benefit	2.81	0.91	2.18	1.21	-0.63	-0.42	ns
		On benefit	14.12	0.37	14.43	0.31	0.32	0.67	ns
	Had loan	No benefit	8.81	0.85	11.79	1.15	2.98	2.08	*
		On benefit	12.11	0.34	14.65	0.25	2.54	6.06	***

For people in their twenties who completed a certificate at levels 1 to 3, with post-completion incomes at the 25th percentile.
The probability that a difference in incomes is greater than zero is indicated by: * p<0.05, ** p<0.01; *** p<0.001; ns – not significantly different.

7.5 Assumptions

As we mentioned, quantile regression makes no distributional assumptions about residuals, and is robust to heterogeneous variances and outliers.

One useful feature of quantile regression is that it is invariant to monotonic transformations, like the logarithmic transformation, meaning that the same results are produced using the transformed variable and un-transforming it after the regression analysis as using the untransformed variable in the first place (Cade and Noon 2003). This meant we could again use untransformed income values in the regression, which facilitates ease of interpretation and simplifies the comparison to the student loan repayment threshold.

[9] The pooled standard error is defined as $\mathrm{se}_{\mathrm{pooled}} = \sqrt{(\mathrm{se}_1^2 + \mathrm{se}_2^2)}$.

8 REFERENCES

Allison, P. (2005) *Fixed effects regression methods for longitudinal data using SAS®*, SAS Institute Inc: Carey, NC, USA.

Australian Bureau of Statistics (2010) *Migration, Australia, 2008-09 technical note: '12/16 month rule' methodology for calculating net overseas migration from September quarter 2006 onwards*, Australian Bureau of Statistics: Canberra.

Bendixen, M. (2003) *A practical guide to the use of correspondence analysis in market research*, Marketing Bulletin, 14, Technical note 2, 1-15.

Cade, B. and B. Noon (2003) *A gentle introduction to quantile regression for ecologists*, Frontiers in ecology and the environment, 1(8): 412–420.

Clark, P., C. Crawford, F. Steele and A. Vignoles (2010) *The choice between fixed and random effects models: some considerations for educational research*, Working paper No. 10/240, Centre for Market and Public Organisation: Bristol.

Engler, R. (2014) *Was it worth it? Do low-income New Zealand student loan borrowers increase their income after studying for a tertiary qualification?* Ministry of Education: Wellington.

Hayes, A. and L. Cai (2007) *Using heteroskedasticity-consistent standard error estimates in OLS regression: An introduction and software implementation*, Behaviour Research Methods, 39(4): 709–722.

Hoffman, D. and Franke, G. (1986), *Correspondence analysis: graphical representation of categorical data in marketing research*, Journal of Marketing Research, 23, 213–227.

Mahoney, P., Z. Park and R. Smyth (2013) *Moving on up. What young people earn after tertiary education*, Ministry of Education: Wellington.

Mooney, C. and R. Duval (1993) *Bootstrapping: a nonparametric approach to statistical inference*, Sage University Paper series on Quantitative Applications in the Social Sciences, Series no. 07-095, Sage: Newbury Park, CA.

Nickel, S. (1981) *Biases in dynamic models with fixed effects*, Econometrica, 29(6): 1417–1426.

Papadopoulos, T. (2012) *Who left, who returned, and who is still away? Migration patterns of 2003 graduates, 2004–2010*, Labour and Immigration Research Centre, Ministry of Business, Innovation and Employment: Wellington.

SAS (2013) *SAS/STAT 12.3 User's Guide, The Corresp procedure*, SAS Publishing: Cary, downloaded from http://support.sas.com/documentation/onlinedoc/stat/123/corresp.pdf.

Singer, J. and J. Willett (2003) *Applied longitudinal data analysis: modelling change and event occurrence*, Oxford: New York.

Sokal, R. and F. Rohlf (1981) *Biometry. The principles and practice of statistics in biological research*, Second Edition, W.H. Freeman and Company: San Francisco.

Wei, Y., A. Pere, R. Koenker and X. He (2006) *Quantile regression methods for reference growth charts*, Statistics in Medicine, 25(8): 1,369–1,382.

9 APPENDIX. DATA TABLES

This section contains various data tables not included in the original report. These are:

1. Example regression output linking the individual growth model's parameters with the regression coefficients.
2. Individual growth model regression output for people in their twenties.
3. Individual growth model regression output for people in their thirties.
4. Individual growth model regression output for people in their forties.
5. Individual growth model regression output for people in their fifties.
6. Distribution of unexplained variance in the individual growth models.
7. Sample sizes for individual growth models for people who may have completed a certificate at levels 1 to 3.
8. Sample sizes for individual growth models for people who may have completed a certificate at level 4.
9. Sample sizes for individual growth models for people who may have completed a diploma.
10. Sample sizes for individual growth models for people who may have completed a bachelors degree.
11. Multiple correspondence analysis results: dimension coordinates for people completing tertiary qualifications using broad fields of study.
12. Multiple correspondence analysis results: dimension coordinates for people completing tertiary qualifications using narrow fields of study.

Table 1. Example of regression output linking the individual growth model's parameters to the regression coefficients; people in their twenties who may or may not have studied for a certificate at levels 1 to 3

Fixed effect	Coefficient	Model parameter	Estimate	Standard error
Pre-study income	Intercept	γ_{00}	3.04	0.06
	ON BENEFIT	γ_{01}	4.04	0.11
	ENROLLED	γ_{02}	-0.34	0.11
	HAS LOAN	γ_{03}	6.35	0.31
	ON BENEFIT x HAS LOAN	γ_{04}	-5.62	0.29
	FEMALE	γ_{05}	-0.16	0.08
	FEMALE x ON BENEFIT	γ_{06}	3.37	0.13
	FEMALE x ENROLLED	γ_{07}	0.27	0.13
	FEMALE x HAS LOAN	γ_{08}	-0.62	0.38
	FEMALE x ON BENEFIT x HAS LOAN	γ_{09}	-0.29	0.37
Income on completion	COMPLETED	γ_{10}	6.34	0.57
	COMPLETED x ON BENEFIT	γ_{11}	-2.74	0.28
	COMPLETED x HAS LOAN	γ_{12}	-1.46	0.58
	COMPLETED x FEMALE	γ_{13}	-1.47	0.61
	COMPLETED x FEMALE x HAS LOAN	γ_{14}	0.91	0.68
Rate of change in income, pre-study	Intercept: linear term	γ_{15}	1.58	0.02
	Intercept: quadratic term	γ_{16}	-0.07	0.00
	STUDIED	γ_{17}	1.46	0.05
	ON BENEFIT	γ_{18}	-0.81	0.03
	HAS LOAN	γ_{19}	0.17	0.07
	STUDIED x ON BENEFIT	γ_{20}	-0.27	0.05
	STUDIED x HAS LOAN	γ_{21}	-0.99	0.07
	ON BENEFIT x HAS LOAN	γ_{22}	0.34	0.08
	STUDIED x ON BENEFIT x HAS LOAN	γ_{23}	0.37	0.08
	FEMALE	γ_{24}	-0.24	0.02
	FEMALE x STUDIED	γ_{25}	-0.73	0.06
	FEMALE x ON BENEFIT	γ_{26}	0.13	0.04
	FEMALE x HAS LOAN	γ_{27}	0.28	0.09
	FEMALE x STUDIED x ON BENEFIT	γ_{28}	0.09	0.07
	FEMALE x STUDIED x HAS LOAN	γ_{29}	-0.19	0.10
	FEMALE x ON BENEFIT x HAS LOAN	γ_{30}	-0.22	0.10
	FEMALE x STUDIED x ON BENEFIT x HAS LOAN	γ_{31}	0.25	0.11
Rate of change in income, post-completion	Intercept: linear term	γ_{32}	0.54	0.16
	Intercept: quadratic term	γ_{33}	-0.10	0.13
	ON BENEFIT	γ_{34}	-0.21	0.13
	HAS LOAN	γ_{35}	-0.09	0.13
	ON BENEFIT x HAS LOAN	γ_{36}	0.56	0.13
	FEMALE	γ_{37}	-0.10	0.16
	FEMALE x HAS LOAN	γ_{38}	-0.03	0.18
Variance components				
Level-1	Within person, ε_{ij}	σ_ε^2	39.45	0.31
Level-2, between people for:	Pre-study income, δ_{0i}	σ_0^2	35.20	0.56
	Rate of change in income before completing, δ_{1i}	σ_1^2	2.37	0.03
	Change in income on completing, δ_{2i}	σ_2^2	62.32	4.14
	Rate of change in income after completing, δ_{3i}	σ_3^2	8.18	0.52
	Co-variance between σ_0^2 and σ_1^2	σ_{01}	0.80	0.08
	Co-variance between σ_0^2 and σ_2^2	σ_{02}	-13.48	2.02
	Co-variance between σ_0^2 and σ_3^2	σ_{03}	-3.71	0.67
	Co-variance between σ_1^2 and σ_2^2	σ_{12}	-2.00	0.58
	Co-variance between σ_1^2 and σ_3^2	σ_{13}	-2.72	0.21
	Co-variance between σ_2^2 and σ_3^2	σ_{23}	3.07	0.96

The coefficients and model parameters are defined in Section 5.4 of this report.
Standard errors are derived using a bootstrap with 500 iterations. See section 5.6 in this report for details.
Source: Original data from Statistics NZ Integrated Data Infrastructure. Analysis by Ministry of Education.

Table 2. Regression results for the individual growth models for people in their **twenties**

Fixed effects	Coefficients	Certificates Levels 1-3		Certificates Level 4		Diplomas		Bachelors	
Pre-study income	Intercept	3.04	***	3.32	***	3.50	***	3.77	***
	ON BENEFIT	4.04	***	3.51	***	3.42	***	3.58	***
	ENROLLED	-0.34	***	0.01	ns	-2.67	***	-4.11	***
	HAS LOAN	6.35	***	5.71	***	5.63	***	5.00	***
	ON BENEFIT x HAS LOAN	-5.62	***	-5.40	***	-5.30	***	-4.61	***
	FEMALE	-0.16	ns	-0.27	***	-0.29	***	-0.09	ns
	FEMALE x ON BENEFIT	3.37	***	3.56	***	3.51	***	3.21	***
	FEMALE x ENROLLED	0.27	ns	-0.26	ns	1.77	***	1.18	***
	FEMALE x HAS LOAN	-0.62	ns	0.05	ns	0.38	ns	0.98	***
	FEMALE x ON BENEFIT x HAS LOAN	-0.29	ns	-0.59	ns	-0.99	**	-1.59	***
Income on completion	COMPLETED	6.34	***	4.07	***	8.20	***	10.12	***
	COMPLETED x ON BENEFIT	-2.74	***	-1.95	***	-2.31	***	-1.90	***
	COMPLETED x HAS LOAN	-1.46	***	0.75	ns	-3.69	***	-3.01	***
	COMPLETED x FEMALE	-0.24	**	0.49	ns	0.79	ns	6.61	***
	COMPLETED x FEMALE x HAS LOAN	0.91	ns	-1.48	ns	0.33	ns	-2.74	***
Rate of change in income, pre-study	Intercept: linear term	1.58	***	1.74	***	1.85	***	1.93	***
	Intercept: quadratic term	-0.07	***	-0.08	***	-0.08	***	-0.09	***
	STUDIED	1.46	***	1.62	***	1.66	***	1.93	***
	ON BENEFIT	-0.81	***	-0.76	***	-0.78	***	-0.81	***
	HAS LOAN	0.17	*	-0.01	ns	-0.07	ns	0.06	ns
	STUDIED x ON BENEFIT	-0.27	***	-0.50	***	-0.34	***	-0.43	***
	STUDIED x HAS LOAN	-0.99	***	-0.99	***	-0.63	***	-0.83	***
	ON BENEFIT x HAS LOAN	0.34	***	0.41	***	0.43	***	0.28	***
	STUDIED x ON BENEFIT x HAS LOAN	0.37	***	0.51	***	0.27	*	0.34	***
	FEMALE	-0.24	***	-0.29	***	-0.32	***	-0.35	***
	FEMALE x STUDIED	-0.73	***	-0.73	***	-0.40	***	-0.18	ns
	FEMALE x ON BENEFIT	0.13	**	0.12	***	0.14	***	0.19	***
	FEMALE x HAS LOAN	0.28	***	0.03	ns	-0.04	ns	-0.19	***
	FEMALE x STUDIED x ON BENEFIT	0.09	ns	0.28	*	-0.04	ns	-0.30	*
	FEMALE x STUDIED x HAS LOAN	-0.19	ns	0.16	ns	-0.11	ns	0.09	ns
	FEMALE x ON BENEFIT x HAS LOAN	0.25	*	-0.13	ns	-0.05	ns	0.09	ns
	FEMALE x STUDIED x ON BENEFIT x HAS LOAN	0.25	*	-0.12	ns	0.05	ns	0.21	ns
Rate of change in income, post-completion	Intercept: linear term	0.54	***	1.10	**	1.04	***	1.66	***
	Intercept: quadratic term	-0.10	***	-0.13	***	-0.19	***	-0.22	***
	ON BENEFIT	-0.21	ns	-0.31	ns	-0.15	ns	0.27	ns
	HAS LOAN	-0.09	ns	-0.14	ns	0.68	**	0.61	*
	ON BENEFIT x HAS LOAN	0.56	***	0.31	ns	0.19	ns	-0.50	ns
	FEMALE	-0.10	ns	-1.06	***	-0.64	ns	-1.60	***
	FEMALE x HAS LOAN	-0.03	ns	0.36	ns	0.19	ns	0.48	ns
Variance components									
Within person		39.45	***	40.06	***	41.24	***	45.14	***
Pre-study income between persons		35.20	***	38.04	***	39.29	***	38.99	***
Rate of change before studying		2.37	***	2.47	***	2.60	***	2.81	***
Change in income on completing		62.32	***	67.99	***	89.80	***	178.25	***
Rate of change after completing		8.18	***	12.51	***	11.55	***	21.49	***
Co-variance components									
Pre-study income and rate of change before studying		0.80	***	1.17	***	1.33	***	1.27	***
Pre-study income and change in income on completing		-13.48	***	-14.32	***	-9.60	***	-4.51	ns
Pre-study income and rate of change after completing		-3.71	***	-3.48	**	-2.80	**	-1.16	ns
Rate of change in income before study and change on completing		-2.00	***	-2.46	*	0.05	ns	-0.16	ns
Rate of change in income before study and after completing		-2.72	***	-3.33	***	-2.57	***	-3.04	***
Change on completing and rate of change after completing		3.07	**	2.72	ns	-2.33	ns	-11.80	***
Sample size		61,041		61,197		64,713		72,954	

The probability (p) that a given coefficient is zero is indicated by: *** p<0.004; ** p<0.01; * p<0.05; ns – not significantly different from zero. Confidence indicators are based on bootstrapped standard errors and confidence intervals.

Source: Original data from Statistics NZ Integrated Data Infrastructure. Analysis by Ministry of Education.

Table 3. Regression results for the individual growth models for people in their **thirties**

Fixed effects	Coefficients	Certificates Levels 1-3	Certificates Level 4	Diplomas	Bachelors
Pre-study income	Intercept	5.60 ***	5.63 ***	5.68 ***	5.72 ***
	ON BENEFIT	4.71 ***	4.64 ***	4.58 ***	4.58 ***
	ENROLLED	1.10 ***	1.62 ***	0.74 *	-1.81 ***
	HAS LOAN	5.83 ***	5.59 ***	5.90 ***	5.71 ***
	ON BENEFIT x HAS LOAN	-5.24 ***	-4.90 ***	-4.96 ***	-4.68 ***
	FEMALE	-0.72 ***	-0.74 ***	-0.75 ***	-0.74 ***
	FEMALE x ON BENEFIT	3.57 ***	3.53 ***	3.53 ***	3.49 ***
	FEMALE x ENROLLED	-0.98 ***	-0.98 ***	-0.52 ns	0.56 ns
	FEMALE x HAS LOAN	-3.67 ***	-3.38 ***	-3.77 ***	-3.87 ***
	FEMALE x ON BENEFIT x HAS LOAN	2.53 ***	2.21 ***	2.37 ***	2.47 ***
Income on completion	COMPLETED	3.05 ***	2.91 *	6.05 ***	13.92 ***
	COMPLETED x ON BENEFIT	-1.77 ***	-1.62 **	-2.33 ***	-4.89 ***
	COMPLETED x HAS LOAN	0.62 ns	0.86 ns	-1.39 ns	-4.61 ns
	COMPLETED x FEMALE	-0.97 ns	-0.76 ns	1.11 ns	-0.10 ns
	COMPLETED x FEMALE x HAS LOAN	0.18 ns	-0.65 ns	-0.17 ns	5.82 *
Rate of change in income, pre-study	Intercept: linear term	1.96 ***	2.10 ***	2.17 ***	2.22 ***
	Intercept: quadratic term	-0.08 ***	-0.08 ***	-0.08 ***	-0.08 ***
	STUDIED	1.23 ***	1.34 ***	1.55 ***	0.89 ***
	ON BENEFIT	-0.93 ***	-0.98 ***	-0.99 ***	-1.01 ***
	HAS LOAN	-0.36 ***	-0.50 ***	-0.55 ***	-0.49 ***
	STUDIED x ON BENEFIT	-0.53 ***	-0.66 ***	-0.63 ***	-0.13 ns
	STUDIED x HAS LOAN	-0.88 ***	-0.82 ***	-0.57 ***	-0.38 *
	ON BENEFIT x HAS LOAN	0.53 ***	0.61 ***	0.62 ***	0.54 ***
	STUDIED x ON BENEFIT x HAS LOAN	0.54 ***	0.45 *	0.31 ns	0.17 ns
	FEMALE	0.01 ns	-0.07 ***	-0.08 ***	-0.09 ***
	FEMALE x STUDIED	-0.92 ***	-0.77 ***	-0.83 ***	-0.12 ns
	FEMALE x ON BENEFIT	0.06 ns	0.11 ***	0.12 ***	0.13 ***
	FEMALE x HAS LOAN	0.32 ns	0.31 *	0.35 ***	0.32 ***
	FEMALE x STUDIED x ON BENEFIT	0.42 ***	0.40 **	0.31 ns	-0.24 ns
	FEMALE x STUDIED x HAS LOAN	0.46 **	0.54 ***	0.38 *	-0.06 ns
	FEMALE x ON BENEFIT x HAS LOAN	-0.51 ***	-0.41 ***	-0.41 ***	-0.40 ***
	FEMALE x STUDIED x ON BENEFIT x HAS LOAN	-0.11 ns	-0.21 ns	-0.26 ns	0.24 ns
Rate of change in income, post-completion	Intercept: linear term	-0.70 **	0.23 ns	-0.52 ns	-0.19 ns
	Intercept: quadratic term	0.00 ns	-0.15 ***	-0.09 ns	-0.09 ***
	ON BENEFIT	0.34 *	-0.39 ns	0.20 ns	0.63 ns
	HAS LOAN	0.29 ns	0.25 ns	0.24 ns	0.82 ns
	ON BENEFIT x HAS LOAN	-0.38 ns	0.32 ns	-0.01 ns	-0.80 ns
	FEMALE	0.41 ns	0.08 ns	0.73 ns	0.53 ns
	FEMALE x HAS LOAN	-0.35 ns	-0.27 ns	0.01 ns	-0.31 ns
Variance components					
Within person		68.94 ***	69.88 ***	70.87 ***	71.37 ***
Pre-study income between persons		57.20 ***	58.00 ***	58.23 ***	58.14 ***
Rate of change before studying		3.57 ***	3.60 ***	3.65 ***	3.68 ***
Change in income on completing		57.12 ***	50.87 ***	121.23 ***	204.52 ***
Rate of change after completing		11.68 ***	14.33 ***	16.07 ***	20.60 ***
Co-variance components					
Pre-study income and rate of change before studying		-2.47 ***	-2.42 ***	-2.42 ***	-2.36 ***
Pre-study income and change in income on completing		-22.88 ***	-21.13 ***	-13.26 ns	6.69 ns
Pre-study income and rate of change after completing		-1.06 ns	1.09 ns	1.98 ns	-1.74 ns
Rate of change in income before study and change on completing		-0.24 ns	-1.00 ns	-1.57 ns	-8.73 ***
Rate of change in income before study and after completing		-5.10 ***	-5.65 ***	-4.90 ***	-5.27 ***
Change on completing and rate of change after completing		5.68 ***	10.96 ***	-0.73 ns	10.43 *
Sample size		46,302	46,764	48,519	50,301

The probability (p) that a given coefficient is zero is indicated by: *** p<0.004; ** p<0.01; * p<0.05; ns – not significantly different from zero. Confidence indicators are based on bootstrapped standard errors and confidence intervals.

Source: Original data from Statistics NZ Integrated Data Infrastructure. Analysis by Ministry of Education.

Table 4. Regression results for the individual growth models for people in their **forties**

Fixed effects	Coefficients	Certificates Levels 1-3	Certificates Level 4	Diplomas	Bachelors
Pre-study income	Intercept	7.42 ***	7.44 ***	7.49 ***	7.52 ***
	ON BENEFIT	3.36 ***	3.30 ***	3.25 ***	3.24 ***
	ENROLLED	0.71 ***	1.07 *	1.42 *	-1.44 **
	HAS LOAN	0.28 ns	-0.03 ns	-0.15 ns	0.21 ns
	ON BENEFIT x HAS LOAN	-0.47 ns	0.19 ns	0.20 ns	0.09 ns
	FEMALE	-0.97 ***	-1.01 ***	-1.02 ***	-1.00 ***
	FEMALE x ON BENEFIT	3.14 ***	3.18 ***	3.17 ***	3.18 ***
	FEMALE x ENROLLED	-1.03 ***	-0.67 ns	-1.82 ***	-0.07 ns
	FEMALE x HAS LOAN	-0.19 ns	0.03 ns	0.29 ns	-0.86 ns
	FEMALE x ON BENEFIT x HAS LOAN	-0.64 ns	-1.24 ns	-1.30 ns	-0.37 ns
Income on completion	COMPLETED	2.14 *	0.47 ns	5.64 **	11.50 ***
	COMPLETED x ON BENEFIT	-1.76 ***	0.04 ns	-2.52 ***	-5.24 ***
	COMPLETED x HAS LOAN	-0.64 ns	2.35 ns	-2.15 ns	-0.84 ns
	COMPLETED x FEMALE	0.23 ns	0.94 ns	-1.15 ns	2.05 ns
	COMPLETED x FEMALE x HAS LOAN	0.65 ns	-2.27 ns	2.39 ns	3.14 ns
Rate of change in income, pre-study	Intercept: linear term	2.63 ***	2.76 ***	2.79 ***	2.81 ***
	Intercept: quadratic term	-0.15 ***	-0.15 ***	-0.15 ***	-0.15 ***
	STUDIED	1.01 ***	0.81 ***	1.26 ***	0.50 *
	ON BENEFIT	-0.84 ***	-0.90 ***	-0.90 ***	-0.91 ***
	HAS LOAN	-0.17 ns	0.08 ns	0.05 ns	0.02 ns
	STUDIED x ON BENEFIT	-0.52 ***	-0.26 ns	-0.67 ***	0.19 ns
	STUDIED x HAS LOAN	-0.13 ns	-0.41 ns	-0.64 ***	0.08 ns
	ON BENEFIT x HAS LOAN	-0.01 ns	-0.11 ns	-0.06 ns	-0.04 ns
	STUDIED x ON BENEFIT x HAS LOAN	0.59 ns	0.26 ns	0.59 *	-0.25 ns
	FEMALE	0.05 ns	-0.04 ns	-0.04 ns	-0.04 ns
	FEMALE x STUDIED	-0.82 ***	-0.43 ***	-0.74 ***	0.21 ns
	FEMALE x ON BENEFIT	-0.07 ns	-0.02 ns	-0.03 ns	-0.03 ns
	FEMALE x HAS LOAN	0.45 *	0.07 ns	0.06 ns	0.14 ns
	FEMALE x STUDIED x ON BENEFIT	0.38 ***	0.05 ns	0.62 **	-0.60 *
	FEMALE x STUDIED x HAS LOAN	-0.33 ns	0.07 ns	0.26 ns	-0.48 ns
	FEMALE x ON BENEFIT x HAS LOAN	-0.23 ns	0.12 ns	0.11 ns	0.02 ns
	FEMALE x STUDIED x ON BENEFIT x HAS LOAN	0.27 ns	0.04 ns	-0.50 *	0.58 ns
Rate of change in income, post-completion	Intercept: linear term	-1.44 ***	-0.44 ns	-1.18 *	0.62 ns
	Intercept: quadratic term	0.01 ns	-0.02 ns	-0.04 ns	-0.06 ns
	ON BENEFIT	0.04 ns	-0.03 ns	-0.41 ns	-0.03 ns
	HAS LOAN	0.86 **	-0.15 ns	1.54 *	0.00 ns
	ON BENEFIT x HAS LOAN	0.09 ns	0.24 ns	0.45 ns	-0.51 ns
	FEMALE	1.29 ***	0.51 ns	1.72 ***	0.85 ns
	FEMALE x HAS LOAN	-1.03 ***	-0.42 ns	-1.47 *	-0.34 ns
Variance components					
Within person		74.45 ***	75.63 ***	76.15 ***	76.30 ***
Pre-study income between persons		64.29 ***	65.24 ***	65.49 ***	65.65 ***
Rate of change before studying		3.63 ***	3.68 ***	3.70 ***	3.74 ***
Change in income on completing		96.79 ***	55.36 ***	84.64 ***	164.24 ***
Rate of change after completing		9.75 ***	12.71 ***	12.88 ***	15.50 ***
Co-variance components					
Pre-study income and rate of change before studying		-2.19 ***	-2.21 ***	-2.26 ***	-2.27 ***
Pre-study income and change in income on completing		-25.32 ***	-19.48 ***	-14.83 *	-6.79 ns
Pre-study income and rate of change after completing		0.39 ns	-1.08 ns	-4.48 *	-1.50 ns
Rate of change in income before study and change on completing		-2.40 ns	-2.05 ns	-0.87 ns	-6.41 ***
Rate of change in income before study and after completing		-4.19 ***	-4.55 ***	-4.35 ***	-4.94 ***
Change on completing and rate of change after completing		2.92 ns	3.19 ns	9.30 ***	8.90 ***
Sample size		43,842	44,070	45,336	46,371

The probability (p) that a given coefficient is zero is indicated by: *** p<0.004; ** p<0.01; * p<0.05; ns – not significantly different from zero. Confidence indicators are based on bootstrapped standard errors and confidence intervals.

Source: Original data from Statistics NZ Integrated Data Infrastructure. Analysis by Ministry of Education.

Table 5. Regression results for the individual growth models for people in their **fifties**

Fixed effects	Coefficients	Certificates Levels 1-3	Certificates Level 4	Diplomas	Bachelors
Pre-study income	Intercept	8.52 ***	8.48 ***	8.49 ***	8.47 ***
	ON BENEFIT	2.26 ***	2.28 ***	2.27 ***	2.31 ***
	ENROLLED	-0.08 ns	0.81 ns	-0.54 ns	-0.98 ns
	HAS LOAN	-0.01 ns	-0.82 ns	0.02 ns	0.01 ns
	ON BENEFIT x HAS LOAN	-0.24 ns	0.59 ns	-0.28 ns	0.02 ns
	FEMALE	-1.05 ***	-0.97 ***	-0.97 ***	-0.96 ***
	FEMALE x ON BENEFIT	1.60 ***	1.52 ***	1.48 ***	1.45 ***
	FEMALE x ENROLLED	-0.29 ns	-0.39 ns	0.94 ns	1.21 ns
	FEMALE x HAS LOAN	0.18 ns	0.51 ns	-0.31 ns	-0.82 ns
	FEMALE x ON BENEFIT x HAS LOAN	-0.62 ns	-1.18 ns	-0.35 ns	-0.33 ns
Income on completion	COMPLETED	-1.27 ns	2.98 ns	-1.03 ns	11.15 **
	COMPLETED x ON BENEFIT	0.31 ns	-0.25 ns	-2.07 ns	-5.27 ***
	COMPLETED x HAS LOAN	-0.13 ns	-2.24 ns	3.12 ns	-1.51 ns
	COMPLETED x FEMALE	2.30 *	-2.51 ns	2.35 ns	-4.92 ns
	COMPLETED x FEMALE x HAS LOAN	-0.36 ns	2.38 ns	-1.36 ns	3.33 ns
Rate of change in income, pre-study	Intercept: linear term	2.27 ***	2.36 ***	2.39 ***	2.41 ***
	Intercept: quadratic term	-0.15 ***	-0.15 ***	-0.15 ***	-0.15 ***
	STUDIED	0.84 ***	0.84 ***	1.11 ***	0.57 ns
	ON BENEFIT	-0.58 ***	-0.64 ***	-0.65 ***	-0.66 ***
	HAS LOAN	-0.38 *	-0.09 ns	-0.19 ns	-0.11 ns
	STUDIED x ON BENEFIT	-0.47 ***	-0.67 ***	-0.45 *	-0.62 *
	STUDIED x HAS LOAN	0.08 ns	-0.39 ns	-0.30 ns	-0.13 ns
	ON BENEFIT x HAS LOAN	0.28 ns	0.11 ns	0.20 ns	0.01 ns
	STUDIED x ON BENEFIT x HAS LOAN	0.07 ns	0.51 ns	-0.14 ns	0.57 ns
	FEMALE	-0.41 ***	-0.47 ***	-0.47 ***	-0.48 ***
	FEMALE x STUDIED	-0.46 ***	-0.25 ns	-0.33 ns	0.11 ns
	FEMALE x ON BENEFIT	0.20 ***	0.24 ***	0.24 ***	0.25 ***
	FEMALE x HAS LOAN	0.72 **	0.40 ns	0.52 **	0.49 ***
	FEMALE x STUDIED x ON BENEFIT	0.16 ns	0.15 ns	0.09 ns	0.20 ns
	FEMALE x STUDIED x HAS LOAN	-0.43 ns	0.36 ns	-0.07 ns	-0.39 ns
	FEMALE x ON BENEFIT x HAS LOAN	-0.46 ns	-0.20 ns	-0.30 ns	-0.13 ns
	FEMALE x STUDIED x ON BENEF x HAS LOAN	0.30 ns	-0.53 ns	0.21 ns	-0.28 ns
Rate of change in income, post-completion	Intercept: linear term	-0.94 *	-1.01 ns	-0.23 ns	0.40 ns
	Intercept: quadratic term	0.00 ns	-0.04 ns	-0.01 ns	-0.23 ***
	ON BENEFIT	0.26 ns	0.56 ns	0.77 ns	-0.50 ns
	HAS LOAN	0.42 ns	0.25 ns	-0.04 ns	1.20 ns
	ON BENEFIT x HAS LOAN	-0.08 ns	-0.05 ns	-0.61 ns	0.02 ns
	FEMALE	0.83 *	1.54 ***	0.77 ns	1.36 ns
	FEMALE x HAS LOAN	-0.35 ns	-0.76 ns	0.11 ns	-0.90 ns
Variance components					
Within person		66.69 ***	66.82 ***	67.44 ***	67.64 ***
Pre-study income between persons		58.43 ***	58.51 ***	58.62 ***	58.94 ***
Rate of change before studying		2.81 ***	2.83 ***	2.87 ***	2.89 ***
Change in income on completing		78.67 ***	60.52 ***	86.69 ***	150.96 ***
Rate of change after completing		10.92 ***	9.54 **	13.26 **	9.06 ***
Co-variance components					
Pre-study income and rate of change before studying		-2.74 ***	-2.75 ***	-2.76 ***	-2.81 ***
Pre-study income and change in income on completing		-13.84 *	-18.78 **	-9.28 ns	-1.59 ns
Pre-study income and rate of change after completing		-0.27 ns	2.75 ns	1.67 ns	-2.41 ns
Rate of change in income before study and change on completing		-2.22 ns	-0.03 ns	-1.84 ns	-2.49 ns
Rate of change in income before study and after completing		-3.18 ***	-4.12 ***	-4.16 ***	-2.07 ns
Change on completing and rate of change after completing		0.06 ns	1.40 ns	7.11 ns	-3.65 ns
Sample size		35,307	35,436	35,856	36,078

The probability (p) that a given coefficient is zero is indicated by: *** p<0.004; ** p<0.01; * p<0.05; ns – not significantly different from zero. Confidence indicators are based on bootstrapped standard errors and confidence intervals.

Source: Original data from Statistics NZ Integrated Data Infrastructure. Analysis by Ministry of Education.

Table 6. Distribution of unexplained variance across the random effects in the models

		Variance component					
Age in 1999	Level of study	Pre-study income	Rate of increase in income before studying	Step change in income on completing	Rate of increase in income after completing	Within person (through time)	Total variance
20	Certificates at levels 1–3	24%	2%	42%	6%	27%	147.5
	Certificates at level 4	24%	2%	42%	8%	23%	161.1
	Diplomas	21%	1%	49%	6%	22%	184.5
	Bachelors	14%	1%	62%	7%	16%	286.7
30	Certificates at levels 1–3	29%	2%	29%	6%	35%	198.5
	Certificates at level 4	29%	2%	26%	7%	36%	196.7
	Diplomas	22%	1%	45%	6%	26%	270.0
	Bachelors	16%	1%	57%	6%	20%	358.3
40	Certificates at levels 1–3	26%	1%	39%	4%	30%	248.9
	Certificates at level 4	31%	2%	26%	6%	36%	212.6
	Diplomas	27%	2%	35%	5%	31%	242.9
	Bachelors	20%	1%	50%	5%	23%	325.4
50	Certificates at levels 1–3	27%	1%	36%	5%	31%	217.5
	Certificates at level 4	30%	1%	31%	5%	34%	198.2
	Diplomas	26%	1%	38%	6%	29%	228.9
	Bachelors	20%	1%	52%	3%	23%	289.5

The total variance is the sum of the individual variance components, not including the covariance components.
Source: Original data from Statistics New Zealand, Integrated Data Infrastructure.

Table 7. Sample sizes; people with pre-study incomes in 1999 below the repayment threshold, studying for certificates at levels 1 to 3 (if anything), by age, student loan status, benefit status, gender, qualification completion status and time

20 years old in 1999

Has loan	Gender	Benefit status	Completed	Year									
				1999	2000	2001	2002	2003	2004	2005	2006	2007	2008
No	Male	Not on benefit	No	20,790	20,946	21,813	22,503	23,091	23,682	24,105	24,243	24,624	24,669
			Yes	s	33	66	102	153	219	291	396	495	582
		On benefit	No	8,151	7,524	6,324	5,382	4,653	3,954	3,459	3,186	2,844	2,808
			Yes	30	60	69	78	90	105	117	138	141	159
	Female	Not on benefit	No	17,802	17,940	18,417	18,795	19,038	19,320	19,530	19,677	19,965	20,094
			Yes	21	42	72	117	165	243	327	429	528	615
		On benefit	No	7,869	7,281	6,504	5,748	5,271	4,674	4,221	3,843	3,447	3,264
			Yes	48	78	96	96	159	234	255	267	270	258
Yes	Male	Not on benefit	No	591	753	1,116	1,368	1,560	1,776	1,887	1,890	1,857	1,731
			Yes	39	108	228	306	399	474	504	555	576	543
		On benefit	No	2,142	2,238	2,016	1,842	1,668	1,419	1,284	1,245	1,137	1,155
			Yes	192	294	324	369	342	324	300	306	279	303
	Female	Not on benefit	No	792	942	1,191	1,425	1,536	1,623	1,686	1,764	1,836	1,815
			Yes	75	168	249	354	402	441	507	564	591	612
		On benefit	No	2,238	2,322	2,196	2,187	2,088	2,010	2,004	1,932	1,809	1,764
			Yes	243	312	363	369	429	543	558	612	636	663
20 year old group total				61,041	61,041	61,041	61,041	61,041	61,041	61,041	61,041	61,041	61,041

30 years old in 1999

Has loan	Gender	Benefit status	Completed	Year									
				1999	2000	2001	2002	2003	2004	2005	2006	2007	2008
No	Male	Not on benefit	No	10,392	10,476	11,037	11,535	11,979	12,357	12,678	12,759	12,960	12,990
			Yes	s	s	s	33	51	108	156	264	324	360
		On benefit	No	7,344	7,128	6,450	5,820	5,283	4,788	4,386	4,152	3,897	3,798
			Yes	s	s	s	24	42	54	57	78	93	123
	Female	Not on benefit	No	18,981	18,924	18,948	19,095	19,113	19,131	19,200	19,314	19,446	19,443
			Yes	s	36	54	90	147	285	408	576	717	843
		On benefit	No	7,422	7,317	7,083	6,642	6,423	6,048	5,679	5,259	4,920	4,743
			Yes	30	36	54	72	117	195	255	303	333	354
Yes	Male	Not on benefit	No	228	258	345	417	459	531	582	585	570	588
			Yes	s	s	36	54	84	123	138	141	138	144
		On benefit	No	792	849	813	768	747	687	651	648	642	621
			Yes	33	63	93	150	168	162	165	183	186	183
	Female	Not on benefit	No	411	465	513	585	603	606	633	672	717	717
			Yes	27	42	81	102	135	174	204	240	264	297
		On benefit	No	573	606	660	759	780	837	867	843	819	825
			Yes	33	66	99	144	174	216	252	285	279	273
30 year old group total				46,302	46,302	46,302	46,302	46,302	46,302	46,302	46,302	46,302	46,302

Table 7 (Continued). Sample sizes; people with pre-study incomes in 1999 below the repayment threshold, studying for certificates at levels 1 to 3 (if anything), by age, student loan status, benefit status, gender, qualification completion status and time

40 years old in 1999

				Year									
Has loan	Gender	Benefit status	Completed	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008
No	Male	Not on benefit	No	9,612	9,735	10,215	10,698	11,112	11,478	11,811	11,865	12,015	12,054
			Yes	s	s	33	57	84	135	201	327	405	462
		On benefit	No	7,410	7,224	6,657	6,078	5,604	5,142	4,710	4,500	4,287	4,185
			Yes	s	s	21	30	39	60	81	93	99	111
	Female	Not on benefit	No	19,575	19,536	19,470	19,473	19,515	19,488	19,578	19,551	19,716	19,749
			Yes	s	33	81	132	240	369	501	690	831	954
		On benefit	No	6,057	5,961	5,847	5,652	5,430	5,232	4,947	4,674	4,377	4,209
			Yes	s	30	48	66	87	135	165	195	207	222
Yes	Male	Not on benefit	No	129	153	186	210	240	276	276	282	282	261
			Yes	s	s	24	45	54	66	75	90	84	84
		On benefit	No	348	357	363	375	339	321	318	324	300	315
			Yes	39	48	57	69	84	81	87	84	87	90
	Female	Not on benefit	No	270	303	345	378	393	387	399	441	420	408
			Yes	27	48	75	108	126	150	165	171	180	186
		On benefit	No	294	324	345	375	384	396	387	417	399	396
			Yes	24	48	75	93	105	129	138	147	150	162
40 year old group total				43,842	43,842	43,842	43,842	43,842	43,842	43,842	43,842	43,842	43,842

50 years old in 1999

				Year									
Has loan	Gender	Benefit status	Completed	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008
No	Male	Not on benefit	No	8,436	8,532	8,871	9,183	9,438	9,648	9,846	9,921	10,041	10,020
			Yes	s	s	s	24	27	75	138	246	297	327
		On benefit	No	6,090	5,970	5,598	5,256	4,980	4,704	4,419	4,221	4,044	4,026
			Yes	s	s	s	s	24	33	42	60	69	78
	Female	Not on benefit	No	15,012	14,952	14,952	14,979	14,976	14,931	14,955	14,916	14,916	14,898
			Yes	s	s	39	72	105	156	207	285	333	396
		On benefit	No	5,241	5,253	5,181	5,070	5,004	4,959	4,854	4,761	4,692	4,641
			Yes	s	s	24	27	45	63	84	108	129	144
Yes	Male	Not on benefit	No	66	75	84	96	108	108	114	111	105	99
			Yes	s	s	s	s	s	24	21	24	21	21
		On benefit	No	156	153	156	150	144	144	150	138	144	147
			Yes	s	s	s	24	21	24	30	39	42	45
	Female	Not on benefit	No	132	138	153	165	168	180	189	186	177	156
			Yes	s	s	24	30	45	42	48	51	57	57
		On benefit	No	132	144	150	162	162	168	168	183	180	192
			Yes	s	s	21	36	42	45	45	60	63	60
50 year old group total				35,307	35,307	35,307	35,307	35,307	35,307	35,307	35,307	35,307	35,307

Source: Statistics NZ Integrated Data Infrastructure.
All numbers are randomly rounded to base 3.
s indicates there are less than 20 people in the particular category.

Table 8. Sample sizes; people with pre-study incomes in 1999 below the repayment threshold, studying for certificates at level 4 (if anything), by age, student loan status, benefit status, gender, qualification completion status and time

20 years old in 1999

Has loan	Gender	Benefit status	Completed	Year									
				1999	2000	2001	2002	2003	2004	2005	2006	2007	2008
No	Male	Not on benefit	No	20,967	21,159	22,107	22,839	23,472	24,123	24,585	24,792	25,221	25,320
			Yes	s	s	21	33	45	57	90	114	150	183
		On benefit	No	8,265	7,668	6,456	5,487	4,716	4,002	3,522	3,264	2,943	2,919
			Yes	s	s	s	s	s	s	s	s	s	s
	Female	Not on benefit	No	17,886	18,036	18,546	18,945	19,206	19,533	19,764	19,965	20,325	20,466
			Yes	s	s	s	24	42	57	96	126	171	216
		On benefit	No	7,923	7,368	6,576	5,820	5,382	4,824	4,410	4,041	3,627	3,438
			Yes	s	s	s	s	s	39	48	54	69	72
Yes	Male	Not on benefit	No	645	840	1,290	1,602	1,845	2,130	2,226	2,253	2,250	2,082
			Yes	s	s	48	75	102	132	138	150	162	162
		On benefit	No	2,280	2,442	2,214	2,097	1,953	1,677	1,560	1,533	1,392	1,431
			Yes	s	42	54	57	51	60	66	69	66	84
	Female	Not on benefit	No	828	1,023	1,341	1,602	1,728	1,869	1,950	2,052	2,139	2,124
			Yes	s	s	39	78	105	129	165	201	240	264
		On benefit	No	2,340	2,502	2,406	2,442	2,424	2,406	2,403	2,370	2,238	2,202
			Yes	s	42	66	78	96	144	162	195	195	219
20 year old group total				61,197	61,197	61,197	61,197	61,197	61,197	61,197	61,197	61,197	61,197

30 years old in 1999

Has loan	Gender	Benefit status	Completed	Year									
				1999	2000	2001	2002	2003	2004	2005	2006	2007	2008
No	Male	Not on benefit	No	10,401	10,506	11,091	11,619	12,096	12,522	12,867	13,023	13,245	13,299
			Yes	s	s	s	s	s	27	42	72	90	117
		On benefit	No	7,416	7,218	6,564	5,937	5,379	4,884	4,473	4,242	3,984	3,891
			Yes	s	s	s	s	s	s	21	27	36	36
	Female	Not on benefit	No	19,404	19,395	19,413	19,602	19,650	19,683	19,809	19,995	20,196	20,193
			Yes	s	s	s	21	48	132	195	288	381	465
		On benefit	No	7,431	7,311	7,137	6,693	6,495	6,171	5,850	5,433	5,082	4,923
			Yes	s	s	s	s	s	60	78	111	111	117
Yes	Male	Not on benefit	No	246	288	384	459	504	591	651	663	681	693
			Yes	s	s	s	s	s	21	36	39	39	45
		On benefit	No	837	882	840	855	870	807	777	792	771	762
			Yes	s	s	s	s	21	39	42	45	60	69
	Female	Not on benefit	No	438	480	564	651	678	720	753	795	849	861
			Yes	s	s	s	24	39	57	90	120	132	162
		On benefit	No	576	645	693	813	876	954	987	999	978	996
			Yes	s	s	21	39	48	75	96	114	132	135
30 year old group total				46,764	46,764	46,764	46,764	46,764	46,764	46,764	46,764	46,764	46,764

Table 8 (Continued). Sample sizes; people with pre-study incomes in 1999 below the repayment threshold, studying for certificates at level 4 (if anything), by age, student loan status, benefit status, gender, qualification completion status and time

40 years old in 1999

Has loan	Gender	Benefit status	Completed	Year									
				1999	2000	2001	2002	2003	2004	2005	2006	2007	2008
No	Male	Not on benefit	No	9,657	9,783	10,293	10,794	11,208	11,604	11,997	12,144	12,345	12,405
			Yes	s	s	s	s	24	45	66	87	99	120
		On benefit	No	7,443	7,278	6,705	6,138	5,655	5,202	4,767	4,569	4,359	4,275
			Yes	s	s	s	s	s	27	24	33	39	42
	Female	Not on benefit	No	19,845	19,839	19,809	19,830	19,923	19,914	20,040	20,130	20,340	20,427
			Yes	s	s	24	45	84	168	246	342	411	495
		On benefit	No	6,039	5,964	5,883	5,700	5,481	5,310	5,061	4,797	4,500	4,335
			Yes	s	s	s	s	24	57	66	81	93	90
Yes	Male	Not on benefit	No	135	159	201	234	276	312	315	321	324	309
			Yes	s	s	s	s	s	s	21	30	39	39
		On benefit	No	366	378	390	411	408	387	396	405	384	393
			Yes	s	s	s	s	s	27	27	30	30	33
	Female	Not on benefit	No	261	297	351	399	432	438	456	489	489	480
			Yes	s	s	s	30	39	54	63	69	81	84
		On benefit	No	294	327	357	420	447	480	477	492	477	465
			Yes	s	s	s	s	24	36	39	60	63	75
40 year old group total				44,070	44,070	44,070	44,070	44,070	44,070	44,070	44,070	44,070	44,070

50 years old in 1999

Has loan	Gender	Benefit status	Completed	Year									
				1999	2000	2001	2002	2003	2004	2005	2006	2007	2008
No	Male	Not on benefit	No	8,445	8,547	8,889	9,207	9,462	9,696	9,939	10,116	10,269	10,275
			Yes	s	s	s	s	s	33	42	51	75	87
		On benefit	No	6,120	6,003	5,637	5,295	5,010	4,740	4,464	4,269	4,098	4,086
			Yes	s	s	s	s	s	s	24	24	27	30
	Female	Not on benefit	No	15,126	15,063	15,093	15,144	15,153	15,129	15,171	15,171	15,189	15,201
			Yes	s	s	s	27	51	96	117	153	177	207
		On benefit	No	5,250	5,280	5,214	5,097	5,037	4,968	4,890	4,812	4,761	4,713
			Yes	s	s	s	s	s	30	42	57	69	75
Yes	Male	Not on benefit	No	72	75	90	114	123	129	129	135	126	114
			Yes	s	s	s	s	s	s	s	s	s	s
		On benefit	No	153	162	162	159	162	159	174	168	171	171
			Yes	s	s	s	s	s	s	s	s	s	s
	Female	Not on benefit	No	129	138	147	174	180	195	201	198	189	174
			Yes	s	s	s	s	s	s	s	21	24	30
		On benefit	No	132	150	162	186	189	195	186	210	210	216
			Yes	s	s	s	s	s	s	21	24	27	27
50 year old group total				35,436	35,436	35,436	35,436	35,436	35,436	35,436	35,436	35,436	35,436

Source: Statistics NZ Integrated Data Infrastructure.
All numbers are randomly rounded to base 3.
s indicates there are less than 20 people in the particular category.

Table 9. Sample sizes; people with pre-study incomes in 1999 below the repayment threshold, studying for diplomas (if anything), by age, student loan status, benefit status, gender, qualification completion status and time

20 years old in 1999

Has loan	Gender	Benefit status	Completed	Year									
				1999	2000	2001	2002	2003	2004	2005	2006	2007	2008
No	Male	Not on benefit	No	21,486	21,576	22,494	23,253	23,892	24,564	25,059	25,308	25,770	25,932
			Yes	s	21	48	84	123	162	198	225	261	297
		On benefit	No	8,601	7,926	6,663	5,637	4,839	4,098	3,585	3,306	2,985	2,967
			Yes	s	s	s	s	s	s	s	s	s	s
	Female	Not on benefit	No	18,243	18,360	18,909	19,317	19,590	19,935	20,199	20,412	20,790	20,931
			Yes	s	21	60	93	126	156	201	231	279	339
		On benefit	No	8,400	7,740	6,822	6,003	5,529	4,968	4,536	4,170	3,750	3,576
			Yes	s	s	s	s	s	21	21	21	21	21
Yes	Male	Not on benefit	No	1,011	1,215	1,740	2,097	2,376	2,694	2,841	2,868	2,868	2,655
			Yes	s	42	120	171	213	270	321	333	330	333
		On benefit	No	2,787	3,081	2,754	2,565	2,370	2,040	1,845	1,803	1,623	1,653
			Yes	30	72	108	129	123	108	84	90	99	93
	Female	Not on benefit	No	1,143	1,350	1,710	2,013	2,175	2,334	2,430	2,559	2,694	2,688
			Yes	s	54	150	231	279	333	390	402	435	444
		On benefit	No	2,898	3,120	2,949	2,958	2,898	2,859	2,838	2,793	2,619	2,592
			Yes	60	108	144	138	156	165	156	183	183	177
20 year old group total				64,713	64,713	64,713	64,713	64,713	64,713	64,713	64,713	64,713	64,713

30 years old in 1999

Has loan	Gender	Benefit status	Completed	Year									
				1999	2000	2001	2002	2003	2004	2005	2006	2007	2008
No	Male	Not on benefit	No	10,602	10,689	11,262	11,802	12,267	12,720	13,071	13,248	13,500	13,578
			Yes	s	s	s	21	33	42	63	69	99	120
		On benefit	No	7,554	7,317	6,639	5,976	5,403	4,896	4,488	4,260	4,020	3,930
			Yes	s	s	s	s	s	s	s	s	s	s
	Female	Not on benefit	No	20,076	20,004	19,989	20,130	20,178	20,274	20,430	20,661	20,889	20,931
			Yes	s	s	24	42	69	105	141	183	234	303
		On benefit	No	7,626	7,503	7,299	6,837	6,618	6,312	5,985	5,592	5,253	5,088
			Yes	s	s	s	s	s	s	30	33	33	39
Yes	Male	Not on benefit	No	306	348	480	573	645	756	828	840	837	843
			Yes	s	s	30	48	66	81	96	105	102	99
		On benefit	No	1,002	1,092	1,023	1,014	1,011	936	888	906	882	867
			Yes	s	27	36	45	54	48	48	45	39	45
	Female	Not on benefit	No	540	618	717	843	867	924	978	1,050	1,137	1,152
			Yes	s	27	60	78	93	117	144	177	198	237
		On benefit	No	744	834	870	1,014	1,104	1,179	1,203	1,212	1,173	1,176
			Yes	24	33	63	72	84	102	117	126	114	108
30 year old group total				48,519	48,519	48,519	48,519	48,519	48,519	48,519	48,519	48,519	48,519

Table 9 (Continued). Sample sizes; people with pre-study incomes in 1999 below the repayment threshold, studying for diplomas (if anything), by age, student loan status, benefit status, gender, qualification completion status and time

40 years old in 1999

Has loan	Gender	Benefit status	Completed	Year									
				1999	2000	2001	2002	2003	2004	2005	2006	2007	2008
No	Male	Not on benefit	No	9,828	9,930	10,431	10,926	11,358	11,772	12,174	12,330	12,558	12,636
			Yes	s	s	s	21	27	33	45	60	72	87
		On benefit	No	7,515	7,341	6,747	6,165	5,685	5,235	4,803	4,605	4,395	4,314
			Yes	s	s	s	s	s	s	s	s	s	s
	Female	Not on benefit	No	20,373	20,331	20,274	20,304	20,403	20,451	20,586	20,688	20,940	21,054
			Yes	s	s	39	63	87	132	192	249	303	354
		On benefit	No	6,219	6,111	6,015	5,802	5,583	5,406	5,163	4,914	4,617	4,452
			Yes	s	s	s	s	24	30	36	30	36	33
Yes	Male	Not on benefit	No	168	198	246	297	339	387	405	414	411	384
			Yes	s	s	s	21	33	45	51	51	57	60
		On benefit	No	450	483	498	516	495	462	465	474	450	459
			Yes	s	s	24	27	33	27	27	33	30	30
	Female	Not on benefit	No	339	405	453	516	564	576	600	660	645	648
			Yes	s	21	54	75	93	117	141	156	168	186
		On benefit	No	393	444	474	543	546	582	573	594	582	570
			Yes	s	21	36	45	60	66	69	63	63	66
40 year old group total				45,336	45,336	45,336	45,336	45,336	45,336	45,336	45,336	45,336	45,336

50 years old in 1999

Has loan	Gender	Benefit status	Completed	Year									
				1999	2000	2001	2002	2003	2004	2005	2006	2007	2008
No	Male	Not on benefit	No	8,541	8,634	8,976	9,291	9,546	9,795	10,050	10,230	10,404	10,419
			Yes	s	s	s	s	s	s	s	24	30	39
		On benefit	No	6,168	6,039	5,667	5,322	5,043	4,785	4,503	4,311	4,134	4,119
			Yes	s	s	s	s	s	s	s	s	s	s
	Female	Not on benefit	No	15,228	15,162	15,186	15,240	15,258	15,270	15,327	15,351	15,387	15,426
			Yes	s	s	s	24	33	42	57	72	93	102
		On benefit	No	5,301	5,310	5,244	5,118	5,052	5,004	4,929	4,866	4,821	4,776
			Yes	s	s	s	s	s	s	s	s	s	s
Yes	Male	Not on benefit	No	84	96	114	138	153	165	174	177	156	147
			Yes	s	s	s	s	s	21	21	21	27	30
		On benefit	No	186	207	207	204	201	192	201	198	204	207
			Yes	s	s	s	s	s	s	s	21	21	21
	Female	Not on benefit	No	162	174	183	210	228	237	240	237	228	213
			Yes	s	s	s	33	39	42	63	75	72	63
		On benefit	No	162	195	195	222	225	228	207	228	234	243
			Yes	s	s	27	24	33	36	42	36	30	36
50 year old group total				35,856	35,856	35,856	35,856	35,856	35,856	35,856	35,856	35,856	35,856

Source: Statistics NZ Integrated Data Infrastructure.
All numbers are randomly rounded to base 3.
s indicates there are less than 20 people in the particular category.

Table 10. Sample sizes; people with pre-study incomes in 1999 below the repayment threshold, studying for bachelors degrees (if anything), by age, student loan status, benefit status, gender, qualification completion status and time

20 years old in 1999

Has loan	Gender	Benefit status	Completed	Year									
				1999	2000	2001	2002	2003	2004	2005	2006	2007	2008
No	Male	Not on benefit	No	22,689	22,368	23,127	23,760	24,363	25,011	25,521	25,785	26,289	26,487
			Yes	s	24	117	282	432	585	723	807	894	969
		On benefit	No	9,015	8,211	6,834	5,757	4,923	4,167	3,645	3,354	3,036	3,012
			Yes	s	s	27	39	51	39	27	24	s	24
	Female	Not on benefit	No	19,701	19,299	19,632	19,872	20,085	20,412	20,688	20,916	21,327	21,513
			Yes	s	72	252	549	789	987	1,158	1,266	1,383	1,515
		On benefit	No	8,928	8,118	7,110	6,192	5,673	5,067	4,644	4,251	3,825	3,636
			Yes	s	27	69	87	66	42	39	27	30	33
Yes	Male	Not on benefit	No	2,133	2,586	3,195	3,501	3,672	4,002	4,107	4,116	4,074	3,837
			Yes	s	33	162	390	615	810	951	1,029	1,068	1,095
		On benefit	No	3,609	4,173	3,879	3,519	3,177	2,640	2,337	2,211	1,974	1,950
			Yes	s	60	120	213	228	207	153	141	108	90
	Female	Not on benefit	No	2,691	3,111	3,387	3,438	3,423	3,543	3,570	3,654	3,771	3,735
			Yes	s	129	483	948	1,371	1,608	1,713	1,830	1,878	1,887
		On benefit	No	4,143	4,599	4,305	4,029	3,774	3,597	3,465	3,348	3,096	3,009
			Yes	s	138	261	375	309	237	216	201	183	168
20 year old group total				72,954	72,954	72,954	72,954	72,954	72,954	72,954	72,954	72,954	72,954

30 years old in 1999

Has loan	Gender	Benefit status	Completed	Year									
				1999	2000	2001	2002	2003	2004	2005	2006	2007	2008
No	Male	Not on benefit	No	10,668	10,743	11,313	11,850	12,318	12,768	13,125	13,308	13,590	13,689
			Yes	s	s	s	33	42	48	63	69	78	90
		On benefit	No	7,593	7,347	6,663	5,994	5,433	4,923	4,518	4,296	4,056	3,960
			Yes	s	s	s	s	s	s	s	s	s	s
	Female	Not on benefit	No	20,700	20,601	20,523	20,637	20,643	20,706	20,862	21,066	21,309	21,363
			Yes	s	s	42	81	114	150	189	228	291	333
		On benefit	No	7,845	7,683	7,473	6,969	6,756	6,441	6,102	5,700	5,367	5,193
			Yes	s	s	s	21	21	21	24	24	21	24
Yes	Male	Not on benefit	No	381	441	591	714	813	942	1,029	1,035	1,026	1,041
			Yes	s	s	36	54	78	96	93	105	117	117
		On benefit	No	1,257	1,347	1,266	1,227	1,194	1,113	1,050	1,062	1,014	984
			Yes	s	27	33	48	39	27	30	33	27	33
	Female	Not on benefit	No	750	831	978	1,140	1,182	1,254	1,314	1,419	1,485	1,518
			Yes	s	45	108	159	219	255	297	354	396	465
		On benefit	No	1,038	1,152	1,182	1,311	1,395	1,485	1,515	1,512	1,440	1,404
			Yes	s	51	60	69	57	72	78	81	84	84
30 year old group total				50,301	50,301	50,301	50,301	50,301	50,301	50,301	50,301	50,301	50,301

Table 10 (Continued). Sample sizes; people with pre-study incomes in 1999 below the repayment threshold, studying for bachelors degrees (if anything), by age, student loan status, benefit status, gender, qualification completion status and time

40 years old in 1999

Has loan	Gender	Benefit status	Completed	Year									
				1999	2000	2001	2002	2003	2004	2005	2006	2007	2008
No	Male	Not on benefit	No	9,876	9,975	10,485	10,974	11,406	11,829	12,231	12,399	12,633	12,726
			Yes	s	s	s	s	s	27	36	42	54	63
		On benefit	No	7,563	7,386	6,780	6,198	5,715	5,256	4,821	4,632	4,422	4,335
			Yes	s	s	s	s	s	s	s	s	s	s
	Female	Not on benefit	No	20,790	20,700	20,631	20,604	20,691	20,742	20,913	21,045	21,306	21,435
			Yes	s	s	42	75	120	153	204	240	288	336
		On benefit	No	6,306	6,192	6,093	5,877	5,652	5,481	5,247	4,983	4,692	4,515
			Yes	s	s	s	s	s	s	s	s	s	s
Yes	Male	Not on benefit	No	213	249	312	366	423	474	498	510	507	483
			Yes	s	s	21	36	42	54	60	60	63	60
		On benefit	No	552	585	600	609	588	558	552	549	519	522
			Yes	s	s	s	s	s	s	s	21	21	24
	Female	Not on benefit	No	504	567	639	717	765	792	804	837	837	849
			Yes	s	42	75	135	174	189	216	246	270	273
		On benefit	No	534	615	624	690	705	732	708	735	699	672
			Yes	s	s	39	39	36	45	42	48	51	51
40 year old group total				46,371	46,371	46,371	46,371	46,371	46,371	46,371	46,371	46,371	46,371

50 years old in 1999

Has loan	Gender	Benefit status	Completed	Year									
				1999	2000	2001	2002	2003	2004	2005	2006	2007	2008
No	Male	Not on benefit	No	8,550	8,640	8,982	9,294	9,555	9,810	10,071	10,254	10,440	10,458
			Yes	s	s	s	s	s	s	s	s	s	s
		On benefit	No	6,171	6,048	5,670	5,331	5,052	4,788	4,509	4,317	4,143	4,131
			Yes	s	s	s	s	s	s	s	s	s	s
	Female	Not on benefit	No	15,324	15,252	15,273	15,324	15,351	15,375	15,438	15,465	15,516	15,561
			Yes	s	s	24	30	36	45	54	60	69	75
		On benefit	No	5,310	5,328	5,259	5,130	5,076	5,022	4,950	4,890	4,839	4,797
			Yes	s	s	s	s	s	s	s	s	s	s
Yes	Male	Not on benefit	No	87	108	123	153	171	186	192	192	177	177
			Yes	s	s	s	s	s	s	s	s	21	21
		On benefit	No	231	243	252	243	237	225	231	237	237	231
			Yes	s	s	s	s	s	s	s	s	s	s
	Female	Not on benefit	No	204	207	228	261	267	282	285	291	273	258
			Yes	s	s	21	30	30	30	33	42	42	45
		On benefit	No	186	216	216	246	261	258	249	267	273	279
			Yes	s	s	s	s	s	s	s	s	s	s
50 year old group total				36,078	36,078	36,078	36,078	36,078	36,078	36,078	36,078	36,078	36,078

Source: Statistics NZ Integrated Data Infrastructure.
All numbers are randomly rounded to base 3.
s indicates there are less than 20 people in the particular category.

Table 11. Correspondence analysis results; coordinates of dimensions for broad fields of study, by level of study and birth year, gender and benefit status

	Certificates at levels 1 to 3		Certificates at levels 4		Diplomas		Degrees	
	Dim1	Dim2	Dim1	Dim2	Dim1	Dim2	Dim1	Dim2
20-29	0.25	0.57	0.65	-0.44	0.52	-0.06	-0.37	-0.13
30-39	-0.32	-0.11	-0.36	-0.01	-0.46	0.39	0.97	0.50
40-49	-0.12	1.07	-0.59	0.72	-0.76	-0.31	1.22	0.23
Female	-0.60	0.06	-0.48	-0.19	-0.55	-0.06	0.49	-0.24
Male	1.07	-0.11	1.17	0.46	1.07	0.11	-0.99	0.48
Had benefit	-0.45	-0.68	0.06	-0.85	-0.37	1.44	1.08	2.08
No benefit	0.27	0.41	-0.03	0.46	0.10	-0.39	-0.10	-0.20
Agriculture, environmental and related studies	1.18	0.65	1.02	-0.19	1.00	-1.26	-0.89	-0.39
Architecture and building	2.04	-0.73	2.29	1.92	1.11	-1.35	-0.91	-0.26
Creative arts	0.32	-0.69	0.07	-1.10	0.45	1.14	-0.43	0.35
Education	-1.19	2.05	-1.38	1.14	-0.96	-0.64	1.01	-0.42
Engineering and related technologies	1.82	-0.45	2.00	0.66	1.86	-0.46	-1.31	0.29
Food, hospitality and personal services	-0.26	-1.28	0.09	-0.69	-0.30	1.07		
Health	0.40	1.87	-0.45	0.20	-1.05	0.02	0.93	-0.34
Information technology	0.02	-0.63	1.15	0.22	1.14	0.55	-1.30	1.67
Management and commerce	-0.52	0.35	-0.25	-0.56	0.12	-0.90	-0.59	-0.89
Mixed field programmes	-0.84	-0.39	0.13	-1.48				
Natural and physical sciences							-0.92	-0.13
Society and culture	0.23	0.05	-0.31	0.28	-0.30	0.37	0.13	0.94
Per cent variance explained	55%	12%	59%	10%	64%	8%	76%	4%
Total sample size		17,634		6,405		6,486		12,156

Blank entries indicate the particular field of study was either not offered at the particular level of study, or was excluded because too few students completed qualifications in this field at that level.
Sample sizes have been randomly rounded to base 3.
Original data from Statistics NZ Integrated Data Infrastructure. Analysis by Ministry of Education.

Table 12. Correspondence analysis results; coordinates of dimensions for narrow fields of study, by level of study and birth year, gender and benefit status

	Certificates at levels 1 to 3		Certificates at levels 4		Diplomas		Degrees	
	Dim1	Dim1	Dim1	Dim1	Dim2	Dim2	Dim1	Dim2
20-29	0.22	0.49	0.65	0.61	0.54	-0.15	0.38	-0.13
30-39	-0.28	-0.09	-0.39	-0.07	-0.47	0.41	-1.01	0.55
40-49	-0.10	-0.97	-0.50	-0.81	-0.89	-0.10	-1.24	0.13
Female	-0.48	-0.17	-0.45	0.15	-0.54	-0.10	-0.47	-0.24
Male	1.14	0.40	1.32	-0.44	1.05	0.19	1.01	0.52
Had benefit	-0.64	0.68	-0.13	0.80	-0.24	1.46	-1.03	2.12
No benefit	0.41	-0.45	0.07	-0.44	0.06	-0.40	0.10	-0.20
Accountancy							0.40	0.07
Architecture and Urban Environment							0.87	-0.25
Automotive Engineering and Technology	1.92	0.83						
Behavioural Science							-0.21	1.69
Biological Sciences							0.62	0.60
Building			2.44	-0.96				
Business and Management	0.14	-0.31	-0.19	-0.15	0.06	-0.93	0.58	-0.74
Communication and Media Studies			1.43	0.47	0.91	0.20	0.29	-0.39
Computer Science					1.11	0.62	1.49	1.55
Electrical & Electronic Engineering & Technology					1.82	0.14		
Employment Skills Programmes	-1.11	0.67						
Food and Hospitality	0.33	0.90	0.58	-0.21				
General Education Programmes	0.48	0.03						
Graphics and Design Studies					0.72	0.45	0.57	0.22
Human Welfare Studies & Services	-0.56	-1.09	-0.70	-0.09	-1.29	1.13		
Information Systems					0.76	0.53	1.13	0.53
Language and Literature	0.75	-0.42	-0.22	-0.46	-0.05	-0.40	-0.18	1.47
Nursing							-1.39	0.12
Office Studies	-0.56	-0.34	-0.37	1.71				
Other Education			-1.13	-0.49				
Other Natural and Physical Sciences							0.86	-0.48
Other Society and Culture							0.15	0.18
Performing Arts					0.70	1.30	0.74	0.58
Personal Services	-0.71	0.42	-0.25	1.35	-0.60	0.85		
Public Health	1.33	-1.42						
Rehabilitation Therapies							-0.03	-1.55
Sales and Marketing							0.71	-0.35
Social Skills Programmes	0.07	-1.05						
Sport and Recreation	1.29	1.88	1.84	0.48	1.03	-0.24	0.65	0.00
Studies in Human Society	-0.86	0.06					-0.09	0.47
Teacher Education			-0.85	-1.47	-0.93	-0.48	-0.90	-0.45
Tourism	0.04	-0.01	0.41	0.57	0.30	-1.17		
Visual Arts and Crafts			0.13	0.94	-0.84	1.57	-0.04	1.49
Per cent variance explained	48%	14%	48%	18%	62%	9%	73%	5%
Total sample size		12,351		4,569		4,667		10,125

Blank entries indicate the particular field of study was either not offered at the particular level of study, or was excluded because too few students completed qualifications in this field at that level.
Sample sizes have been randomly rounded to base 3.
Original data from Statistics NZ Integrated Data Infrastructure. Analysis by Ministry of Education.

MINISTRY OF EDUCATION

Te Tāhuhu o te Mātauranga

newzealand.govt.nz