University of Southampton

Faculty of Law, Arts & Social Sciences

School of Social Sciences

The educational and labour market expectations of adolescents and young adults

by

John Jerrim

Thesis for the degree of Doctor of Philosophy

January 2011

Abstract

Understanding why some suitably qualified young adults go on to enter higher education and others do not has been the subject of extensive research by a number of social scientists from a range of disciplines. Economists suggest that young adults' willingness to invest in a tertiary qualification depends upon what they believe the costs and benefits of this investment will be. On the other hand, sociologists stress that an early expectation of completing university is a key driver of later participation in higher education. Children's subjective beliefs of the future (their "expectations") are a consistent theme within these distinctively different approaches. Researchers from both disciplines might argue that children's low or mistaken expectations (of future income, financial returns, their ability to complete university) might lead them into making inappropriate educational choices. For instance, young adults who do not have a proper understanding of the graduate labour market may mistakenly invest (or not invest) in tertiary education. Alternatively some academically talented children may not enter university if they do not see it as realistic possibility, or that it is 'not for the likes of them'. I take an interdisciplinary approach within this thesis to tackle both of these issues. Specifically, I investigate whether young adults have realistic expectations about their future in the labour market and if disadvantaged children scoring high marks on a maths assessment at age 15 believe they can complete university.

Contents

List of Tables

Appendix Tables.		Page

List of Figures

List of Figures continued

Declaration of Authorship

I, John Jerrim, declare that the thesis entitled "The educational and labour market expectations of adolescents and young adults" and the work presented in the thesis are both my own, and have been generated by me as the result of my own original research. I confirm that:

- this work was done wholly or mainly while in candidature for a research degree at this University;
- where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
- where I have consulted the published work of others, this is always clearly attributed;
- where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
- I have acknowledged all main sources of help;
- where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;

Signed: . . . .

Date: . . . .

Acknowledgements

Firstly, I would like to thank all members of my supervisory team (John Micklewright, Sylke Schnepf and Dave Holmes) for all their help, hard work and advice. Likewise, I would like to thank Tim Smeeding, Karl Scholz and Jane Cooley for their warmth and hospitality during my two month stay in Madison, where I was putting together the finishing touches on this thesis. Bob Michael also deserves special thanks, for reader through a provisional draft of my third chapter and making a series of useful suggestions. Feedback received from other University of Southampton staff during the course of this work has been critical to both the development of my own skills and each of the chapters included in this work. Otherwise, special thanks goes to my family and friends who supported me over the last three years, especially my loving girlfriend Jackie.

Chapter 1

Introduction:
The educational and labour market expectations of adolescents and young adults

Whether to invest in higher education is one of the most important decisions that young people make. Such an investment can offer substantial financial rewards; the UK government often cites work that suggests graduates earn, on average, £100,000 more over their lifetime than if they entered the labour force upon finishing their A-Levels (Department for Business, Innovation and Skills 2008). Yet completing higher education is also becoming more expensive. At the time of writing, the £3,000 cap on university tuition fees is under review, with many parties expecting it to be lifted. Hence educational decisions typically made at age 18 are having ever greater financial implications, with increasing responsibility being placed onto relatively inexperienced shoulders.

Many social scientists claim that adolescents' subjective beliefs about the future (their “expectations”) are a central determinant of such choices. From the one side, economists suggest that young adults’ willingness to invest in a tertiary qualification depends upon what they *expect* the costs and benefits of this investment to be. On the other, sociologists and social psychologists stress that an early expectation of completing university is a key motivational factor behind actual later attainment. Recent theories have emerged that integrate both of these views. In any case, social scientists from a broad range of disciplines might argue that children's low or mistaken expectations (of future income, financial returns or chances of completing university) could lead them into making an inappropriate educational choice. For instance, young adults who do not have a reasonable understanding of the graduate labour market may mistakenly invest (or not invest) in tertiary education. Alternatively some academically talented children may not enter university if they do not see it as realistic possibility, or that it is 'not for the likes of them'. Under both scenarios, such mistaken decisions are likely to be detrimental for output, and possibly inequality, within the UK.

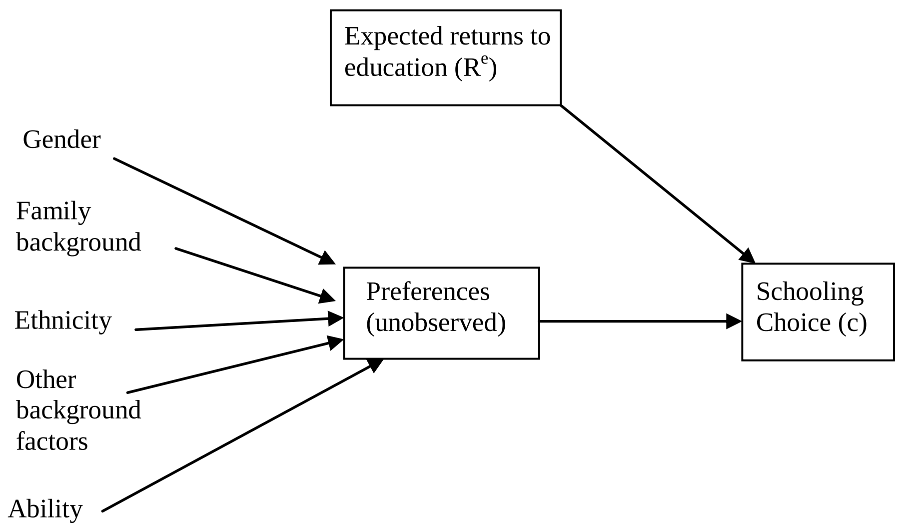
Policymakers should therefore carefully consider children's expectations when designing their educational initiatives. For instance, any scheme to create a market place within UK higher education will only be efficient if students are realistic about the future, understanding the economic costs and benefits of the different options that are available. Likewise, if the current university tuition fee cap is to be lifted, then this cannot be allowed to create a perception amongst disadvantaged children that higher education is the preserve of the rich. In other words, it is vital for policymakers to ensure that young adults are making well informed human capital investment decisions and that all those who can benefit from higher education believe it is an obtainable goal, regardless of their family background.

But are young adults capable of making such rational assessments of the future, and do they hold enough information to make such complex decisions? I take an inter-disciplinary approach to provide an insight into these issues. Specifically, I investigate whether young adults are realistic about their future in the labour market and if children from disadvantaged backgrounds are less optimistic about their prospects of completing university than their more affluent peers. I begin in this introduction by describing economic and sociological theories of schooling choice and the role that expectations play within these frameworks. This includes a description of how social scientists measure children's beliefs about the future and incorporate such factors into their empirical models of educational choice. A brief review of existing studies follows, accompanied by a discussion of the contribution that this thesis makes to the wider literature. To conclude, I will provide an overview of the findings from this work, along with a summary of what they imply for educational policy in the UK.

1.2. The role of adolescents' expectations in economic models of schooling choice

In an economic framework, a central determinant of young adults' educational choice is what they *expect* the financial returns of this investment to be (R^e). Indeed, several studies have tried to assess how important this factor is for children's decisions regarding higher education (Willis and Rosen 1979, Berger 1988, Kaufmann and Attanasio 2009, Arcidiacono et al 2010, Montmarquette et al 2002). In such work, expected returns are usually decomposed into the higher wages an individual expects to receive over their working lifetime (as a result of this human capital investment) less the expected costs of completing this extra level of education. The latter consists of both direct (e.g. tuition fees) and indirect (e.g. foregone income/ opportunity costs) components. All figures are typically adjusted to take into account the time value of money (i.e. a discount rate, r, is applied). Economists also recognise that adolescents' tastes or “preferences” play an important part in their educational behaviour. For instance, an individual who enjoys studying will be more likely to continue their investment in education for a given level of financial return (i.e. education has both investment and consumption value for them). However, such preferences are usually unobservable to the analyst; hence economists include a set of other exogenous variables (X) into their models to try and take children’s educational tastes into account. A graphical representation of this framework can be found in Figure 1.1.

Figure 1.1 Economic model of schooling choice



Notes:

1 Source –authors own diagram

In empirically implementing this framework, economists usually set up a statistical model as illustrated (algebraically) below:

$C = f(X, R^e)$

$$R^e = \sum_{t=0}^{t}\left[\left(W^e_{t,c=j} - W^e_{t,c=k}\right)/(1+r)^t\right] - \sum_{t=0}^{t}\left[\left(Co^e_{t,c=j} - Co^e_{t,c=k}\right)/(1+r)^t\right]$$

Where:

C = The educational choice (e.g. whether to go to university or not)

X = A set of characteristics that capture children's educational preferences (e.g. gender, family background, ability)

R^e = The *expected* returns to that educational choice

r = Discount rate (i.e. time discounting)

$W^e_{t,c=j}$ = Expected income received in year t under educational choice j (e.g. go to university)

$W^e_{t,c=k}$ = Expected income received in year t under educational choice k (e.g. do not go to university)

$Co^e_{t,c=j}$ = Expected costs (direct and indirect) associated with educational choice j in year t (e.g. forgone income, tuition fees)

$Co^e_{t,c=k}$ = Expected costs (direct and indirect) associated with educational choice k in year t (e.g. forgone income, tuition fees)

j = choice j (e.g. go to university)

k = choice k (e.g. do not go to university)

t = time, starting at the point the investment is made (t=0), running until the costs and benefits of the human capital investment end

This clearly demonstrates the theoretical importance of expectations in economic models of schooling behaviour; a child is thought to be more likely to continue their schooling (ceteris paribus) the greater the financial return they expect to this investment. Yet, despite its central role in such models, economists have rarely attempted to measure children's expected returns to education. Consequently, subjective measures of R^e generally do not appear in economists' empirical models of schooling behaviour. Rather, information collected “ex-post” (after the decision has been made) is typically used to make inferences about the decision making process itself. In other words, subjective measures of young adults' *expected* returns to education (R^e) are generally not available (because they are not collected), so economists estimate the *actual* returns to education (R^a) instead and incorporate this into their models of schooling choice[1]. Berger (1988) and Willis and Rosen (1979) are two well known examples that use this strategy. Hence the model set out above is rarely estimated; rather economists typically prefer the specification:

$$C = f(X, R^a)$$

$$R^a = \sum_{t=0}^{t}\left[\left(W^a_{t,c=j} - W^a_{t,c=k}\right)/(1+r)^t\right] - \sum_{t=0}^{t}\left[\left(Co^a_{t,c=j} - Co^a_{t,c=k}\right)/(1+r)^t\right]$$

where:

R^a = The *“actual”* returns to that educational choice

$W^a_{t,c=j}$ = “Actual” income received at time t under educational choice j (e.g. go to university)

$W^a_{t,c=k}$ = “Actual” income received at time t under educational choice k (e.g. do not go to university)

$Co^a_{t,c=j}$ = “Actual” costs (direct and indirect) associated with educational choice j in year t (e.g. forgone income, tuition fees)

$Co^a_{t,c=k}$ = “Actual” costs (direct and indirect) associated with educational choice k in year t (e.g. forgone income, tuition fees)

[1] Note that this quantity (the actual returns to education) must be estimated from the data. When doing so, economists face the well known problem of not observing what a child would go on to earn under the counter-factual (one does not know what a child who went to university would have earned had they not made this investment). I will not go into detail about this problem here, as it has already received much attention in the literature (see Angrist and Krueger (1991), Card (1995) and Dickson (2009)). Yet it is worth remembering that it is not a trivial problem to estimate this value (R^a) in itself.

A critical assumption that is therefore invoked in most economic studies of educational choice, particularly ones which attempt to estimate the impact of children's expected returns on their schooling behaviour, is:

$$\mathrm{R}^{e} = \mathrm{R}^{a} + \varepsilon$$

where:
ε = random error, assumed to be mean zero and independent of all other covariates (X) that are included in the model

This is sometimes called the "rational expectations" hypothesis. It states that if young adults have "full information" (that they fully understand the cost and benefits of different educational options), and process this information "rationally" (they can use it to make reasonable predictions of the future), then there should only be random differences between their expectations and later realisations. In other words, economists usually assume that there are no *systematic* differences between children's ex-ante labour market expectations and their later ex-post realisations[2]. This assumption has recently been criticised in the literature (Manski 2004):

"If experts disagree on the returns to schooling, is it plausible that youth have rational expectations? I think not"

but has remained the subject of relatively little empirical research (particularly outside the US). Indeed, due to the dearth of data on young adults' financial expectations, no study has been able to confirm whether the assumption above (that $\mathrm{R}^{e} = \mathrm{R}^{a} + \varepsilon$), vital to the inferences made in many economic models of schooling behaviour, is credible.

Yet there is a growing desire amongst economists to understand whether young adults are realistic about their prospects in the labour market. Hence a small but developing literature, briefly reviewed below, has considered whether children hold realistic expectations of future *wages* (rather than returns) at one particular point in time

[2] If this assumption is violated, then R^{a} can be thought of as an error prone measure of R^{e}. This in turn violates the error-in-variables assumption of regression analysis, with the estimated effect of the returns to education on schooling choice becoming biased.

(rather than the discounted value of all future streams). In other words, although not being able to show if:

$$\mathrm{R}^{e} = \mathrm{R}^{a} + \varepsilon \quad \Rightarrow \quad \sum_{t=0}^{t}\left[\left(W^{e}_{t,c=j} - W^{e}_{t,c=k}\right)/(1+r)^{t}\right] - \sum_{t=0}^{t}\left[\left(Co^{e}_{t,c=j} - Co^{e}_{t,c=k}\right)/(1+r)^{t}\right]$$

$$=$$

$$\sum_{t=0}^{t}\left[\left(W^{a}_{t,c=j} - W^{a}_{t,c=k}\right)/(1+r)^{t}\right] - \sum_{t=0}^{t}\left[\left(Co^{a}_{t,c=j} - Co^{a}_{t,c=k}\right)/(1+r)^{t}\right] + \varepsilon$$

is a credible assumption, one can get a grasp of whether young adults have a reasonable understanding of a related (and, indeed, much simpler) issue:

$$W^{e}_{t=x,c=j} = W^{a}_{t=x,c=j} + \varepsilon$$

where x = some specified point in the future

Despite the limitations of this approach, such studies still provide an intriguing insight into the methodological framework described above. For instance, if wage expectations differ systematically from realizations (i.e. some groups make better predictions of the future than others) then, although not falsifying the assumption that $\mathrm{R}^{e} = \mathrm{R}^{a} + \varepsilon$, it is at least consistent with academics like Manski's doubts. To put this another way, if even the accuracy of young adults' wage expectations differs between groups at just one point in time, then a question mark must surely also hang over the assumption that there are no systematic differences in the accuracy of their expected returns. Furthermore, if groups who probably hold quite detailed and specific labour market information (e.g. students with access to public pay scales, those about to graduate, the labour market active) make better predictions of future income than those with relatively little, broad information (e.g. university freshmen, art students) then this would be consistent with the idea that some young adults do not fully understand their future career prospects when making their educational choices.

However, even analysis of young adults' wage expectations is relatively rare within the literature. To my knowledge, there are only four published studies in Europe (Wolter 2000, Webbink and Hartog 2004, Wolter and Zbinden 2002 and Brunello et al 2004), and six in the US (Smith and Powell 1990, Betts 1996, Blau 1991, Carvajal et al 2000, Rouse 2004, Dominitz and Manski 1996). The conclusions drawn from this work have been largely inconsistent. Some indicate that students “overestimate the university wage premium” (Brunello et al 2001). Others suggest that “expected wages after graduation at university are rather accurate” (Wolter 2000), while Webbink and Hartog go a step further, boldly entitling their paper: “Can students predict their starting salary? YES!”.

1.3 The Breen and Goldthorpe model of educational choice

The model of educational choice laid out in section 1.2 is, of course, just one (disciplinary-specific way) of conceptualising the university decision making process. Valuable alternatives have been developed by leading academics from other disciplines, which provides key insights from other perspectives. This sub-section is devoted to one particularly important framework that has been widely studied in the sociological literature (the Breen and Goldthorpe 1997 model) which is also based upon rational choice.

The origins of this model stem from the above authors’ desires to explain a widely cited empirical finding by sociologists – that in most developed countries, there has been ‘little change in socio-economic inequality of educational opportunity’ over time (Breen 2001, Blossfeld and Shavit 1993). Many other international comparisons of social class fluidity have reached a similar conclusion (Erikson and Goldthorpe 1992), although there are some who might disagree (Ganzeboom et al 1989)[3].

Nevertheless, in 1997 Breen and Goldthorpe published a paper in which they tried to explain this finding. Breen (2001) describes a simplified version of this framework, and states the original model had three key components:

[3] In other work, Breen (2004) has called for more balanced view regarding social fluidity versus absolute mobility

1. That there are points in all educational systems where young people must choose between options with differing levels of risk

2. That individuals strive to obtain a minimum threshold of educational attainment this is acceptable to them. They thus attempt to minimize the probability of not reaching this level

3. Young people hold subjective beliefs surrounding the probability of success in each of these risky educational options

The second of these points is perhaps the most central to the model proposed by Breen and Goldthorpe. In particular, they assume that children choose a minimum threshold of education so that they maximise their chances of securing at least the same class position as their parents. By implication this means that, as young people from different social classes differ in terms of the point where this minimum education threshold is located, they will end up making different schooling choices. This is the sociological theory of "relative risk aversion", which stresses the costs of not obtaining at least the same class position as ones parents. Breen (2001) goes on to say that this is a special case of "prospect theory" put forward by Kaheman and Tversky (1979). In particular, the model empthasises that individuals tend to see outcomes as a gain or loss relative to some reference point (e.g. their parents).

On top of relative risk aversion, Breen and Goldthorpe state two further mechanisms give rise to class differences in educational attainment. The first is that there exists "primary effects", meaning that when making educational decisions (such as whether to go to university) advantaged and disadvantaged children will differ in terms of their measured academic ability. They will, consequently, also differ in their views on the probability of educational success (e.g. they have differing subjective probabilities of their ability to successfully complete higher education). Secondly, children from the higher social classes will have access to greater financial resources to meet their educational costs (that will, for example, limit their entry into higher education). This brief overview has meant to give the reader an alternative perspective to human capital theory when thinking about the university decision making process. Testing this model is beyond the scope of this thesis. Yet the framework laid out above has a number of implications for the work that I shall present. In particular:

- A large part of the following two chapters will focus on a specific aspect of human capital theory and the importance of expected future wages and economic returns in the educational decisions of young adults. This section has served as an important reminder that this is just one of several frameworks in which one may work – and that academics from other disciplines have produced valuable research where it is other factors that take a central role

- Expectations of the future are not only important in economics, but also other disciplines. Indeed, subjective beliefs about future outcomes play an important part in the Breen and Goldthorpe model (Breen 2001 states that differences in expectations of success accentuates the relative risk aversion effect)

- Leading on from the point above, it is not only expectations of future wages that are likely to be important (as often emphasised in human capital theory), but also other subjective views such as the probability of educational success.

Indeed, I shall expand on the last of these points in the section that follows. In this I describe another model of educational decision making from the sociological and social-psychological literatures, where expectations of the future again play a key role in young peoples' scholarly attainment.

1.4 The role of adolescents' expectations in sociological models of schooling choice

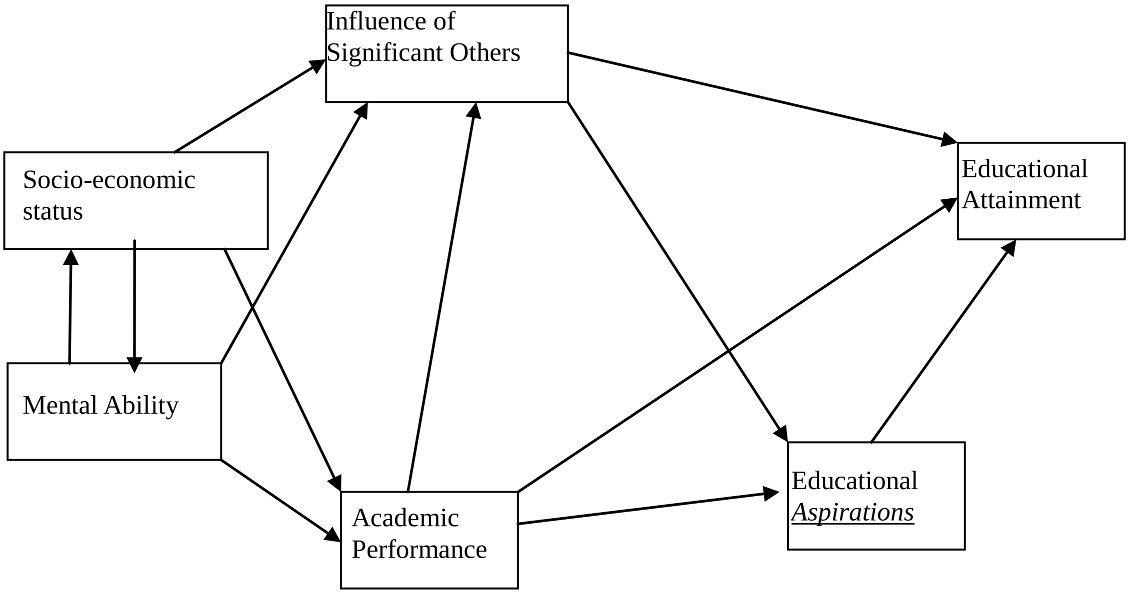
A key difference between traditional economic and sociological approaches to schooling behaviour is that the former emphasize the role of children's *expectations* (in terms of earnings returns), while the later identify *aspirations* (in terms of educational and occupational goals), as the key driver of educational attainment. Although occasionally confused in the literature, these are two distinct (but interlinked) concepts. Sociologists sometimes refer to educational expectations as “realistic aspirations” (Gutman and Akerman 2008) – a realistic assessment of the level of schooling a child believes they will achieve. This is supported by Morgan (1998) who states that adolescents' educational expectations are not “irrational fantasies”, but are rather “grounded in reality and logical thinking”. On the other hand, educational aspirations are typically thought of as what children would like to do; their desired goals and ambitions. Specifically, aspirations are thought to reflect children's 'motivational orientation' (Morgan 2005), which drives the effort they put into their schooling and willingness to complete a tertiary qualification.

A particularly famous sociological theory, which emphasises the importance of children's aspirations for their later educational attainment, is the Wisconsin model of status attainment (Sewell et al 1970). This is set out in Figure 1.2. Without going into the finer details of this model, children’s *aspirations* play a vital part in determining educational and occupational outcomes[4]. In one dimension, high academic performance is (partly) converted into educational attainment through children’s desire to complete more schooling. Yet, in another dimension, educational aspirations are a vital (though indirect) link between socioeconomic background and attainment. Specifically, children from advantaged backgrounds tend to perform better academically than their disadvantaged peers, while also receiving more encouragement from “significant others” (teachers, parents, peers) to continue their investment in education. This has a subsequent (positive) effect on their educational

[4] I have not presented the full model of Sewell et al (1970) for brevity. In particular a further outcome (occupational attainment) is specified within this model which is also partly determined by the aspirations children hold for the future.

aspirations, which drives their high attainment. Disadvantaged children, with lower levels of academic performance and less encouragement, do not hold as high aspirations for the future and thus do not generally reach such advanced levels of schooling.

Figure 1.2 Wisconsin model of status attainment



Notes:

1 Source: Sewell et al (1970), adaption of Figure 2

The sole focus of this thesis, however, is children's expectations (their realistic appraisals of the future, not their hopes and dreams). How does this fit into sociological theory?

Since regaining favour in sociological research during the 1990's, a group of contemporary status attainment researchers have begun to develop the original models of Sewell and his co-authors. A key part of this movement is Stephen Morgan (Morgan 1998, 2004, 2005, 2007) who claims (Morgan 1998) that "there have always been some rational expectations within theory of status socialization" in the sense that young adults are assumed to base educational decisions partly upon logical (i.e. well-informed and efficient) cost-benefit calculations of their future. He points, for example, to the survey questions used to capture children's aspirations in empirical representations of the model above. These typically take the form:

"how far do you expect to go in school?"

Rather than capturing children's aspirations, Morgan argues that responses represent educational expectations, and that these have a rational core, reflecting children's logical thoughts about their future[5]. Thus he moves away from educational *aspirations*, and towards children's *expectations*, as the causal mechanism within the status attainment framework. He sums up in Morgan (2005):

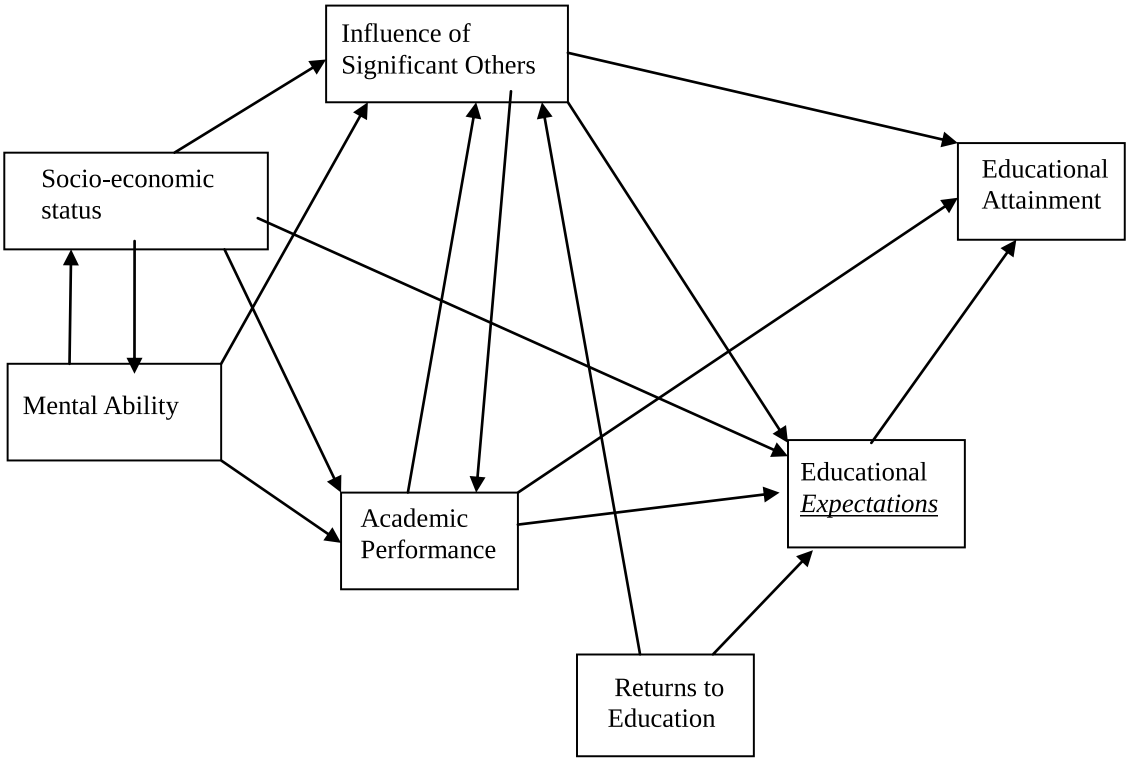
"educational expectations have an effect on educational attainment (or whatever outcome is modelled) and not just as a mediator for the indirect effects of prior variables"

In Figure 1.3, I incorporate the developments of Morgan (1998) into the status attainment model shown in Figure 1.2. The most obvious alteration is that it is now educational *expectations*, rather than *aspirations*, that help determine children's schooling outcomes. However, an additional point of interest is that Morgan includes children's expected returns to education as one of the drivers of these educational plans[6].

---

[5] Indeed, in Morgan (1998) he suggests that children's reported educational expectations vary with the marginal costs and benefits of the available human capital investments (similar to the definition of expectations applied by economists) and that this provides evidence of rational thoughts.
[6] I have not included the determinants of the expected returns to education in this framework for brevity.

Figure 1.3 Status attainment framework, combining Morgan (1998) with Sewell et al (1970)



Notes:

1 Source: Authors own diagram, based upon models presented by Sewell et al (1970) and Morgan (1988)

This model plays an important role in motivating the fourth chapter of this thesis. Specifically, coming from a disadvantaged background (low socio-economic status and little encouragement from one’s parents, teachers and peers) leads to lower educational expectations, which reduces motivation in school and increases the chances of engaging in risky teenage behaviour (drinking, drug use and early sexual activity). This eventually means that children who once performed the same academically may not go on to achieve the same level of education because of differences in such beliefs. As there is likely to be constant feedback between expectations and attainment (so that lower expectations leads to lower attainment, which leads to continued low expectations, and so forth), establishing both the strength and direction of such relationships has proven to be methodologically challenging. Nevertheless, if this theory is correct (that childhood expectations really do have a *causal* influence on later attainment) then enhancing disadvantaged children’s educational expectations should be a pressing concern for any group attempting to widen access to university or promote social mobility. Specifically, disadvantaged children's low educational expectations may be part of the reason why they are less likely to complete higher education and enter professional jobs than their advantaged peers (even if they are equally academically talented).

Based on such theories, governments across the world have introduced policies to try and ensure children hold high expectations for the future and aim for university from an early age. These schemes are particularly prevalent in the US, with examples including the “Gear-up” and “I have a dream” programmes. Yet similar initiatives have emerged during the last decade in the UK. This includes the “Gifted and Talented” scheme, which give disadvantaged children near the end of compulsory schooling (around age 16) the chance to interact with professionals and to be “mentored” by current university students. However, although there seems particular concern amongst policymakers in these two countries, I have seen little evidence to suggest that the socio-economic gap in expectations is bigger here than in other parts of the developed world. Thus, we do not know whether this is a *specific* reason why disadvantaged children tend not to participate in higher education in Britain and America, or a more general barrier to university across many different nations.

1.5. Overview

The following two chapters contribute to the small literature on young adults' subjective labour market expectations. Specifically, I begin by considering the wage expectations of university undergraduates from England and Wales; to my knowledge the first such study to take place in the UK. In this chapter I show how students' expectations vary, before considering whether the beliefs they hold are, on average, realistic. I contribute to the existing European literature by making a more reliable comparison of expectations and realisations, and in attempting to represent a wider student population. My main finding is that, on average, UK undergraduates overestimate their starting salary by around 10-15%. Yet those who take Art, Humanities and Social Science courses hold particularly unrealistic labour market expectations, overestimating their starting wage by almost 20% (£3,500). This is in comparison to maths and education students, where the figure is closer to 5% (£900).

Of particular note, I find that undergraduates who are about to enter the labour market hold more realistic expectations than their peers who have just begun university. I discuss how this may be related to the different costs that young adults face in accessing details about future possible careers. In other findings, I suggest that the accuracy of UK students' wage expectations is unrelated to their gender or ethnicity, but that it may be associated with the quality of institution they attend. Nevertheless, I stress the need to provide all young adults with detailed, yet accessible, labour market information to ensure that they are making well informed educational decisions.

I build on this work in my third chapter by using US panel data to investigate whether 20 year old American men hold realistic expectations of their annual income at age 30. This adds value to my second chapter by considering whether results hold within a different setting and over a longer time horizon, and whether young adults with direct experience of the labour market hold more realistic expectations than their student peers. I also exploit the longitudinal structure of the data to overcome methodological problems that hinder comparisons of expectations and later realisations in the existing literature. Moreover, due to the rich nature of this data, I am able to explore whether young adults' are also unrealistic about other aspects of the labour market – such as their future occupation. My results suggest that, on average, 20 year old American men make poor predictions of both their income and occupation at age 30 and that, on average, young adults in the labour force hold just as unrealistic expectations as their peers enrolled in higher education. Consistent with my findings for the UK, I estimate that young adults who graduate from a mathematical discipline overestimate their future income by an average of just 7%, compared to over 60% for those who studied Art, Humanities or Social Sciences. There is little to suggest that these results differ between those from advantaged and disadvantaged backgrounds, but I do find large and statistically significant differences between ethnic groups, and that the accuracy of young adults' labour market expectations is associated with their ability in mathematics.

In chapter four, I investigate the proportion of 15 year old children who expect to complete higher education and how this varies amongst socio-economic groups. This work is closely linked with educational policy in the UK; particularly schemes like AIMHIGHER that explicitly attempt to raise disadvantaged children's expectations of, and perceptions about, going to university. I contribute to the existing literature by exploring this topic across all the OECD countries, rather than in just a single national setting. Consequently, I am able to place the educational expectations of young adults in the UK into an internationally comparable perspective. My results show that there are large differences between advantaged and disadvantaged children's educational plans, and that this holds true across all countries in the developed world. I find that only part of this divide in children's expectations can be explained by differences in their test performance at age 15 and the schools that they attend. However, perhaps my most concerning finding is that a large proportion of 15 year olds who score highly on an internationally recognised assessment do not expect to complete university. In both England and the US, an advantaged child of scoring around the national average on this test is *more* likely to expect to complete higher education than a disadvantaged child achieving a mark in the top quintile. Indeed, I predict that less than one in two disadvantaged English children who score in the top quintile of the PISA maths test distribution believe they will obtain a bachelor's degree. Yet the same is true across many industrialised nations, with little evidence that the socio-economic gap is atypically big in the UK compared to other members of the OECD. Nevertheless, I argue that English policymakers still have a need to encourage disadvantaged children into higher education, although I suggest that widening access schemes should perhaps start at young ages.

I conclude with a summary of my findings in chapter five, reiterating the main conclusions that I draw and possible directions for future research.

Chapter 2

Wage expectations of UK students: How do they vary and are they realistic?

"The contempt of risk and the presumptuous hope of success are in no period of life more active than at the age at which young people choose their professions."

Adam Smith, The Wealth of Nations (1776), Page 109

The returns to education have been widely explored in England and Wales. Some recent studies suggest students in particular subjects receive, over their lifetime, poor financial returns to their investment in university education (Vignoles 2007). The fact students still decide to enter university and take these courses is often explained by the value of non financial benefits, such as the joy of learning and experiencing independence. However, an often overlooked possibility is that students may be unrealistically optimistic about their future opportunities when deciding to go to university, as noted by Adam Smith in the quote above. This chapter investigates the variations in students' expectations and whether these expectations are in line with wages in the graduate labour market.

This topic has a small literature within Continental Europe, although no recent academic research on students' wage expectations has been conducted within the UK. These existing European studies are severely limited by their reliance on convenience samples drawn from a small number of subjects and institutions, with results unlikely to generalise to the wider student population. The first part of this chapter attempts to address this issue by using data designed to be nationally representative (the Department for Education and Skills Student Income and Expenditure Survey) to estimate a model of students' wage expectations. I test the hypothesis that students near the beginning of their course expect significantly higher wages than those about to graduate, and investigate the impact of several characteristics relating to students and their institutions.

The second half of the chapter considers whether students hold realistic wage expectations. Existing European studies use of unrepresentative samples causes particular difficulty in comparing students' wage expectations to the actual earnings of graduates. A highly selective cross sectional survey on wage expectations is usually compared with historical data on graduate wages. Almost no attention is paid to whether the surveys are comparable, with problems such as selectivity, induced by convenience sampling or non-response bias, largely ignored. This severely hinders the existing studies assessment of how realistic students are. In comparison, this study compares the wage expectations from a national survey of students to average realisations of groups from the same cohort, drawn from an attempted census of all graduates, providing a better basis for comparison. The comparability of the two data sources is discussed, with the results checked for robustness using the Labour Force Survey. I find that, on average, full-time students overestimate their first salary after university, though this varies with the subject that they study.

The chapter begins by reviewing the current literature and describing the available datasets. A model of UK students' wage expectations follows in section 2.3, with discussion of results in relation to the seniority of the student and various background characteristics. The final two sections compare students' expectations with actual graduate wages, before a discussion of what the findings imply for European higher education policy.

2.2 Current literature and research questions

There have been a small number of studies investigating students' wage expectations across America and Continental Europe. A common theme is that students who are further through their course have lower wage expectations than those at the beginning, reflecting better knowledge of their own ability and chances in the graduate labour market. Betts (1996) finds that students do not gather information until a late stage. He therefore concludes that students near the beginning of their course have reasonably poor labour market knowledge. Brunello et al (2001) show a similar pattern in their study; students further through their course tend to not only expect lower wages, but are also less optimistic about their employment prospects. This begs the question, will the difference in wage expectations, based on the seniority of the student, remain once views of employability have been controlled for? Indeed will the same pattern be observed within the UK at all, using data designed to be representative of the undergraduate population?

A topic that has received rather less coverage is the role that university prestige plays in students' wage expectations. Smith and Powell (1990) took samples from two institutions in America that differed in quality. Students at the elite university were found to have higher wage expectations, conditional on their prior high school rank. Brunello et al (2001) looked at the expected wage gain in relation to university status, and found that only tighter admission criteria had a significant impact on expectations. However neither study draws their sample from a large number of institutions within one country. The UK is a particularly interesting setting for such research, with the number of universities having grown dramatically since government expansion of the higher education sector in 1992, creating large variation in terms of standards and prestige. Moreover, previous research in the UK suggests that the quality of an institution may influence graduates' salaries. For instance, Chevalier and Conlon (2003) find a premium of going to a prestigious UK university above and beyond the influence of ability. They conclude that this presents an economic argument for these institutions being allowed to charge higher tuition fees. However do students expect such a premium in their wages?

Many studies have also considered how wage expectations vary between students from different family backgrounds. Webbink and Hartog (2004) found students from high income families expect significantly more than those from poorer backgrounds, but that they are also more likely to overestimate their future wage. Smith and Powell (1990) also found this positive association between parents' income and students' wage expectations. One piece of work conducted in the UK by Williams and Gordon (1981) looked at the impact of socio-economic variables on the wage expectations of students at the end of compulsory education[7]. However they found that socio-economic status had little direct influence on students' expected lifetime gain from going to university. Other variables typically investigated include gender, age and the education and occupation of parents. However less attention has been paid to differences based on characteristics such as ethnicity. For instance, do ethnic minorities anticipate some form of discrimination in the labour market and therefore lower their wage expectations?

Some of these existing studies go onto make a rough comparison between students' expectations with wages in the graduate labour market. For instance, Wolter (2002) shows that students tend to overestimate their wage with a degree. Smith and Powell (1990) suggest students are well informed about average wages, but tend to overestimate their own returns. The one known European study that uses longitudinal data, by Webbink and Hartog (2004), comes to a different conclusion; students can

[7] I know of two studies based within the UK to have considered students' wage expectations. One is by Williams and Gordon (1981) and the second by Bosworth and Ford (1985). Also Brunello et al (2001) contained some information from the University of Sterling and University of Essex, though the sample sizes were very small.

accurately predict their starting salary. However a difficulty encountered in most of this work is that the authors compare a highly selective convenience sample of students' expectations with historical data on graduate wages. The characteristics of the two samples covered in each survey are usually not even discussed, even though the sampling designs mean that they relate to very different, and possibly incomparable, populations. Even Webbink and Hartog (2004) advise caution generalising results in their longitudinal survey, due to the highly selective nature of follow-up. Moreover, their study compares the salary students expect in their first job to the wages of graduates who have been employed for up to three years post university (the wage they are receiving after three years in employment). As such, no paper thus far convincingly illustrates whether students' expectations are realistic. Indeed there is some disagreement. Whereas Webbink and Hartog (2004) boldly label their paper 'Can Student's Predict their Starting Salary? YES!', Betts (1996) concludes that, on average, students overestimate their starting wage by 10%, while Brunello et al (2001) suggest the figure could be even higher than this. This chapter hopes to resolve this conflict by comparing two surveys that cover largely comparable populations with issues of non-response and comparability addressed directly and checked against an additional data source.

The results have substantial policy implications for Europe's higher education sector. In particular, as noted in the introduction to this thesis, young people may be entering university based on unrealistically high expectations of their prospects upon graduation. This may be exacerbated by government policy that highlights the magnitude of possibilities that students will have on graduation, which may actually never materialise. A further possibility is that students are willing to take on high levels of debt because they believe their future wages will enable repayment of their student loans. Gustman and Stafford (1972) show that the higher the income expectations of students, the more they tend to consume. However if their expectations are unrealistic, students may over consume during university, leading to difficulties and debt in later life. In concluding this chapter, I thus argue that policymakers need to provide young adults with more information about the graduate labour market so that such situations do not arise.

On the basis of the international literature and current policy interest, the research questions to be explored in this chapter are as follows:

1. Do students who are further from graduation have greater wage and employment expectations?

2. Do students at elite universities have significantly higher wage expectations than those at less prestigious institutions?

3. Are parental income and ethnicity associated with students' wage expectations?

4. Do students have realistic expectations? Do students who are studying a subject directly leading to a career have more realistic wage expectations than their peers who are likely to enter the wider graduate labour market?

To my knowledge, this chapter provides the first study of students' wage expectations in Europe using data designed to be representative of a national undergraduate population. The first question follows much of the existing research, but extends the analysis to show how students' views of their employment prospects influence their wage expectations. On the other hand, the second question has received little attention in the existing literature, due to the reliance on convenience samples taken from a small number of institutions[8]. Question three attempts to look at some previously neglected variables such as differences in wage expectations between ethnic groups. Finally, I investigate whether students are realistic, the first such study conducted in the UK.

---

[8] The definition of 'elite' in this work is whether the institution belongs to the 'Russell Group'; a self-selected alliance of 20 research-intensive universities. For further details see http://www.russellgroup.ac.uk/

2.3 Data on wage expectations

One reason why more research has not been done in this area is the lack of available data. The Association of Graduate Recruiters Graduate Career Survey is one possible source. However this study only targets the "top 30" UK universities, and therefore does not cover the whole UK student population, leading to an unrepresentative sample[9]. Several methodological problems also exist with the sampling strategy used and with the reliability of responses. An alternative is the 2004/5 Student Income and Expenditure Survey (SIES). This study was carried out using face-to-face interviews between January and March 2005 by the Institute of Employment Studies and the National Centre for Social Research on behalf of the then Department for Education and Skills.

The purpose of the survey was to generate a representative sample of all higher education students in England and Wales, in order to investigate income and expenditure patterns. One strength of using this dataset is that it contains detailed information on a number of potential explanatory variables. This allows analysis of potential sources of variation in wage expectations that have been neglected in previous studies. Information is provided on the students' current year of study and the length of their course, providing valuable information regarding the first research question. The number of universities included in the survey provides a large sample of students from a range of institutions. This allows a detailed investigation across both universities and subjects within one country; a further topic with little coverage in the existing literature. There is also information on students' background, including ethnicity, social class and previous schooling. Other controls such as gender and whether the student is classed as 'dependant', meaning they are in full-time education and had their parents' income taken into consideration when applying for student support, are included[10]. For 'dependant' students, there is also an approximate measure of family income, though it can only be taken as a proxy due to the way this data has been collected and recorded[11]. Unfortunately some other important

[9] A "top 30" university in this case is defined by the Association of Graduate Recruiters. The majority of universities included in the survey are Russell Group institutions, known for their excellence in research.

[10] Full details are given in Appendix 2.1 about the survey definition of this variable.

[11] Further details are given in Appendix 2.2

information is missing; in particular there is no indicator of student ability.

A complex sampling design was used to ensure a representative cross-section of students was selected (I allow for this complex survey design by making the appropriate adjustment to the standard errors in all figures that I present). Universities were sampled using a probability design based on the size of the institution. There was also stratification by region and whether it was a "Pre-1992" or "Post-1992" university[12]. A sample of 80 universities, from a population of 132, was drawn, with probability proportional to size. In total, 69 universities agreed to take part with the intention of contacting 240 randomly selected students from each institution. Separate samples of full-time and part-time students were drawn, with special previsions made for those institutions with medical schools[13]. 25 Further Education Colleges (other degree awarding institutions) were also approached, with 19 entering the final sample. From each of these institutions, 60 students were randomly selected. Across all institutions, a total of 16,524 students were selected to take part. These students were each mailed an initial "opt-in" questionnaire, where they were asked to provide some basic information. 7,548 (45%) opt-in questionnaires were returned, with 5,810 (35%) agreeing to participate in the study. In total 4,570 names were issued with 3,548 interviews achieved[14]. I choose to drop students who did not report their expected starting salary, along with those studying at further education colleges or for qualifications other than at degree level. I exclude a further 142 (5%) observations where the expected starting salary was below £8,000[15]. The final dataset contains 2,659 observations, with the sample selection rules presented in Table 2.1.

[12] A "post-1992" university is an institution that achieved university status in 1992 or later. This date marks a major change in the UK higher education sector, when several polytechnic institutions were given degree awarding powers. This increased the number of students at universities dramatically.
[13] Further details can be found in the 2004/2005 SIES technical report (Finch et al 2006b)
[14] Another institution that mainly involves part-time distance learning, The Open University, was in the original dataset but was dropped as these students did not give details on their wage expectations.
[15] Almost a quarter of these values were at £1, and thus largely reflect illogical answers. Results were also checked for robustness using £3,000 as the minimum allowed expected salary. When this is done, all the substantial conclusions in the following sections remain intact.

Table 2.1 Sample selection rules

Rule	Sample remaining
Initial sample	3,548
Observations where salary expectations missing dropped	3,375
Further Education colleges dropped	3,170
"Other" degrees dropped	2,791
Expectations below £8,000 dropped	2,659

Notes:
1 This table illustrates observations dropped from the analysis because of item non-response or my sample selection criteria. It does not include information on unit non-response, which is described in the text above.
2 The selection rules are applied cumulatively, hence the final figure of 2,659 refers to when all four selection rules are applied (i.e. I drop 173 observations from missing expectations data, then a *further* 205 where respondents attend further education colleges (and so forth)).
3 Source: Authors calculations from the SIES data

The level of non-response is not negligible, and obviously has implications for the generalisability of results. Those that take part in the survey could be systematically different to those who opt out. To address this, the survey organisers modelled the probability of student response using the rich data available on the sampling frame and other auxiliary information. These data included the students' age, gender, previous qualifications (e.g. the number of A-Levels sat), quality of institution they attend and whether their parents went to university, amongst other things. Estimates from this model were used to create sampling weights that attempt to correct for the probability of a student being selected and responding. A second stage of weighting was also conducted to ensure the sex and age profile of students matched that of Higher Education Statistics Authority records[16]. It is evident that significant effort has been put into investigating and correcting any bias in the sample, to ensure it represents the student population in England and Wales (though it should be noted that the use of sample weights can only correct estimates in terms of observable characteristics). Nevertheless, the SIES 2004/2005 report (Finch et al 2006a) proceeds to state that:

[16] Weights were calculated as the *inverse* of the probability of being both selected and responding to the survey, and were the product of five conditional probabilities. Further details are provided on non-response and weighting in the 2004-2005 Student Income and Expenditure technical report (Finch et al 2006b).

“As can be seen, this was an ambitious methodology but one which succeeded in producing the objective of a nationally representative student sample for interviews.” (P 10)

Indeed, in comparison to most of the studies on wage expectations discussed in section 2.2, the SIES data has the advantage that it is designed to draw a representative sample from the population of students, rather than relying on a simple convenience sample. Moreover, data is drawn from around 70 institutions across the whole spectrum of subjects, whereas most of the existing research can only boast a handful of subjects from a couple of universities. Although non-response does cause some limitations, it is reasonable to say the SIES is more likely to be representative of the wider student population than any previous study and therefore provides a better source for analysis.

Another critical part of the survey is how students report their wage expectations. They were each asked the following question:

“What sort of salary do you expect to be earning in the first job you take once you have graduated?”
Interviewer comments: If not sure of the exact amount, please give your best estimate.

Students are asked for their expected salary, to be recorded in an open text field, allowing precise estimates to be made. This is interpreted as students giving the mean of all possible outcomes they face. In other words, it is assumed that students are providing the arithmetic mean for the entire distribution of all possible outcomes[17]. A further issue is that the question asks students about the first job they take after university. Students are not asked explicitly whether they expect this to be full-time or part-time work, or if this will be temporary while they look for a job directly related to their career aspirations. Nevertheless, Manski (1996) suggests that students interpret questions regarding future salary expectations on the assumption that they will be in full-time employment. Thus it seems reasonable to assume expected salary

[17] Ideally, a precise definition would be provided to the students as Manski (2004) suggests when eliciting students’ median expectation of their future wage distribution. However I feel my assumption that students are providing the arithmetic mean for the entire distribution of all possible outcomes they face is reasonable reflection of how students would interpret the question posed.

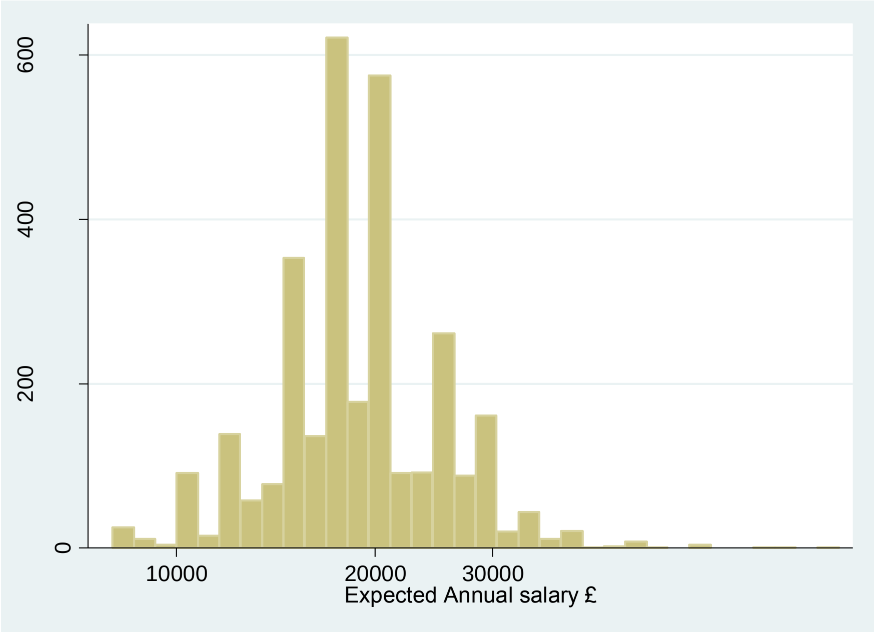
corresponds to students' first full-time job after university.

When interpreting the data it is also assumed that students are providing a gross, yearly figure. Although ideally this would be made explicit in the question, it seems reasonable to assume students would report figures in this way, as it is the standard method of advertising salaries in the UK. Assumptions must also be made about how students deal with inflation when forming their wage expectations. The most common assumption is that students do not consider inflation, and are thus reporting in 2005 prices, as discussed in Wolter and Zbiden (2002) and Brunello et al (2001). This is the approach also taken in this chapter.

Brunello et al (2001) and Dominitz and Manski (1996) also note that respondents tend to round their estimates to questions surrounding expectations to the nearest 5 or 0. The histogram of expected salaries in Figure 2.1, along with the data presented in Table 2.2, shows that the distribution of log expected wages is broadly symmetric, though there is bunching of estimates to the nearest £1,000. Large spikes are especially prelevant at multiples of £5,000 (for instance £15,000, £20,000 and £25,000), with other instances at £12,000 and £18,000 (equivalent to a salary of £1,000 and £1,500 per month). Brunello et al (2001) go on to say that there is no evidence to suggest students do not take care when completing their questionnaire[18]. However no existing study tries to take this explicitly into account. This chapter goes a stage further and checks the robustness of results to this heaping in the data. A description of the techniques used can be found in Appendix 2.3.

[18] Manski (2004) finds a similar phenomenon when eliciting individuals subjective probabilities, where individuals round to the nearest 5% or 10%.

Figure 2.1 Distribution of log expected starting salary



Notes: 1 Source – SIES. Sample size 2,659

Table 2.2 Peaks in the distribution of expected starting salary

Expected salary £000	Frequency	%
8	22	1.4
10	75	2.7
12	129	4.7
13	43	1.6
14	68	2.5
15	304	11.1
16	115	4.2
17	135	4.9
18	364	13.3
19	92	3.4
20	473	17.2
21	37	1.4
22	70	2.6
23	35	1.3
24	40	1.5
25	202	7.4
26	30	1.1
27	26	1.0
28	49	1.8
29	10	0.4
30	126	4.6
35	28	1.0
40	16	0.6
50	8	0.3

1 Refers to unweighted figures, 2 Source – SIES

A final point that often concerns economists when using subjective data is that respondents have little incentive to report their expectations accurately. A recent study by Botelho and Pinto (2004) tested this issue in an experimental setting, and found financial incentives have little impact on the accuracy of students' reported expectations. Moreover Manski (2004) puts over compelling evidence on the measurement and use of expectations in economic research. These studies conclude that economists' scepticism of subjective data is largely unwarranted.

I proceed by using this data to explore the factors associated with students' wage expectations via a multivariate OLS regression model. The purpose of this model is to attempt to answer research questions one to three. Ideally, one would want to control for all the factors that potentially confound the relationship between students' wage expectations and the variables of interest (i.e. the variables that relate to the research questions I have posed). So, for instance, one would want to try and control for student ability, as more talented pupils should expect to achieve higher wages, but my also disproportionately come from certain groups (e.g. households with higher income). Likewise, to explore differences between whites and ethnic minorities, one would want to account for the different social positions and educational choices made by these groups. Hence the need to include various background characteristics into the model, such as age, gender, subject and university entry qualifications. One may also want to include some subjective measures of respondents' views on the graduate labour market into the model. For instance, certain groups (e.g. those in their final year) may not be as positive about their chances of employment (because, for example, the realities of the labour market have struck them) which mean they lower their wage expectations. It may also be of interest to see how results differ with and without the inclusion of such subjective factors, in order to gain a better understanding of why wage expectations might vary between certain groups. Yet, on the other hand, it may be that those with higher wage expectations also hold higher reservation wages, which makes them less optimistic about their post university prospects (making this variable potentially endogenous). Hence whether one includes variables like "students' views on the labour market" into the model also depends upon concerns about endogeneity (and the limitations that this puts on the inferences one can draw). Taking the above into account, I estimate several different specification of the regression model below:

$$\text{Log}(W_{ij}) = \alpha + \boldsymbol{\beta}\mathbf{X}_i + \boldsymbol{\psi}\mathbf{P}_\mathbf{i} + \text{٣}\mathbf{U}_\mathbf{j} + \boldsymbol{\lambda}\mathbf{V}_i + \xi_{ij}$$

with W = Student's expected wage

X = Vector of background characteristics (e.g. gender)

P = Vector of dummy variables reflecting how many years left to completion of their course (i.e. their "proximity to graduation"). Reference - final year.

U = A vector of dummy variables reflecting university type (Ref: – Post-1992).

V = A vector containing two sets of dummy variables reflecting student's view on their employability (described in more detail below).

ξ = error

i = for individual i

j = for university j

In this model, "Proximity to graduation" is a set of dummy variables reflecting how many years the student has left until the end of their course (1 year, 2 years or in their final year as the reference). University type refers to the quality of university the student attends. This is defined as either "Russell Group" (old, research intensive institutions), "other Pre-1992" (old but generally smaller and less research intensive institutions) or "Post-1992" (modern, teaching based institutions), with the latter as the reference group. The vector V refers to two variables that capture students' views on their employability. Specifically, they were asked to what extent they agree with the statement that 'the growing number of graduates will make it hard to get a graduate job', which I code into three categories (agree, neutral and missing) and include into the regression as a set of dummy variables ('agree' as the reference). Students are also asked about their post-university plans, where they were provided with various options (e.g. get a career job, take temporary work, further study, travel). I recode this information to create four groups indicting what the respondent has considered doing after they graduate: get a career job (reference), take temporary work, undecided between career job and temporary work, further study or travel. Other variables include the total income of the student, how much the student earned from work that academic year and their highest qualification upon entry to university (A-levels, GNVQ, other).

Three specifications of this model will be fitted. Specifically, I will exclude the vector $\mathbf{V}_i$ that refers to students views of their employability in the first specification, and then include it in the second[19]. In these two initial models, I will allow for the complex survey design (clustering and stratification) by making the appropriate adjustment to the estimated standard errors. In the final specification, I will then estimate a fixed effect model to test the robustness of previous results (and as an alternative way of dealing with the complex survey design employed in the SIES)[20]. In all specifications, the error term is assumed to be normally distributed. Summary statistics for the explanatory variables, and their relationship with wage expectations, are presented in Table 2.3 (this, and all the following analyses, use the sample weight provided in the SIES dataset[21]).

[19] One may suggest the variables in matrix $\mathbf{V}_i$ are potentially endogenous. In this model, I assume that if individuals are less optimistic about the labour market, they will lower their wage expectations. However, another possibility is that students who expect higher wages, and who possibly also have a higher reservation wage, limit their labour market opportunities and are therefore less optimistic about their post university prospects. Leaving $\mathbf{V}_i$ out of the first specification gives an indication of results if one considers this possible endogeneity to be a problem.

[20] I note that, in doing so, certain parameters from the initial specifications (e.g. the dummy variables relating to university type and university location) will drop out of the model and not be estimated

[21] Table 2.3 contains the unweighted sample size, while the summary statistics are provided using the weights. The weights have a substantial influence on the gender composition, increasing the proportion of men from 33% to 46%.

Table 2.3 Distribution of expected wage for a selection of groups

Category	Observations (Unweighted)	P90 (£000)	P10 (£000)	Mean (£000)	Median (£000)	Standard deviation (£000)
Proximity to graduation						
Final year	916	25.0	12.0	18.1	18.0	6.4
1 year	728	25.0	12.5	18.6	18.0	6.2
2 or more years	1,015	26.0	13.0	19.4	19.0	7.7
Gender						
Male	854	27.0	13.0	19.5	19.0	7.5
Female	1,805	25.0	12.0	18.1	18.0	6.5
University type						
Russell	593	30.0	13.0	20.0	20.0	7.3
Other Pre-1992	502	25.0	12.0	18.0	18.0	7.1
Post-1992	1,564	25.0	12.0	18.4	18.0	6.5
Parents income (£ per annum)						
Below 20,000	356	23.0	12.0	17.0	17.0	5.4
20,001-40,000	559	23.0	12.0	18.0	18.0	6.1
40,001+	635	25.0	12.0	18.8	18.0	6.3
Dependant student/ No data[1]	1,109	29.5	15.0	20.0	19.5	7.8
Ethnic Group						
White	2,260	25.0	12.0	18.5	18.0	6.9
Asian	134	28.0	12.0	19.9	20.0	6.5
Black	110	30.0	14.0	20.9	20.0	7.4
Mixed	155	30.0	14.0	20.0	19.0	6.2
All groups	**2,659**	**28.0**	**12.5**	**18.6**	**18.5**	**6.5**

Notes:
1 See Appendix 2.2 for further details on parental income
2 The sample size given is unweighted. The summary statistics are provided after applying the SIES sample weights
3 Source: Authors calculations from the SIES data. Sample size 2,659

2.4. Proximity to graduation

A number of existing studies have found junior students to be more optimistic than those nearing graduation. Betts (1996) concluded that students lowered their expectations due to "learning effects", where individuals discover more about their ability and the labour market as they move through tertiary education. Brunello et al (2001) found similar results, identifying senior students to be less optimistic about wage levels and employment prospects. The initial research question follows these studies and asks if students further from the labour market are more positive about their employment prospects and have higher wage expectations. This is then extended by investigating whether wage expectations still differ after controlling for students' views on their post graduation prospects.

Full results are provided in Table 2.4. The first specification follows the traditional approach in the literature and does not contain students' views of their employability after university (**V**) as explanatory variables. These are then included in the second specification[22]. Finally, I also include a university level fixed effect.

Table 2.4a Regression results

	Specification 1		Specification 2		Fixed effects (Specification 2)	
	Coefficient	**Standard error**	**Coefficient**	**Standard error**	**Coefficient**	**Standard error**
Future plans (Ref: Career job only)						
Temporary job only	-	-	-0.224*	0.026	-0.219*	0.025
Either a career or temporary job	-	-	-0.126*	0.033	-0.116*	0.031
Further study or travel	-	-	-0.007	0.013	-0.006	0.013
Hard to get graduate job (Ref: Agree)						
Neutral	-	-	0.013	0.018	0.015	0.017
Disagree	-	-	0.056*	0.014	0.054*	0.013
Missing	-	-	0.078*	0.031	0.069*	0.031
Proximity to graduation (Ref: Final year)						
1 year	0.030*	0.014	0.013	0.014	0.009	0.014
2 or more years	0.066*	0.017	0.034*	0.017	0.031	0.017
University type (Ref: Post-1992)						
Other Pre-1992	-0.006	0.019	-0.003	0.017	-	-
Russell group	0.074*	0.020	0.065*	0.019	-	-
Parents earnings (Ref: Below £20,000)						
£20,001-£40,000	0.053*	0.022	0.046*	0.021	0.045*	0.021
£40,001+	0.071*	0.024	0.065*	0.022	0.058*	0.022
Independent student or missing data	0.119*	0.030	0.094*	0.029	0.094*	0.029
How parents earns (Ref: Work)						
Benefits	-0.052	0.046	-0.054	0.042	-0.063	0.042
Investments	0.118*	0.031	0.109*	0.028	0.117*	0.029
Ethnic group (Ref: White)						
Black/Asian	0.064*	0.025	0.063*	0.024	0.059*	0.024
Mixed/Other	0.033	0.024	0.025	0.024	0.018	0.023

Notes:
1 * Indicates significance at the 5% level
2 Results have been split into two tables. Table 2.4a contains variables directly relating to the research questions posed, while 2.4b contains the other control variables.
3 A chow test was conducted, but provided no evidence that, the results should be reported separately for men and women.
4 A fixed effects model has also been developed for specification 1. All substantial results remain in place, though the coefficient for students one year away from graduation is reduced to 0.025 and is outside the range of statistical significance.
5 Where a dash appears in the fixed effect specification, it indicates that the parameters were not estimated due to colinearity
6 Source: Authors calculations based on the SIES data. Sample size in all regressions is 2,659. Dependent variable is the natural logarithm of students' expected wages

[22] The results and interpretation presented in this section are robust to each of the subsequent specifications used, including the introduction of fixed effects.

Table 2.4b Regression results continued

	Specification 1		Specification 2		Fixed effects (Specification 2)	
	Coefficient	Standard error	Coefficient	Standard error	Coefficient	Standard error
Total income						
Mean centred (per £10,000)	0.038*	0.017	0.038*	0.018	0.030*	0.018
Study mode (Ref: Full-time)						
Part-time student	-0.047	0.026	-0.044	0.025	-0.039	0.026
Earnings from work						
Mean centred (per £10,000)	0.026	0.022	0.026	0.021	0.030	0.020
Part-time student *Earnings from work	0.166*	0.032	0.152*	0.031	0.150*	0.030
Subject area (Ref: Medicine)						
Allied To Medicine	-0.206*	0.033	-0.192*	0.031	-0.201*	0.029
Sciences	-0.271*	0.033	-0.215*	0.031	-0.225*	0.030
Maths, Computer Science	-0.209*	0.037	-0.171*	0.038	-0.194*	0.037
Engineering, Technology	-0.170*	0.034	-0.153*	0.033	-0.166*	0.029
Architecture, Building	-0.233*	0.034	-0.209*	0.034	-0.210*	0.034
Social Studies	-0.233*	0.032	-0.189*	0.030	-0.199*	0.027
Law	-0.172*	0.044	-0.135*	0.041	-0.154*	0.041
Business	-0.230*	0.034	-0.182*	0.034	-0.184*	0.033
English, Languages, Classics	-0.302*	0.039	-0.247*	0.037	-0.260*	0.034
History, Philosophy	-0.326*	0.042	-0.264*	0.043	-0.277*	0.042
Arts	-0.362*	0.037	-0.322*	0.035	-0.321*	0.033
Education	-0.220*	0.032	-0.202*	0.030	-0.225*	0.028
Combined	-0.275*	0.039	-0.229*	0.038	-0.242*	0.037
Other	-0.309*	0.054	-0.268*	0.051	-0.265*	0.049
Entry qualification (Ref: A-levels)						
GNVQ/AVCE	-0.060*	0.023	-0.068*	0.024	-0.064*	0.023
Other	-0.010	0.015	-0.017	0.015	-0.016	0.015
Age						
Mean centred	0.013*	0.005	0.014*	0.005	0.015*	0.005
University location (Ref: Other England)						
London	0.091*	0.018	0.093*	0.020	-	-
Wales	-0.057*	0.022	-0.059*	0.021	-	-
Gender (Ref: Male)						
Female	-0.053*	0.013	-0.053*	0.013	-0.050*	0.012
University dummies	-	-	-	-	X	X
Constant	10.032*	0.042	10.041*	0.041	10.034*	0.061

Notes:
1 * Indicates significance at the 5% level
2 X indicates that university dummy variables have been included but values not reported
3 Source: Authors calculations using the SIES data. Sample size in all regressions is 2,659. Dependent variable is the natural logarithm of students' expected wages

Initial results, using the first specification, support the existing studies. Students one year away from graduation expect around 3% more on average than final year students, while those two or more years away expect around 7% more.

An interesting extension is whether students who are further from the labour market also hold more positive views on their employability after university. Two variables within the SIES shed some light on this issue. Firstly, respondents were asked whether it will be hard for them to get a graduate job. In particular, students were asked to what extent they agree or disagree with the statement:

'The growing number of graduates will make it hard to get a graduate job'

Secondly, students were asked about their post university plans, including whether they have considered taking a temporary job[23]. These two questions provide a reasonably good indication of student views on their employability. Tables 2.5 and 2.6 contain cross-tabulations between these variables and how close the student is to graduation[24].

Table 2.5 Student response to whether they believe 'the growing number of graduates will make it hard to get a graduate job'

	Final year %	1 year %	2+ years %
Strongly agree or agree	56.5	46.4	48.0
Neutral	14.5	15.4	16.4
Strongly disagree or disagree	26.5	34.6	33.2
Observations	916	728	1,015

Notes: 1 Source: Authors calculations using the SIES dataset. Total sample size is 2,659.

Table 2.6 Response to whether students have considered getting a temporary job

	Final year %	1 year %	2+ years %
No	90.6	94.9	98.1
Yes	9.4	5.2	1.9
Observations	916	728	1,015

Notes: 1 Source: Authors calculations using the SIES dataset. Total sample size is 2,659.

[23] Students were asked if they planned to get a job related to their future career, a temporary job, continue studying or go travelling. They could identify more than one option; therefore this gives a rough indication of students' future plans.
[24] I also estimated a set of logistic and ordinal regressions to investigate the association between students' background characteristics and the responses to these two questions. However, cross-tabulations have been presented rather than the results from these models for ease of interpretation, with little evidence lost in relation to the research question posed.

Both sets of results indicate final year students generally hold more negative views. Only 2% of students who are two or more years away from graduation considered taking a temporary job after university, compared to nearly 10% of those in their final year. This could be interpreted in several ways. It is suggested that final year students are less positive about their labour market prospects. However an alternative explanation could be that these students are looking to delay the coming of the “real world”. Yet, Table 2.5 shows that a large proportion of final year students, compared to those one or two years away, agree or strongly agree that the growing number of students will make it hard to get a graduate job. The combined evidence does indeed suggest final year students are less optimistic about the graduate labour market. This may be due to the fact that they probably know more about their expected grade, and link this to their employability[25]. Alternatively, given the survey is conducted between January and March, they may well have already started their hunt for a graduate job, and have thus far been unsuccessful.

A question that is ignored in the current literature is whether, after controlling for views on employability, students at the beginning of their course still expect a higher starting wage than those in their final year? The two additional variables, analysed in the cross-tabulations, are included in the second model specification. The impact of being further from graduation on wage expectations has been significantly reduced. Students who are a year away from graduation now only expect wages 1.4% higher than their final year peers; a difference that is not statistically significant. Previously students two years or more away from graduation expected a 7% premium compared to final years. This almost halves to 3.5% when their future plans and opinions about the graduate labour market are taken into consideration.

[25] Indeed a recent paper by Chevalier et al (2009) indicates that first year undergraduate students over-estimate their ability. It could be that as students move through university they learn more about their ability and alter their expectations on the receipt of this new information.

This pattern could represent either a cohort or age effect. Given that other research offers similar results, it seems reasonable to suggest this represents a changing of students' views as they progress through university, rather than a difference between particular cohorts. At first, it seems there is a large difference in wage expectations between year groups. However students appear to differ in their views of the labour market in at least two aspects, namely their employability and the wages on offer. Once views of employability have been controlled for, the difference in wage expectations appears to be reasonably small.

2.5 University prestige, family background and ethnicity

I begin this section by exploring how students' wage expectations vary with the quality of institution they attend. One fundamental drawback in the existing literature is the lack of available data drawn from a range of institutions within one country. Convenience sampling, generally of small sizes, also makes research on background characteristics rather limited. On the other hand, the SIES contains information from students at almost 70 institutions, making it possible to look into the association between wage expectations and institutional quality. Moreover, the UK is a particularly interesting setting to investigate the effect of institution on wage expectations, due to its large and expanding higher education sector and the drive to widen participation. My hypothesis is that those studying at older, more prestigious universities have higher wage expectations. Caution is required, however, when interpreting the results that follow. In particular, due to the SIES data not containing any information on students' academic history, I can not separate out the "effect" of institutional quality on wage expectations from the ability of the students who attend. In other words, the parameter estimates for the institutional dummies presented in Table 2.4 will suffer from omitted variable bias. I can therefore only show the association between going to a higher quality institution and wage expectations, and <u>not</u> the causal effect of the former on the latter.

When exploring this association, I divide universities into three broad groups: Post-1992, Pre-1992 and Russell Group institutions (this is one common way to classify higher education providers in the UK). The first group refers to institutions that gained their university status after 1992 and are generally not research intensive. They also tend to admit, on average, students with lower university entry scores (i.e. lower A-Level grades). The latter two groups gained their university status before 1992, are generally more research intensive and tend to admit students with higher levels of academic attainment. The "Russell Group" institutions are a particular subset of this "Pre-1992" groups, referring to a self-selected alliance of the UK's largest and most prestigious institutions (e.g. it includes Oxford, Cambridge, Warwick, London School of Economics etc). It is this group of institutions that I take to indicate the highest quality.

The results from the Table 2.4 indicate that students at other Pre-1992 universities expect very similar wages to those at Post-1992 institutions. This is despite the fact that the former tend to admit higher ability students and generally have a better reputation within the UK higher education sector. However, as expected, students at Russell Group universities expect significantly more than the other groups – roughly 7% more in their first salary after finishing university. Yet I again remind the reader that this result needs to be interpreted carefully. Specifically, as there is no measure available on students' academic ability, these estimates will suffer from omitted variable bias. Consequently, these results only show the association between institutional quality and wage expectations (unconditional on "ability"), rather than a causal "effect".

Next, I turn to the background of the student, in particular the association between students' wage expectations and their socio-economic background (measured via parental income) and ethnicity[26]. Regarding the former, it is hypothesised that parents are a critical source of labour market information, and thus that students' wage expectations will increase with family income. For the latter, ethnic minorities may, for example, expect to suffer some form of discrimination in the labour market and

[26] It should be remembered that information available on parental income is likely to suffer a non-trivial degree of measurement error. Further details are available in Appendix 2.2. This information is also only available for those classed as "Dependant" students (59% of the total).

thus, ceteris paribus, have lower expectations. Alternatively, Rouse (2004), notes that ethnic minorities in the US tend to be overly ambitious. Is this the same in the UK?

The results in Table 2.4 are consistent across all specifications and support the hypothesis that students from richer backgrounds have higher wage expectations. All groups presented expect significantly more than those with parents earning below £20,000. There is also a monotonic trend; the higher the income group, the greater the wage expectation[27]. It could be that students from a rich background expect this high salary in order to maintain a high standard of living. Alternatively as university participation rates have increased over time, students may use their parents' income as a lower bound, and expect a higher salary due to their better education. Another possibility is that students may think their parents have connections in the labour market that will help them secure a lucrative job. However this variable could also be reflecting unobserved factors, such as parents' influence on intelligence and work ethic, which are also correlated with wage expectations.

On the other hand, the hypothesis that students from a minority background may expect to suffer some form of discrimination in the labour market does not seem to hold. Results suggest that Black and Asian students expect a significantly higher starting wage than white students[28]. This conclusion holds within all specifications. Hence it seems that ethnic minorities do not expect to suffer discrimination in the graduate labour market. Indeed quite the opposite appears to be true, these groups tend to be more optimistic about their future earning potential than their white peers.

[27] Out of the 1838 dependant students, 164 (9%) did not report a figure for parents income. These students had higher wage expectations than the highest category included in the regression (those with parents earning over £40,000). A logistic regression, not presented, was carried out to investigate if certain groups are more likely to not report a figure for parents' income. The results of the logistic regression suggested that students who have parents generating most of their income from pensions or investments are less likely to report a figure. To the extent that this exhibits wealth, for instance early retirement or being able to live off investments without working, the result fits quite well with the observed pattern; the better off the students' parents, the higher their wage expectations.

[28] The Black and Asian groups were combined due to small sample sizes within each. In initial regressions the two groups were entered separately, producing similar coefficient estimates and standard errors, significantly different from Whites at the 10% level. A test was performed of whether Black and Asian have equal regression coefficients, resulting in the null hypothesis not being rejected (one should, however, exercise caution when interpreting this lack of significance precisely because of the small number of observations within the two groups).

2.6 Data on realised wages

The preceding analysis illustrates that there are quite large differences in students' wage expectations. Yet this may simply reflect the different labour market opportunities that students face. On the other hand, it may be that certain groups of students hold unrealistic wage expectations and significantly overestimate starting salary. Manski (2004) describes various ways in which expectations and realisations can be compared, with longitudinal data the most direct method of comparison. Unfortunately, there is almost no suitable panel data available across Europe and none in the UK[29]. Thus attention is turned to Manski's second method; using repeated cross sectional surveys to evaluate average expectations and realisations. Therefore a second data source is needed that contains information on wages in the graduate labour market for the same cohort of students. Data on graduate wages, corresponding to the same year, is drawn from the Higher Education Statistics Authority (HESA) Destination of Leavers Survey (DLHE). This section describes the additional data source and methods for comparing the two surveys.

The DLHE is an attempted census of all 2004-2005 graduates' employment circumstances, including their current salary, six months after completing university. Students are contacted directly by the institution they studied at by postal questionnaire, with non-respondents followed up in a telephone interview. This results in a survey response rate of around 80%. The results are then linked with administrative data about the student collected by HESA, providing a rich source for analysis. Variables within the dataset include socio-economic status, university entrance (UCAS) score, degree classification, subject of study and where the student lived while at university.

The target of this survey is obviously a great deal wider than that of the SIES. Several sample selection procedures, available in Table 2.7, were applied to the data to ensure the two sources were comparable.

[29] Even the one longitudinal study by Webbink and Hartog (2004) suffers quite substantial methodological problems. Their study suffers both high non-response to the survey and missing data on the wage variables. Moreover they show that the missing wage data comes disproportionately from groups who are likely to have lower wages in the graduate labour market. Likewise, they compare university students' starting salary expectations to the wages of graduates with up to three years labour market experience.

Table 2.7a Sample selection rules for full-time students

Sample selection rules	DLHE (realisation) sample size	SIES (expectations) sample size
Start	256,507	2,659
England & Welsh universities only	224,226	2,659
First degree only	180,911	2,393
Salary above £8,000	178,491	2,339
Employed within UK only	168,673	2,339
Over 24 excluded	145,517	1,923
Doing "something else" (e.g. further study/travel) excluded	117,660	1,923
Only those working full-time	87,327	1,923
Missing salary data excluded	45,906	1,828
Medics excluded	44,436	1,666
Final	**44,436**	**1,666**

Notes:
1 "First degree" in the DLHE survey means excluding those in postgraduate study and foundation courses. The SIES collected expectations data only from those doing a first degree or a foundation course and not from postgraduate students.
2 Source: Authors calculations using the DLHE and SIES dataset.

Table 2.7b Sample selection rules in the DLHE for part-time students

Sample selection rules	DLHE sample size
Start 2006 prices	59,965
Start 2005 prices	59,965
England & Welsh universities only	52,606
Postgraduate students excluded	28,416
Salary above £8000	28,416
Employed within UK only	27,353
Doing "something else" (e.g. further study/travel) excluded	24,219
Missing salary data excluded	9,842
Medics excluded	9,834

Notes: 1 Source: Authors calculations using the DLHE dataset.

As in the SIES, only students who attended a university in England and Wales were considered. The data was also restricted to only those students who had finished their first undergraduate degree; with those completing postgraduate study, or university courses below the level of bachelors degree, excluded.[30] It is assumed that respondents to the SIES were reporting their first wage expected after their undergraduate degree. However it is impossible to rule out the possibility that some

[30] In the UK this includes HND, HNC, foundation and access courses, amongst others. These are generally thought of as a qualification below degree level.

students reported their wage expectation under the assumption that they were going to continue in full-time education and gain a post graduate qualification. Only students who reported salaries of £8,000 or more and were working full-time are included[31]. Since students were asked for their full-time annual equivalent wage, £8,000 was identified as the lower bound for logical responses due to minimum wage laws in the UK[32]. The same rule is also applied to the SIES data, so that only expected and actual wages above £8,000 are considered. Moreover, to limit the potential influence of previous labour market experience, and to target the particular group of interest, in both surveys the sample was restricted to those below the age of 25. This has also been done so that an additional data source, the labour force survey, can be used as a check for robustness of results[33]. Finally, Medical students have been excluded from many parts of the analysis (i.e. when I investigate average expectations and realisation across all students) due to the different proportion of these students in the SIES compared to the DLHE[34]. I do, however, return to this group (medics) when looking at breakdowns by subject (later in this section).

The DLHE has many features that make it a strong candidate to compare with the SIES data. The information was collected in January 2006 and specifically refers to graduates who were final year students in 2004-2005 when the SIES was conducted. However in this analysis, to ensure a reasonable sample size, wage expectations from students in all year groups are used[35]. It is again assumed that students in all years report their expectations in current (2005) prices[36,37].

---

[31] Dominitz and Manski (1996) shows that students tend to report their expectations conditional on working full-time.

[32] The adult minimum wage at this time was £5.05. Assuming the minimum amount of time required in a full-time job is 30 hours per week, for 52 weeks a year, this generates a full-time annual income of around £7,900. Only around 1% of observations were dropped from the DLHE using this sample selection rule, with 2% dropped from the SIES.

[33] The wording used in the labour force survey means only those under 25 can be considered. This is discussed later in this section.

[34] Initial analysis, not presented, suggested a difference in the proportion of medical students contained in the two surveys (8% of the SIES compared to 2% of the DLHE). Thus medical students will be excluded in many parts of the analysis, due to the difference between the two surveys and the quite different labour market these individuals face.

[35] Indeed it may be argued that I am not using repeated cross sectional surveys in the strictest definition. In part, I am assuming that all students in the SIES face the same distribution of actual wages, represented by the salary recorded for the graduating 2004/05 cohort. However it seems highly likely the distribution of graduate wages will remain stable considering the short space of time. Manski (2004) discusses this assumption in more detail.

[36] Since the SIES was conducted in early 2005, while the DLHE recorded actual wages in 2006, an adjustment has been made for inflation. The wages in the DLHE were scaled back to 2005 prices using the Retail Price Index (2.8% for the year in question).

The questions posed in each survey also relate closely to one another:

SIES

"What sort of salary do you expect to be earning in the first job you take once you have graduated?"

Interviewer comments: If not sure of the exact amount, please give your best estimate.

DLHE

"What was your annual pay to the nearest thousand £, before tax?"

If you were employed less than a year or were part-time, please estimate your pay to the full-time annual equivalent.

The SIES asks about salary expectations in students' first job after graduation and the DLHE records information on salary sixth months after finishing university. In the vast majority of cases, the difference between these two definitions is likely to be minimal. Previously it was stated that students in the SIES are thought to provide estimates of a gross, annual salary. The DLHE survey asks students to provide an estimate for their full-time equivalent annual wage before tax, providing a closely matched definition. A final issue is that the DLHE survey asks students for their wage to the nearest thousand, while expectations in the SIES were recorded in an open text cell[38]. However section 2.3 described how students' expectations tend to bunch around the nearest thousand, meaning this is unlikely to induce any substantial bias.

Yet, for all the benefits of the DLHE, there are some difficulties with response rates and the selectivity of respondents. Although the DLHE is an attempted census of graduates, there is quite a large degree of non-response to the question about salary[39]. Moreover, the salaries for those students who go into postgraduate courses, or those

---

[37] Ideally, students would have been formally instructed not to consider inflation in the wording of questions, as in Dominitz and Manski (1996). They report that students generally adhere to this, and do not consider inflation in their wage expectations. Moreover Brunello et al (2001) use similar wording to the SIES, in that students are not directly informed how to deal with inflation. They also assume students report their expectations in current prices, and find inconsistencies in their data with the idea that respondents make an adjustment to their responses to try and account for inflation.

[38] The mid-point of this band has been used for all subsequent analysis. The top band in the DLHE was £50,000 and above (which I treat as £60,000). However, after the sample selection rules in Table 2.7 have been applied, only 0.1% of observations were in this category, with this choice having negligible impact on results.

[39] For the sample selection criteria described above, 87,327 individuals are in full-time work and have responded to the survey. Of these, 45,906 (53%) report a salary.

that are unemployed, are also unobserved. To illustrate the potential difficulty this may cause, after the sample selection procedures have been applied, 145,517 observations remain. Of these only 45,906 (32%) individuals are in the labour market and report their salary. Missing data is either due to self-selection out of the labour market, not responding to the survey or missing salary information. As a result, there may be a selectivity problem when comparing the two surveys. Differences recorded could be a result of who is responding to each of the surveys, rather than actual differences between students' expectations and realisations. A further issue maybe that certain groups have higher drop out rates than others, leading to different proportions recorded in each survey. For instance students from lower socioeconomic backgrounds may be more likely to drop out of university. Therefore one would observe a higher proportion of this group in the SIES, with data recorded during university, than the DLHE, with data recorded for graduates.

Perhaps the greatest worry is that the missing realised salary data in the DLHE comes disproportionately from groups who are earning a particularly high or low wage, creating bias in my estimate of the average graduate starting salary. To investigate this further, a probit regression for the item non-response to the salary variable has been conducted[40]. Results can be found in Appendix 2.4, while Table 2.8 uses the model results to illustrate the predicted probability of certain hypothetical groups responding to the question regarding salary.

[40] Here only item non-response is considered and not unit non-response to the entire questionnaire, though it should be remembered the overall response rate to the questionnaire is around 75%.

Table 2.8 Predicted probability of responding to DLHE salary question for a set of hypothetical individuals based on a probit model (see Appendix 2.4)

	Person A "good prospects"	**Person B**
Gender	Male	Male
Degree class	1st	1st
University type	Russell Group	Russell Group
Work status	Full-time	Full-time
Subject	Engineering	Engineering
Ethnicity	White	White
Disability	None	None
Graduate job	Graduate level	*Non-graduate job*
University location	England	England
Term time accommodation	Away from parents' home	Away from parents' home
Tariff (mean=300)	430 (1 S.D above mean)	430 (1 S.D above mean)
Home location	London	London
Degree a job requirement?	Formal requirement	*Not required*
Type of job	Managerial	Managerial
Probability of responding	76%	62%

	Person C	**Person D "poor prospects"**
Gender	Male	*Female*
Degree class	*2.2*	*3rd*
University type	Russell Group	*Post-1992*
Work status	Full-time	*Full-time*
Subject	*Art*	*Art*
Ethnicity	White	*Black*
Disability	None	*Yes*
Graduate job	*Non-graduate job*	*Non-graduate job*
University location	England	*Wales*
Term time accommodation	Away from parents' home	Away from parents' home
Tariff (mean=300)	430 (1 S.D above mean)	*170 (1 S.D below mean)*
Home location	London	*North East*
Degree a job requirement?	*Not required*	*Not required*
Type of job	*Administrative job*	*Administrative job*
Probability of responding	45%	31%

Notes:
1 Calculated using the probit model in Appendix 2.4.
2 Words in italics illustrate how the hypothetical individual is different to "person A" (who has "good" labour market prospects).
3 Source: Authors calculations using the DLHE dataset. Total sample size in the regression was 87,327. Dependent variable in the regression model these figures were based upon was a binary indicator of whether the graduate reported their salary (coded 0 if they did not and 1 if they did).

Person A has the characteristics of someone with excellent labour market prospects and thus expected to be a high wage earner. They have a good degree from a top university and are now in a graduate job. This individual has a 76% chance of responding to the salary question in the DLHE. Individuals B and C illustrate the impact of certain characteristics on salary response rates. Person B only differs from A in that he is in a non-graduate job, yet there is only a 62% chance of him responding. When I add further characteristics that are likely to mean lower wages in the labour market, such as gaining a 2.2 in Art and now being in an administrative job, the probability of responding drops to 45%. Person D is someone who is likely to have a particularly difficult time in the graduate labour market. Her probability of responding is just 31%, compared to 76% for someone who is likely to be a particularly high earner. This clearly illustrates that, if anything, the DLHE is likely to provide an *upwardly* bias estimate of the true graduate wage. To investigate this further, response weights were created by calculating the inverse of the predicted probability of students responding to the salary question. When these were applied to the individuals data, the estimated average graduate wage, for full-time students, fell from £16,455, to £15,996 (a drop of 3%)[41,42].

The selected SIES and DLHE samples, for those who reported wages, were then compared in terms of characteristics that could be observed within both populations. The results are shown in Table 2.9, both with and without the response weights for the two surveys applied[43].

[41] Response weights have also been calculated for part-time students.
[42] It should be noted this is significantly lower than the official figure HESA publishes. Appendix 2.5 fully documents why this is the case, and how the figures presented here are calculated.
[43] The SIES response weight refers to those described earlier in section 2.3 (contained in the dataset provided). The DLHE weight refers to those I have created via the probit model just described.

Table 2.9a Comparison of full-time students in the SIES and DLHE samples (weighted proportions in brackets)

	SIES %	DLHE %
Gender *		
Male	33.2 (46.5)	38.9 (41.4)
Female	66.8 (53.5)	61.1 (58.6)
Ethnicity *		
White	86.0 (86.7)	88.3 (87.8)
Asian	5.7 (5.6)	8.0 (8.4)
Black	2.9 (2.4)	1.6 (1.8)
Mixed/Other	5.5 (5.4)	2.1 (2.0)
University group		
Russell Group	22.9 (28.4)	23.4 (22.7)
Other Pre-1992	20.6 (18.9)	21.4 (20.3)
Post-1992	56.5 (52.8)	55.2 (57.0)
Social class (parents occupation) *		
Managerial/Professional	58.9 (60.0)	56.7 (56.3)
Intermediate	20.9 (20.1)	28.6 (28.4)
Routine/Manual	20.3 (19.1)	14.7 (15.3)
Living arrangement		
Parental home	21.9 (19.4)	21.2 (22.4)
Living away from home	78.1 (80.6)	78.8 (77.6)
Subject*		
Allied To Medicine	5.5 (4.5)	7.0 (6.3)
Sciences	9.5 (10.4)	8.7 (8.1)
Maths, Computer Science	5.8 (6.6)	9.3 (8.9)
Engineering, Technology	4.0 (5.0)	5.1 (4.7)
Architecture, Building	1.7 (2.0)	1.6 (1.7)
Social Sciences	13.7 (13.9)	10.1 (9.7)
Business	8.5 (8.8)	14.6 (14.1)
Languages	6.1 (5.8)	8.7 (8.5)
History	4.7 (5.3)	4.8 (5.0)
Art	13.9 (13.3)	8.8 (11.6)
Education	8.2 (6.9)	5.8 (5.6)
Combined	4.2 (3.9)	0.6 (0.5)
Psychology	1.9 (1.4)	4.1 (4.0)
Sports Science	1.7 (1.7)	3.2 (3.1)
Law	5.0 (5.2)	2.5 (2.6)
Mass Communication	3.0 (2.9)	3.3 (3.7)
Other	2.6 (2.4)	1.8 (1.9)
Observations	1,666	44,436

Notes:
1 Proportions when weights are used are reported in brackets. The weights used in the SIES correct for unit non-response and ensures the age-sex profile of the sample matches that of the student population.
2 These are described in section 2.3. The DLHE weights refer to non-response weights I have calculated to take into account missing data for graduates' salary. These are based on the probit model in Appendix 2.4.
3 Social class given for the 20,156 (46%) observations with data available in the DLHE
4 DLHE and SIES figures are for those who reported wages and after sample selection procedures have been applied
5 Medics have been excluded after initial analysis showed a greater number in the SIES than DLHE
6 * Indicates whether chi-squared test (of differences in proportions between surveys) is statistically significant at 5% level.
7 Source: Authors calculations using the DLHE and SIES dataset. Sample size in the former was 44,436 and 1,666 in the latter

Table 2.9b Comparison of part-time students in the SIES and DLHE samples (weighted proportions in brackets)

	SIES %	DLHE %
Gender*		
Male	33.1 (46.3)	42 (43.4)
Female	66.8 (53.6)	58 (56.6)
Ethnicity*		
White	87.7 (89.4)	90.9 (88.9)
Asian	4.1 (2.0)	4.2 (4.9)
Black	3.6 (4.1)	3.5 (4.3)
Mixed/Other	1.4 (4.1)	1.4 (1.7)
University group*		
Russell Group	12.1 (5.6)	3.4 (4.5)
Other Pre-1992	15.5 (11.1)	27.6 (19.0)
Post-1992	72.3 (83.3)	69 (76.5)
Subject*		
Medicine	3.3 (2.9)	0.1 (0)
Allied To Medicine	11.2 (8.4)	16.6 (18.4)
Sciences	3.1 (4.0)	3.3 (3.0)
Maths, Computer Science	5.6 (6.4)	6.9 (6.9)
Engineering, Technology	8.6 (16.9)	8.9 (8.2)
Architecture, Building	5 (6.7)	5.7 (6.1)
Social Sciences	9.6 (6.8)	11.1 (8.9)
Business	8.1 (9.0)	10.9 (11.0)
English, Languages	2.8 (2.4)	1.7 (1.8)
History	4.1 (2.8)	3.0 (3.2)
Art	4.0 (4.6)	1.2 (2.0)
Education	21.8 (18.0)	14.4 (16.3)
Combined	3.9 (3.3)	9.2 (7.0)
Psychology	0.4 (0.3)	2.5 (2.1)
Law	3.7 (3.3)	2.6 (2.5)
Mass Communication	0.9 (0.6)	0.6 (0.1)
Other	3.9 (3.8)	1.6 (1.6)
Observations	784	9,842

Notes: 1 See notes to the table above

There are some differences in the social class and gender composition of the two surveys, though the use of the SIES weights (see section 2.3) alters the proportions of the latter quite significantly. For social class, there is a problem with missing data in the DLHE, which is likely to be causing some mismatch between the two surveys. However, even with this difficulty, the differences in observable characteristics are not of great magnitude (although statistically significant due to the large sample size). For example, 46% of the SIES is male, compared to 41% in the DLHE. Alternatively, 53% of the SIES sample attended a Post-1992 university, compared to 57% of the DLHE. To check robustness, estimates are presented both with and without the sampling weights, to analyse the sensitivity of results.

One further issue is that although the populations appear broadly similar in terms of characteristics that are observable in *both* surveys, there may still be differences in characteristics that are unobservable. These characteristics could go unmeasured in either, or both, of the datasets. As a further check for robustness, the UK Labour Force Survey will be used as an alternative source of information on graduate wages.

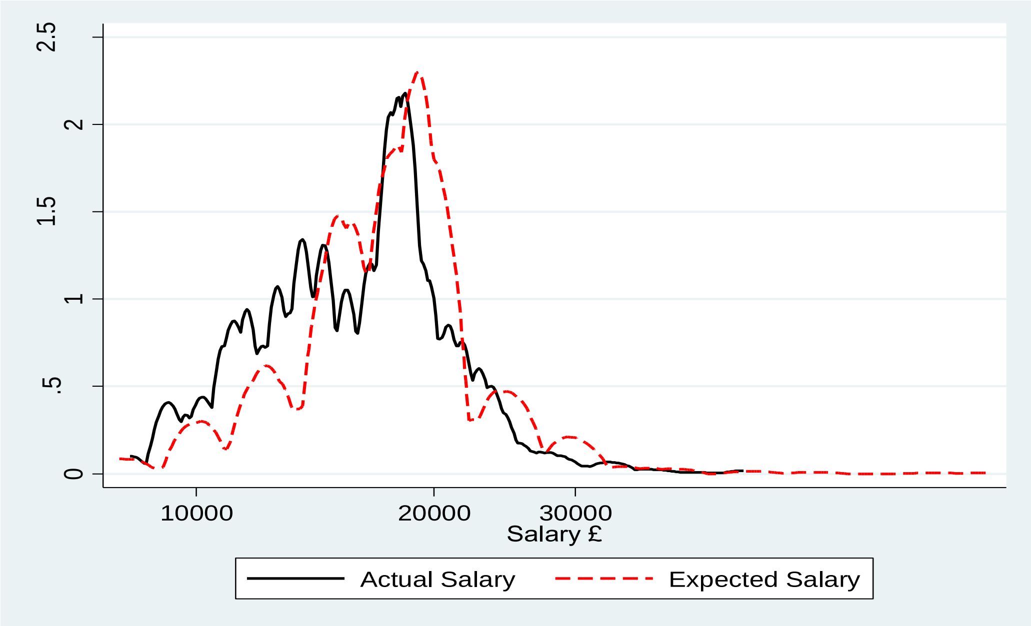
2.7 Comparison of expected and actual wages

Two methods are used to compare students' average wage expectations with the average actual wage. Firstly I calculate the ratio of the mean (median) expected to actual wage. The second method is to graphically represent the distributions, via kernel density estimates, of the actual and expected wage to identify differences.

Do students have realistic expectations?

I initially investigate whether students overestimate wages in the graduate labour market at the population level. The SIES and DLHE samples were divided into full and part time students, with summary statistics and kernel density estimates reported in Figure 2.2.

Figure 2.2 Kernel density estimates of actual versus expected log wages
Full-time students



Part-time students[44]

Notes:
1 Kernels and ratios are estimated with unweighted data
2 Figure in brackets denotes when both SIES and DLHE weights are applied to data
3 The kernel density estimates have used the default smoothing applied by Stata.
4 Authors diagrams produced using the DLHE and SIES dataset. Sample size in the DLHE is 44,436 for the full-time diagram and 9,842 in the part-time diagram. Sample size in the SIES is 1,666 in full-time diagram and 784 in the part-time diagram

[44] For part-time students, the sample selection rules differ slightly, to ensure adequate sample sizes. In particular, students of all ages and those doing “other” types of university courses, though not higher degrees, are included. Table 2.7b gives further details.

The ratio of the mean (median) expected salary to actual salary for part-time students is 1 (0.95), suggesting that, on average, part-time students have realistic expectations. Furthermore, the kernel density estimate illustrates how closely the distributions of part-time students' expectations and actual wages match. However a different conclusion is reached when looking at the results for full-time students. The kernel density estimate of students' expected wage appears to be to the right of the estimate for actual wages, especially near the lower tail, suggesting overestimation by students. This overestimation is also illustrated by the ratio of average expected to average actual salary in Table 2.10. Expectations are around 10% (£1,600) higher than average wages in the graduate labour market, using the means, and 12% (£1,900) higher using the medians. However when using the sampling weights to correct for non-response, the overestimation appears even higher at around 14% (£2,200). This means that students (on average) overestimate starting wages by over half the yearly fee now charged in the UK for university tuition[45]. Indeed, I find that this result holds under several robustness tests, including different sample selection rules to those presented. Moreover, in all estimates, expectations are statistically significant to actual wages at the 5% level. Thus there seems sufficient evidence to support the hypothesis that, on average, full-time students have unrealistic expectations of future wages.

As a check for robustness, the Labour Force Survey (LFS) has been used as an alternative data source for wages of recently qualified graduates. I use data from 10 quarterly surveys, running from September 2005 to March 2008. These dates were chosen as they relate to when the students covered in the SIES would have graduated and entered the job market. Moreover, from the September 2005 survey onwards, respondents were asked the question:

"Which, if any of these qualifications did you gain in the last 12 months?"

With one of the response options being:

"Degree level qualification including foundation degrees, graduate membership of a professional institute, PGCE, or higher"

[45] These particular students would have actually paid an upfront tuition fee of around £1,200 per year. Tuition fees changed for students starting after 2005 to a maximum of £3,000 per year, payable after graduation.

Respondents were also asked what their highest qualification is. I begin by restricting the LFS sample to respondents who hold a bachelors degree, and who were obtained this qualification in the past twelve months. Moreover, as per my previous selection rules for the SIES and DLHE, I restrict the LFS sample to those individuals who are under 25, working full-time and earning over £8,000 a year. Gross weekly wages are used from respondents in their fifth, and final, wave of the survey, and scaled up to the annual equivalent[46]. For those surveys conducted between 2006 and 2008, wages have been deflated to 2005 prices using the Retail Price Index, under the previously stated assumption that students report expectations in today's (2005) prices[47]. The final sample size is 194 observations[48]. Although this is a reasonably small number of observations, the LFS remains a useful resource to check whether results are robust. My findings are presented in Table 2.10.

Table 2.10 Comparison of average salary in the SIES, DLHE and LFS for full-time students (weighted estimates in parentheses)

	SIES expected wage £000	DLHE actual wage £000	LFS actual wage £000	Ratio SIES:DLHE	Ratio SIES:LFS
Mean wage	18.1 (18.3)	16.4 (16.0)	16.1	1.10 (1.14)*	1.13*
Median wage	18.0	16.0	15.1	1.12*	1.19*
Observations	1,666	44,436	194		

Notes:
1 SIES weighted by those provided in the dataset to correct for unit non-response. DLHE weights are those calculated in section 2.7 that correct for item non-response to the salary question.
2 * Indicates that the ratio is statistically different from 1 (where average expectations equals average realisations) at the 5% level (using a two sample t-test, assuming independent samples). Equality of medians tested using the Wilcoxon rank sum-test. However, this latter test does not take into consideration the complex survey design
3 Authors calculations using the DLHE, SIES and LFS dataset. Sample size in DLHE was 44,436, SIES 1,666 and LFS 194.

[46] Respondents to the LFS are asked for their wages in the first and fifth wave. However since I am using 10 consecutive waves, there would be a problem of double counting people if wage data was taken from both wave one and five. For instance, someone who was in wave 1 during September-December 2005 would be in wave five during September-December 2006, and hence have their wages recorded twice. Hence only wages in wave five are used.
[47] Even without this assumption, the ratio of expected to actual median wage is 1.15 using the labour force survey.
[48] Two outlying observations have been dropped due to their large influence on the mean wage in this small dataset. These individuals had wages over two times bigger (over £80,000) than the next largest observation (£40,000), and were over six standard deviations higher than the mean. Robustness was checked by including these two observations, with all the substantial results remaining intact (see the following footnote)

The estimate for mean starting wages from the LFS is £16,073 (standard error £370), very close to that recorded in the DLHE (£16,455), particularly after the DLHE has been weighted for non-response (£15,996). Consequently, the ratio of the means from the SIES to LFS (1.13) is only slightly larger than when comparing the means from the SIES to the DLHE (1.14). The difference when considering the median is, however, slightly larger (1.12 for SIES:DLHE compared to 1.19 for SIES:LFS)[49]. Moreover the difference between the average SIES expected salary and the average LFS wage is statistically significant, even with the limited sample size. Hence it appears that the preceding results are indeed robust to the data source used.

Does the realism of students' wage expectations vary by background characteristics?

Previously, I illustrated how students' wage expectations vary with several background characteristics. An interesting question is whether students who expect higher wages actually secure this premium in the labour market, or are they, on average, more unrealistic? This proves challenging methodologically without longitudinal data. The samples selected in the SIES and DLHE can be restricted further, for example to look at men and women separately, though this can obviously only be done for characteristics observable in both surveys. Hence factors such as parental income cannot be explored. Moreover there are likely to be further compositional issues, similar to those discussed in Table 2.9, particularly with the reduced sample sizes. This, coupled with the non-response in the DLHE, limits my ability for a more in-depth analysis. Nevertheless, Table 2.11 provides results, though these should largely be treated as indicative rather than definitive.

[49] Note with the outliers the mean salary is £16,749, and the median £15,163. With the outliers trimmed to £50,000 the salary is £16,420 and mean ratio 1.10, while the median remains at 1.19. Even with including these trimmed outliers, there is still a statistically significant difference between the average expected and average LFS salary.

Table 2.11 Comparison between mean expected and mean actual wages for full-time students, based on background characteristics (weighted estimates in parentheses)

	Mean expected wage £000	Mean actual wage £000	Ratio
All full-time students			
Final year	17.4 (17.7)	16.5 (16.0)	1.06 (1.10)*
1 year	17.9 (18.2)	16.5 (16.0)	1.09 (1.14)*
2 or more years	18.8 (18.9)	16.5 (16.0)	1.14 (1.18)*
Ethnic group			
Black/Asian	19.2 (19.6)	17.3 (17.2)	1.11 (1.14)*
White	17.9 (18.1)	16.1 (15.8)	1.12 (1.13)*
Mixed	18.6 (19.0)	16.8 (16.9)	1.11 (1.13)*
University type			
Russell Group	19.2 (19.4)	17.3 (17.1)	1.11 (1.13)*
Other Pre-1992	17.6 (17.8)	16.6 (16.5)	1.06 (1.08)*
Post-1992	17.8 (17.9)	15.6 (15.4)	1.14 (1.16)*
Gender			
Male	18.7 (18.9)	16.9 (16.6)	1.11 (1.14)*
Female	17.8 (17.8)	15.8 (15.6)	1.12 (1.14)*

Notes:
1 See notes to Table 2.9 for more details on the weighted estimates.
2 Excludes medical and part-time students, due to different proportions found within the two sources.
3 For all variables the mean expected and actual salary are significantly different at the 5% level
Social class has not been investigated due to a large amount of missing data in the DLHE for this variable.
4 Indicates the ratio is significantly different from 1.0 at the 5% level (using a two sample t-test, assuming independent samples)
5 Authors calculations using the DLHE and SIES dataset. Sample size in the former was 44,436 and 1,666 in the latter.

Table 2.11 illustrates that junior students are less realistic than their peers who are approaching the labour force. Whereas final year students tend to overestimate their starting salary by only 6%, on average, those who have just entered university overestimate by around 14%. An important implication is that students who have just made the decision to invest in university education have especially inflated expectations. In the introduction to this thesis, I explained the importance of students understanding the graduate labour market when making their educational decisions. However this analysis indicates that students are not even particularly aware of starting salaries when making their human capital investments. In other results, there is little evidence that students at Russell Group universities are any less realistic than those at Post-1992institutions, who overestimate on average by 14% and 11% respectively. However students at other Pre-1992 universities appear to be much more accurate, overestimating by only 6%, on average. Similarly, there is no evidence to suggest differences based on gender or ethnicity, with overestimation by each group, on average, of around 11%.

Does the accuracy of students' wage expectations depend on the subject they study?

Though full-time students seem to be too optimistic, on average, in their wage expectations, those in certain subjects may have access to better information on wages, and thus make better predictions. It is hypothesised that students who are studying a subject leading to a particular career will be more realistic, as they are likely to research specific jobs and have better knowledge of the labour market they face. Alternatively students who take language and art based courses are likely to enter a far more general labour market, with less certainty about their future career prospects. Results are provided in Table 2.12, with kernel density estimates for various subjects shown in Figure 2.3[50].

[50] Again, there may be some compositional issues surrounding the two surveys that may hamper comparison. Caution is thus advised when interpreting the results, though I believe the general pattern to hold.

Table 2.12a Comparison between mean expected and mean actual wages for full-time students by subject groups (weighted estimates in parenthesis)

Subject	Sample size SIES (expectations)	Mean expected wage £000	Mean actual wage £000	Ratio
Medicine	162	26.5 (25.8)	28.8 (28.5)	0.92 (0.91)
Education	137	17.9 (17.9)	17.5 (17.4)	1.02 (1.03)
Engineering, Maths, Computer Science	164	19.5 (19.2)	18.4 (17.5)	1.06 (1.10)
Allied To Medicine	92	18.9 (19.0)	17.1 (16.7)	1.10 (1.14)
Social Sciences	230	18.5 (18.9)	16.6 (16.4)	1.11 (1.17)
Art	238	16.3 (16.4)	14.4 (14.2)	1.13 (1.15)
Business, Management	143	18.4 (18.9)	16.3 (16.1)	1.13 (1.17)
History, English & Languages	232	17.5 (17.7)	15.3 (15.1)	1.15 (1.17)
Psychology, Sports Studies, Combined, Other	204	17.7 (17.7)	15.3 (15.3)	1.15 (1.16)
Sciences	161	18.3 (18.2)	15.8 (15.7)	1.16 (1.16)
Law	83	20.7 (20.6)	14.8 (14.8)	1.35 (1.39)

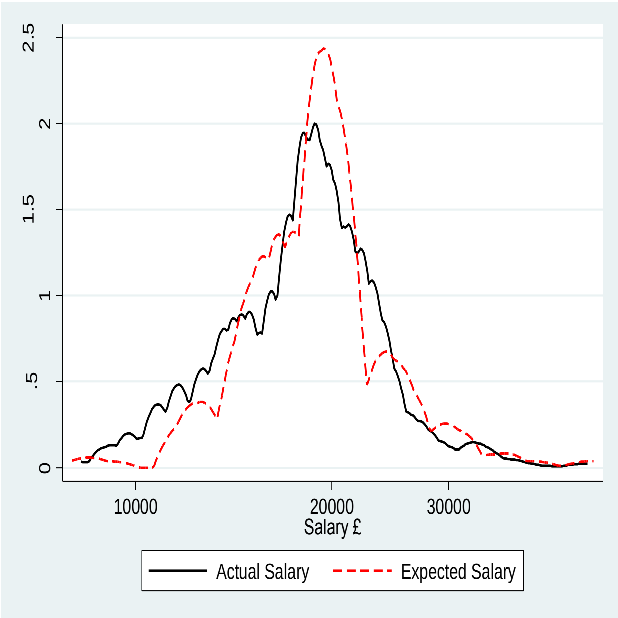
Table 2.12b Comparison between median expected and median actual wages for subject groups

Subject	Median expected £000	Median Actual £000	Ratio
Medicine	26.0	28.7	0.91
Education	18.0	18.0	1.00
Allied To Medicine	18.0	17.0	1.06
Art	15.0	14.1	1.06
Engineering, Maths, Computer Science	20.0	18.0	1.11
Social Sciences	18.0	16.0	1.12
Psychology, Sports Studies, Combined, Other	17.0	15.1	1.13
History, English, Languages	18.0	15.1	1.19
Sciences	18.0	15.1	1.19
Business, Management	18.0	15.1	1.19
Law	20.0	14.1	1.42

Notes:
1 See notes to Table 2.9 for more details on the weighted estimates
2 For all variables the mean expected salary is significantly different from mean actual salary at the 5% level, except for Education
3 “All full time students” excludes Medical students
4 Law is somewhat of an outlier, due to the large number of students continuing into postgraduate training (hence only those with quite low wages are observed in the DLHE). Excluding this group from the aggregate analysis has little influence on the overall result
5 Authors calculations using the DLHE and SIES datasets

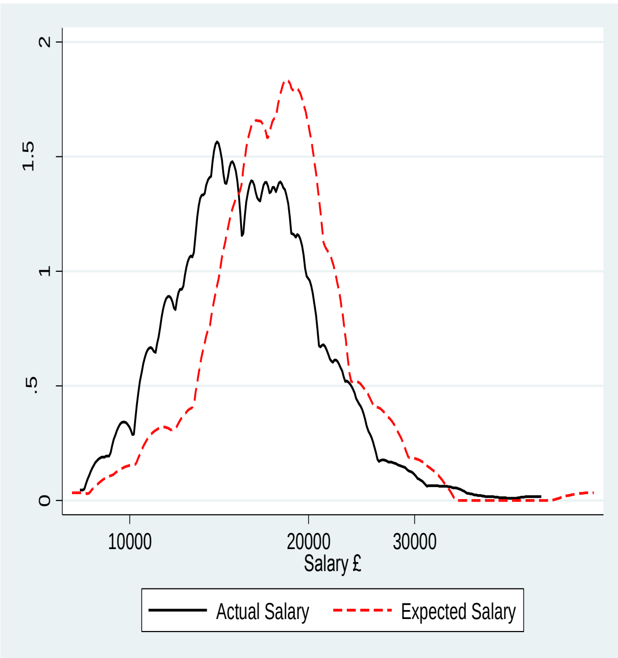
Figure 2.3 Kernel density estimates of actual versus expected log wages, by subject group

Engineering, Computer Science, Maths

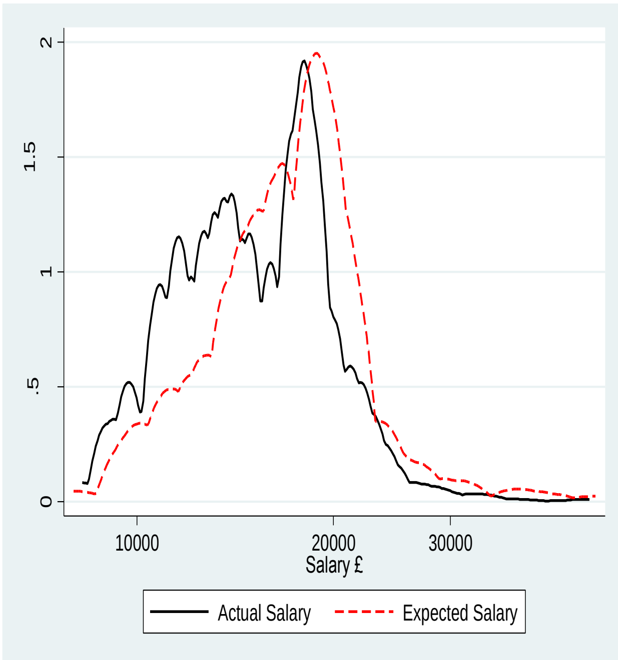


Education

Business and Management



History, English and Languages



Notes: 1 Diagrams produced by the author using the DLHE and SIES datasets

The results show that, on average, there is overestimation in all subjects, except Medicine, where there is some evidence of underestimation[51]. Law also appears as an outlier, with especially large overestimation. One possibility may be that students wishing to enter a legal profession have to complete additional study at a Law school after university. Therefore the vast majority will not be in the labour market yet and their wages unobserved; there is the problem of selectivity. Those that have entered work straight from their course are likely to be in a much less lucrative position than Law graduates who continue their legal training and later enter the profession[52].

As hypothesised, students who are studying a subject that directly leads to a career have more realistic wage expectations. Medicine and Education both appear at the top of the table, and suggest these groups are (on average) very close, or even underestimate, their future wage. In contrast to the hypothesis, students in subjects Allied to Medicine appear no better than the average. However during the period studied, the number of positions available for newly qualified physiotherapists and nurses was particularly low. Thus these students, may have struggled to find jobs in their desired fields and thus had to compete in the more general graduate labour market.

The kernel density estimates in the upper panel of Figure 2.3 illustrate how closely the distributions for actual and expected salaries match for Education and Engineering, Maths and Computer Science. A comparison to the distributions for Business and Management, and History, English and Languages, clearly illustrates the superior estimates made by those in the former groups. Interesting patterns also occur between the subjects that lead to the more general labour market. Those subjects where the errors seem to be largest generally involve language, rather than technical, skills. For example, the Social Sciences, Language and Business courses mainly involve writing essays, while subjects teaching more mathematical skills, such as Computer Science and Engineering, appear to contain more realistic students. Science can be identified as an exception to the discussion above; it is a course that develops technical skills, but whose students appear unrealistic. This could reflect students believing that they

[51] Expected salaries are significantly different from actual wages in all cases, except for education. This suggests sampling variability is not responsible for the observed differences.
[52]Excluding this group from the analysis for “All full-time students” does not substantially change the results.

will receive a high wage because they have chosen to take on a traditionally challenging subject. In reality, however, the technical skills they have built up may be required by relatively few employers, or they end up in a field unrelated to their degree. This may mean they suffer an unanticipated need to retrain in an unrelated discipline and have to accept a lower starting salary, or perhaps move into a non-graduate job.

One may be tempted to suggest that the very different gender composition of subjects is leading to these differences in overestimation. Table 2.13 provides the ratio of average expected to actual wages broken down by subject and gender.

Table 2.13a Comparison between mean expected and mean actual wages of *men* in each subject group

Subject	**Sample size SIES**	**Mean expected wage £000**	**Mean actual wage £000**	**Ratio**
Medicine	55	27.5	29.0	0.95
Education	21	18.3	17.2	1.06
Engineering, Maths, Computer Science	115	19.6	18.7	1.04
Allied To Medicine	16	19.9	16.7	1.19*
All full-time students (excluding Medicine)	559	18.7	16.9	1.11*
Social Sciences	74	19.7	18.1	1.09*
Art	71	16.7	14.8	1.13*
Business, Management	60	18.9	17.1	1.10*
History, English, Languages	60	18.5	15.5	1.19*
Psychology, Sports Studies, Combined , Other	63	17.3	16.4	1.05
Sciences	56	18.1	16.5	1.10*
Law	23	20.9	15.9	1.31*

Table 2.13b Comparison between mean expected and mean actual wages of *women* in each subject group

Subject	Sample size SIES	Mean expected wage £000	Mean actual wage £000	Ratio
Medicine	107	26.0	28.6	0.91*
Education	115	17.8	17.8	1.00
Engineering, Maths, Computer Science	49	19.1	18.1	1.06
Allied To Medicine	76	18.7	17.0	1.10*
All full-time students (excluding Medicine)	1,125	17.8	15.8	1.12*
Social Sciences	156	17.8	16.1	1.10*
Art	167	15.8	14.7	1.08*
Business, Management	83	18.0	15.9	1.13*
History, English, Languages	172	17.1	15.6	1.10*
Psychology, Sports Studies, Combined, Other	141	17.7	15.1	1.18*
Sciences	105	18.3	15.8	1.16*
Law	60	19.8	14.7	1.34*

Notes:
1 SIES sample size refers to unweighted data
2 Sample size for men and women in SIES differ substantially due to differential response rates. See Section 2.3 and Table 2.9 for further details
3 * Indicates ratio is statistically different from 1 at the 5% level (using a two sample t-test, assuming independent samples)
4 Authors calculations using the DLHE and SIES dataset. Total sample size in the former was 44,436 and 1,666 in the latter.

Again, due to the small sample sizes in particular subjects, these results are indicative rather than definitive. Nevertheless the results do seem to suggest similar patterns for men and women. Both, on average, underestimate in Medicine, and overestimate in the Sciences, Business and Language based subjects. In many instances, when the ratios do differ for men and women, this can probably be explained by the small sample size. This is the case for Education and subjects Allied to Medicine, particularly for men. On the other hand, there is a tentative suggestion that men and women, on average, have less accurate expectations when they are studying a subject traditionally associated with the other gender. For instance, men appear to overestimate, on average, more than women in the Arts and Languages. Meanwhile women make worse estimates in the Sciences.

2.8 Discussion

This research has been limited to a single estimate of future wages made by students relating to one specific time in the future. How much I can say about the realism of students about their future earnings over the course of their whole lifetime is therefore quite restricted. Nor do I have students' views of the counterfactual; what they would expect to earn had they not gone to university. This study has also only considered the financial aspects of studying for a degree, and not the non-monetary benefits of university to both the individual and society. Nevertheless, the findings suggest that students have a tendency to, on average, overestimate future wages and that, consequently, one must at least begin to question whether they are making well informed human capital investment decisions. Indeed, the fact that certain groups can predict future wages (on average) better than others may suggest economists' assumption of there being no systematic differences between young adults' expectations and later realisations may be based on a rocky foundation.

Of course, it is still possible for university to be a good investment, even under such conditions. Many individuals will still find higher education both a financially and culturally profitable experience, even if they do not obtain the wage they once expected. Nevertheless it is equally plausible that by overestimating future wages, some students may mistakenly choose to go to university, who will not receive the benefit they expected on enrolment. The UK Class of 99' report by Purcell, Elias, Davies and Wilton (2005) illustrates such feelings in qualitative research, as shown below:

'I would have still ended up in the position I'm in now if I would have carried on working full-time.... I applied for over two hundred jobs, I felt this degree was a total waste of time; I was a self-funding student, which was a waste of money. I'm still paying for it now, I'm a single parent and to be honest it was the biggest waste of time and money that I've ever spent'.

……..everyone tells you if you do a degree the world will be your oyster, you'll earn loads of money. No'. Page 194

Other aspects of this research may also have importance for higher education policy. Students build up debt while at university, when income is low, and expect to pay this back when they have a job after graduation. Gustman and Stafford (1972) also show that students with higher wage expectations tend to consume more at university. From an economic point of view, students are using credit markets, in part, to smooth their consumption over time. However, if students overestimate their future wage, they may also be overestimating their ability to pay back the money they borrowed. This may lead to students taking on too higher levels of debt that they later struggle to repay, due to the fact they are not in as well a paying job as expected. It may also mean they are willing to take on debt to pay for high tuition fees when entering university, but regret this decision in hindsight when paying back the money is harder than they once expected.

Another important issue is how this relates to widening access schemes proposed by European governments, and in particular the UK target of getting half of all school leavers to experience higher education. The benefits of university are widely promoted by governments, and in particular career prospects, to encourage individuals to continue their education. However, this practise could enhance students' unrealistic expectations (which certainly seem to be the case in the quote above) if the fruits of the graduate labour market are over-empathised. For example, a recent UK government publication as part of the AIMHIGHER programme discusses various careers children could enter if they obtain a degree, including becoming a "Hotel Manager"[53]. It suggests a degree in Leisure and Tourism as a possible qualification, and that their "potential earnings" are around £80,000 (presumably meaning per year, at the peak of one's career). Although this is possible, this seems a very high figure to actively promote.

[53] See http://www.bis.gov.uk/assets/biscore/corporate/migratedD/publications/H/HigherEd-DontStop

A related point is whether students are being given accurate information about salaries and employment prospects from the various available sources. For instance, the UK media reported that the average graduate starting salary in 2005 was £22,000[54]. But this was based on the average salary offered by just 226 of the UK's highest paying graduate employers[55]. Moreover this research has illustrated the non-response bias in the salary figures within the DLHE; often quoted by the media and government ministers as the "true" graduate wage. This information may well have an impact on students' views of the graduate wage and inflate their expectations to unrealistic levels. More information for students on the distribution of starting salaries, and their likely place on the scale, is required to make sure that individuals are making well informed decisions when continuing their university education.

2.9. Conclusion

This chapter set out to explore heterogeneity in UK students' wage expectations and to identify whether they held "realistic" views of the graduate labour market. In doing so, this provides the first study in Europe to explore wage expectations using data that is designed to be nationally representative of the UK student population. The results highlight:

- Wage expectations vary significantly based on how far the student has progressed through university, though this is largely explained by their differing views on employment opportunities.

- Students' idiosyncratic characteristics play an important role in determining their wage expectations.

- Quality of the institution is associated with students' wage expectations, although this could be reflecting the fact better institutions admit students of higher ability

[54] See http://www.guardian.co.uk/education/2005/jan/06/highereducation.uk1, who lead with the headline "value of a degree now £22,000 per year"
[55] See http://www.agr.org.uk/Content/AGR-Graduate-Recruitment-Survey-2005-Winter-Review for further details

- While part-time students seem realistic in their wage expectations, those studying full-time tend, on average, to overestimate their starting salary by around 10 - 15%.

Specifically, this chapter shows how, as students' progress through university, their views on life as a graduate change. Final year students are less optimistic about their ability to land a "career job" and their starting salary. This indicates that students probably learn about their own ability and the labour market through their time in higher education, and that prospects may not be as bright as they once expected. There also appears to be a significant difference based on family income, with students from high income backgrounds expecting a greater salary than their low income peers. However the initial hypothesis that students from ethnic minority backgrounds may expect some form of discrimination in the labour market, and therefore predict a lower wage, is rejected. Ethnic minorities actually expect a higher salary than their white peers (even after conditioning on the student they study, views on the labour market and quality of institution they attend).

The second half of the chapter turned to whether such wage expectations are realistic. The only other study to use a balanced sample (collecting data from students in a range of subjects and institutions), Webbink and Hartog (2004), concludes that students can accurately predict their starting salary. My results suggest the exact opposite; on average full-time students overestimate wages in the graduate labour market. In some cases, I find this to be quite severe. Students who have just entered university overestimate their starting salary by almost 20%. On the other hand, part-time students and those studying a subject leading to a particular career expect salaries reasonably close to the observed wages. There are three explanations for why I find this result. Firstly, I can not completely rule out the possibility that selectivity of response is generating part of the observed gap between expectations and realisations. However, I have very carefully checked the quality of my data, and performed several robustness tests, to minimize the impact of this on my results. A second possibility is that students are not "rational". They may be quite knowledgeable about the graduate labour market, but are unable to use this data to accurately form predictions of their future wage. Alternatively, my final explanation is that students simply do not hold enough information to predict their future earnings. This is perhaps the most compelling argument, given the other results in this chapter. For instance, the superior predictions of university seniors could be driven by a greater incentive to collect information as the labour market approaches. This group may also have lower costs of acquiring such information, due to their close interaction with recent graduates. Similar arguments can be made for other groups who make relatively good predictions. For instance education students may be realistic about future wages because national pay scales provide them with relevant and accurate labour market information at a very low cost.

If this final explanation is correct, then why do students not collect such information in the first place? It may be that for the majority of students, the costs of acquiring such information are simply too high. To ensure young adults are making well informed choices regarding higher education, I advise policymakers to lower this cost by increasing the availability and detail of data on graduate wages.

Chapter 3

Who has realistic income expectations: Students or workers?

"Many have argued that attitudes of investors in human capital are very different from those of investors of physical capital because the former tend to be younger, and young persons are especially prone to overestimate their ability and chance of good fortune"

Becker (1993) Human Capital: A Theoretical and Empirical Analysis with Special Reference to Education (Third Addition), page 93

Economic models of schooling choice are often based on the assumption that young adults have a realistic idea of what their future income will be. This assumption can be decomposed into two parts. The first is that young adults hold "full information" – that they understand the economic benefits of the different educational options available. The second is that they hold "rational expectations"; that they are able to use this information efficiently to produce realistic assessments of the future. If this is true for all members of the population, then there should be no systematic differences (on average) between individuals' apriori expectations and later realisations.

It is under these assumptions that economists often use realised ("ex-post") income data to estimate how young adults' expected financial returns to education influence their schooling choices. Notable examples include Berger (1988), Willis and Rosen (1979) and Boudarbat (2004). They all find that expected returns have a large and statistically significant effect on young adults' decisions, whether this is to enter higher education at all or the specific subject they take. However, these results are heavily reliant on the strong assumptions described above. If these are violated, then one may question the robustness of such results.

In chapter 2 I illustrated that UK students, on average, overestimate their starting salary. Yet I also found that the accuracy of labour market predictions varies substantially between different groups. In particular, those who probably held the least labour market information were, on average, the least realistic. This has raised some question marks over whether the assumptions described above hold. However, this analysis was limited by the fact I only considered the accuracy of students' expectations in a single setting over a short time horizon. From an economic perspective, it is lifetime income, rather than starting salaries, that influences people's decisions. An interesting extension of the work completed in chapter 2 is therefore whether students in other settings, and over longer time horizons, are just as unrealistic about their future income.

Moreover in chapter 2 I only considered the expectations of young adults in higher education. Indeed, all other studies, that I am aware of, have done the same; those who have chosen to enter the labour market straight from school have typically been ignored. Yet this group is both relatively large in size and of substantial interest. In particular, workers may have the opportunity to collect valuable information about the labour market from their employers, colleagues and the job search process itself. A student trying to access the same information may face much higher costs. If both groups are "rational", processing all the information that they hold efficiently, one would expect workers' additional information to translate into more accurate expectations. On the contrary, one may argue that students are less myopic than their peers who enter the labour force, and thus less readily discount their future income. Consequently, they may have more incentive to collect information about long-term labour market outcomes and hence hold more realistic expectations. Likewise, it maybe that children who invest in higher education are focused on one particular career, while 20 year olds in the labour force perhaps move somewhat haphazardly between different types of job. Thus it may be that students seek out more specific and relevant information than their working peers, leading to a more realistic assessment of future labour market opportunities.

Indeed this argument may hold true for some groups of students, but not for others. Those studying Education, Nursing or Engineering are being trained for specific jobs. With career councillors widely available in almost all Higher Education institutions, these students are probably well informed about the graduate labour market. On the other hand, someone in a course not leading to one particular career, such as Arts, Humanities, Languages or Social Sciences, may only have a vague idea about the type of job they will pursue. Thus these students receive only quite broad, low quality labour market information and will thus be prone to either under or overestimation of their future income. Indeed, I have already found some evidence of this in the previous chapter. Nevertheless, the discussion above illustrates the interesting insights that a comparison between different groups of students and workers might bring.

At present, such topics have received very little attention in either America or Europe. There are some small scale US studies that compare *students*' expectations to actual labour market outcomes. However, these suffer methodological difficulties, and results can not be generalised to the wider student population. Furthermore students are generally treated as a homogeneous group. There is little discussion of the association between dropping out of college, idiosyncratic ability, subject being studied and the accuracy of future expectations. Of particular note, no comparison is made to the expectations of their peers in the labour force. I make a significant contribution to the small US literature by using a detailed, nationally representative sample of both students and young workers to consider how the factors listed above influence the accuracy of 20 year old men's income expectations ten years into the future.

My results suggest that the US student population over-estimate their future income. However, unlike existing studies, I show that this result only holds for certain groups; I find that some students actually make quite good long range predictions, overestimating their income ten years into the future by (on average) less than 10%. Moreover, I find that differences between students and workers are not as pronounced as one may expect; under certain conditions, students actually hold more realistic expectations than their peers in the labour force.

I begin in section 3.2 by reviewing the current literature on income expectations and motivating the need for this research. In sections 3.3 and 3.4 I describe the National Education Longitudinal Survey (NELS) data. This is followed in section 3.5 by my analysis of young adults' labour market expectations. I conclude with a discussion of my key findings, and argue that either young adults do not hold enough labour market information to predict their income at age 30, or simply choose not to incorporate it into their expectations.

3.2 Literature and research questions

To my knowledge, there are six published US studies that investigate *students'* income expectations. The first to consider this topic was Smith and Powell (1990). They asked 400 students at two mid-western universities how much they expect to earn when they graduate and after 10 years in the labour market. Respondents were quite realistic about pay in their first job, but overestimate wages in 10 years time. Betts (1996) asked 1,000 students at the University of California to predict wages for a hypothetical individual under several different scenarios. He finds that students quite accurately predict the wages of young workers, but overestimate the pay of those with ten or more year's labour market experience. Blau and Ferber (1991) collected data from 351 students studying in the Business faculty at the University of Illinois. Again, students seem quite realistic about starting wages, but become progressively unrealistic over long time horizons. Carvajal et al (2000) analyse the expected starting salary of 219 Business students at Florida International University. They find over-estimation of around 10%. Rouse (2004) investigates the wage expectations of 69 high school seniors from the Baltimore City Public School District[56]. She finds these high school seniors to be quite unrealistic about their future income at age 30. Dominitz and Manski (1996) take a different approach. They asked 110 Madison students several questions to try and not only capture individuals' expectations, but also their uncertainty about future outcomes. They find that male students are reasonably realistic, but girls less so.

[56] Rouse (2004) also uses the NELS data analysed in this paper. In particular, she compares the NELS expected income data to external estimates on actual wages drawn from the 1990 Census. I analyse the NELS data in greater depth than Rouse, and focus on a set of quite different hypotheses. In particular, she is concerned with differences between ethnic groups, where my concern is the accuracy of students expectations compared to workers.

These studies suggest that students have a reasonably good understanding of starting salaries, but are less realistic about their future income over longer time horizons. However, this relatively small literature is somewhat limited by the scope and design of the aforementioned studies. Data are typically:

(a) Collected from students at just one (or at most two) universities
(b) Drawn via convenience sampling, rather than a probabilistic method
(c) Over-represent students from mathematical subjects (Economics, Finance, Engineering) and under represent those studying Art
(d) Of limited sample size

This causes several methodological problems. Firstly, as samples are often drawn from one university and a handful of subjects, it is difficult to generalise results to the wider US student population. This leads to problems when the authors try to assess whether students hold "realistic" expectations. Wage expectations, drawn from a highly selective survey, are compared to data on national graduate wages from an external data source, such as the Current Population Survey (CPS) or Census. The wage expectation and realisation data often represent two populations that could differ in all manner of characteristics. This may clearly bias any assessment of whether expectations are realistic. Secondly, small sample sizes mean that the wage expectation data suffers from quite large sampling error. But, as data are usually drawn via convenience sampling, reliable standard errors can not be calculated. Hence the true extent of sampling error on results actually often goes unknown.

In addressing my first research question, I attempt to overcome these limitations by analysing the income expectations of a large, nationally representative sample of American adults using NELS 1988 data. As this survey was designed to be nationally representative, my results should generalise to the wider US student population. Furthermore, as expectations and realisations are collected from the same individual over time (i.e. this is a panel dataset), my results should also be driven less by the composition and selectivity of the sample than the small scale studies cited above. Moreover, as the NELS data was collected using a large, probability based sample design I can adequately demonstrate the influence that sampling variation has on my estimates. In summary, my first research question is:

Q1. Do 20 year old male students in the US, on average, have realistic expectations of their income at age 30?

However students in countries like the USA, with its large and diversified higher education system, are not a homogeneous group. The accuracy of their labour market expectations is probably related to the subject they study, whether they actually graduate from university and their underlying cognitive ability. For instance, young adults who begin, but do not complete, university probably form their expectations based on the belief that they will obtain a degree. They may not adequately account for the possibility of dropping out, and hence (ex-post) their expectations will appear overly ambitious[57]. Similarly, given the results for the UK in chapter 2, one may suspect the accuracy of students' expectations to vary substantially with the subject they study. Indeed, as noted in the introduction, students studying certain subjects may hold more information about the labour market than others. This may be because they are already being trained to enter a specific job (teacher, nurse, engineer) and hence collect specific and detailed information compared to their peers entering the more general graduate labour market. Alternatively, it may be that wages within these jobs, or the wages of previous graduates from similar disciplines, have quite low variability. Hence students within these subjects face less uncertainty than some of their peers. In either case, one might expect to find similar patterns to those observed in the UK, where students studying mathematical and vocational subjects make better predictions than their peers in more creative disciplines. Such details have rarely been discussed in the US literature, hence my second research question:

Q2. Do students in the US who drop out of university hold particularly unrealistic labour market expectations? Are maths and vocational students more realistic about their future income than those studying more creative subjects?

[57] This is something that I shall go on to explore in section 3.5.

As the reader may have noted, all the studies reviewed focus on students at university. I do not know of any work that investigates the expectations of young adults who have chosen to enter the labour market rather than continue their education[58]. Apriori, one may expect young workers to hold more realistic expectations than students, as they probably have greater access to relevant labour market information. For instance, they will have contact with older workers who, either formally or informally, pass on sector-specific details of future pay and progression. Alternatively, organisations themselves could make information on career progression and pay freely available to their staff. Another factor is that workers have been through the job search process at least once. They should have found out about wages and career opportunities during this time. Indeed, these individuals may have held unrealistic expectations before this experience, but actually going out and trying to find a job may have taught them the realities of their employment opportunities[59]. Many university students would not have had a similar experience of searching for a full-time job, and may still be holding onto their unrealistic expectations. Self-selection into the labour market or higher education may also play a role. Educational attainment is linked to migration (see Borjas 1999). The less educated (who have self-selected into work) are more likely than students to stay in their age 20 location. Workers therefore gather information about wages in the local labour market that they incorporate into their expectations. Students, on the other hand, may well expect to work in other areas of the country. The local labour market will be less informative for many of them. Hence one would expect the labour market to have more salience to those who are already actively employed[60].

[58] Dominitz (1998) assesses the accuracy of American workers' wage expectations. However he does not specifically investigate the expectations of young workers, or how realistic they are compared to students. He also focuses on wage expectations for the year ahead, whereas this paper looks over a longer time horizon.

[59] Recent work in the sociological literature by Morgan (2005) depicts young adults as "Bayesian learners". In particular, he illustrates the accuracy of a "fast" and "slow" learner's expectations over time. Morgan suggests the difference between fast and slow learners could be to do with the different timing of key life events. This could include entry into the labour market, a period when young adults should receive a lot of information that will lead them to quickly (and more accurately) updating their expectations.

[60] Counter-arguments to this hypothesis have been presented in the introduction to this chapter.

Based on these discussions, my third and fourth (inter-related) research questions are:

Q3. Can 20 year old US workers, on average, make realistic predictions of their income at age 30?

Q4. Do 20 year old US workers, on average, make better predictions of their income at age 30 than their peers in higher education?

Although these hypotheses are similar, they do pertain to slightly different things. In particular, it is possible that both workers and students overestimate their future income (reject the null hypothesis of no difference between expectations and realisations in Q1 and Q3) but for workers to still make better predictions than students (reject the null hypothesis that students and workers are equally unrealistic in their labour market expectations in Q4).

This work adds significantly to the literature reviewed at the start of this chapter. I know of no other study that analyses nationally representative data on young adults' income expectations. Moreover, to my knowledge, I am also the first author to use panel data to compare income expectations and realisations over a 10 year time horizon. Thus I can more accurately compare students' expectations to later realisations, with my results being more likely to generalise to the wider US population. I am also able to tackle several new and interesting hypotheses that put the existing work on students' expectations into a wider context.

3.3 Data

To address these research questions, a nationally representative data source is required that follows young adults from their initial predictions of future income to their later success in the labour market. This needs to follow an entire cohort of young adults and not just those who continue on to university. One source is the National Educational Longitudinal Survey (NELS) from the US. This study's aim was to provide data about adolescents at critical points in their development and later into their careers using a nationally representative sample of adolescents. Children were initially interviewed in 1988 when the majority were 14 years of age. They were then followed up four further times, at ages 16, 18, 20 and 26. Parents and teachers of the pupils also completed the first three rounds of the survey.

In the first wave (age 14), a two-stage stratified sampling design was employed, with schools as the primary sampling unit, and probability of selection (of schools) proportional to size. 1,052 schools participated in the survey, including some oversampling of private institutions. A random sample of 26 students was then selected from each school. 26,432 students were eventually selected with 24,599 taking part (93%). In each of the next two waves (age 16 and 18) students who participated in the initial survey were followed up. The sampling process added some newly selected students (1,043 at age 16 and 244 at age 18) [61]. This was done to create a valid probability sample (a nationally representative cross section) of students in each of the respective years. In total, 20,923 18 years olds took part in the third wave. The fourth wave took place when students were 20 years old. To reduce costs, a sub-sample was selected based on demographic characteristics and response history. It is important to note that this reduction is not the result of sample attrition, but from a conscious effort of the survey design to limit burden and cost[62]. This led to the age 20 sampling frame being reduced to 15,964 individuals. In total, completed responses were available from 14,915 (93%) 20 year olds. Further details are given in Appendix

[61] These students were not randomly selected, but drawn from schools where there were other second and third wave respondents. More details can be found in Appendix 3.1 and page 56 of Curtin et al (2002).

[62] Around 5,000 individuals were dropped from the study. 2,000 of these were classed as "poor responders", who were basically excluded because of the low chance of future contact. Hence it may be more appropriate to consider these 2,000 observations as non-respondents. The other 3,000 individuals dropped were not classified as poor responders, but excluded purely to lower costs. Further details are given in Appendix 3.1.

3.1. The final survey took place when most sample members were 26. Data are available for 12,144 individuals (76% of the age 20 sub-sample).

There are obviously some issues of non-response due to sample attrition. One way to help correct for differential response rates in terms of observable characteristics is the use of survey weights. The NELS dataset contains a cross-sectional weight for those who took part in the final survey, and various panel weights. Unfortunately a panel weight is not provided for those who completed the final *two* surveys (ages 20 and 26). Instead a panel weight is available for those who had completed the final *three* surveys (ages 18, 20 and 26)[63]. This refers to the population of high school seniors in 1992. The National Centre for Education Statistics describes this panel weight:

This is the second, third, and fourth follow-up panel weight, which applies to the 12th grade cohort. It applies to fourth follow-up respondents (i.e. 2000) who were also respondents in the second and third follow-up rounds (i.e., 1992, 1994). It estimates longitudinal parameters that describe the population of spring 1992 12th graders.

This weight shall be used in all subsequent analyses to help adjust for unit non response and over sampling of certain minority groups. Therefore, the population I am describing in this analysis is those who were high school seniors in 1992[64].

A vital question is how respondents were asked to report their future income expectations. When respondents were 20 years old, they were asked:

"What do you expect your total annual income to be when you are 30 years old?"

[63] Around 98% of those who responded at 26 also responded at ages 18 and 20.
[64] Some young adults drop out of school before this point (senior high school year). Consequently, I may not be representing this quite small sector of the US population. This is further highlighted where I compare the NELS sample to CPS data in Appendix 3.2.

This question is comparable to those asked in the other major studies of students' labour market expectations[65]. Respondents are clearly asked to predict *their* future income. However this raises several issues about how people actually respond to this type of question. Do they take into account inflation? Do they report gross or net earnings? Is this conditional on having a full-time job? To shed some light on these issues, it is important to consider the ordering of survey questions. Indeed *directly before* they were asked for their income expectations, respondents were given the following question:

*"What was **YOUR** total income from all sources, before taxes, in 1993?* [i.e. the year prior to interview] *This figure should include salaries, wages, pensions, dividends, interest, unemployment compensation, grants, financial aid, scholarships, government assistance (AFDC) and all other income"* [Capitalisation in original question]
Write in dollar amount, write in 0 if no income $..

I assume that respondents follow the same criteria to the above when reporting their income expectations. For example, I assume respondents use their answer to the question above as a reference point for their income expectation and answer with current prices in mind. Indeed, this assumption is consistent with the existing literature. Therefore all reported expectations are assumed to be in 1994 prices. It is also quite clear that respondents should be reporting gross (pre-tax) figures. The question also asks for total annual *income.* This would suggest respondents should not only take into account future wages, but other sources of income such as receipts from benefits or interest payments when reporting their expectations. Another point to note is that the question on income expectation asks for the respondent's ("your") expected income. This is made even clearer in the preceding question, with **"your"** in bold, capital letters. It seems that the respondent should only be considering their own, personal income, and not their partner's or other family members'.

[65] For instance, Webbink and Hartog (2004) phrase their question "How much will your net starting salary be after graduation?" Betts (1996) asks students the question "Below, please circle your estimate of the national average for annual starting salaries". Bruenllo et al (2001) ask "What do you expect to earn right after finishing your degree (first degree possible at your university). State an approximate amount per month (net, i.e. after paying taxes)?"

Dominitz and Manski (1994) state that both men and women respond to questions on income expectations conditional on holding a full-time job. This seems a reasonable assumption for male respondents. However women may expect to have children and be out of the labour force, or working part-time, by the time they are 30. Alternatively women may report their expectation based on working full-time, but by the time they are 30 and have a child, self-select out of the labour market or into part-time jobs. Hence any comparison of women's expectations to later realisations has additional complications. In particular, I do not know how or whether women incorporate selectivity out of the labour force (due to childbearing) into their expectations. I could proceed by assuming that their reported figures are conditional on working full-time (as suggested by Dominitz and Manski). Yet even then I would face the non-trivial task of trying to control for this self-selection in the observed income data. Thus, although I note the potential interest in this issue, and that the problems discussed above are perhaps solvable, I focus on only the 5,782 male responses in this analysis for brevity.

From the initial sample of 5,782 male observations I exclude 633 respondents with missing expectations data from the analysis. A further 39 observations are excluded when an individual reported expected income below $6,000. On the assumption of full-time working, stated above, figures below this level would violate minimum wage laws in the USA. Another 71 observations were dropped where expectations were over $250,000. In total, 743 (13%) male observations have been excluded due to difficulties with the income expectations data.

I investigate the characteristics of those excluded with a logistic regression of item non-response. The results, presented as odds ratios, can be found in Table 3.1.

Table 3.1 Logistic regression of item non-response to question on income expectation

	Specification 1		Specification 2		Specification 3	
	Odds ratio	**Standard error**	**Odds ratio**	**Standard error**	**Odds ratio**	**Standard error**
Socio-economic status index (Ref: Lowest quartile)						
Second quartile	0.85	0.22	0.85	0.21	0.86	0.20
Third quartile	0.58*	0.16	0.56*	0.16	0.59	0.16
Top quartile	0.62	0.16	0.59*	0.16	0.62	0.16
Family income parents reported when respondent was age 18 (Ref: Bottom quartile)						
Second quartile	0.47*	0.11	0.48*	0.11	0.48*	0.11
Third quartile	0.40*	0.10	0.41*	0.11	0.40*	0.10
Top quartile	0.24*	0.07	0.22*	0.07	0.22*	0.07
Race (Ref: White)						
American Indian or Alaska Native	4.52*	2.69	4.13*	2.17	3.16*	1.44
Asian or Pacific Islander	1.44	0.67	1.16	0.56	1.39	0.59
Black, not Hispanic	0.80	0.25	0.75	0.23	0.68	0.20
Hispanic or Latino	1.27	0.26	1.24	0.26	1.24	0.28
More than one race	0.52	0.18	0.54	0.18	0.50	0.18
Reported health problems at 20 (Ref: Yes)						
No	1.00	0.28	1.17	0.37	1.45	0.46
Labour force status at 20 (Ref: Student who does not have a job)						
Student who also has a job	0.84	0.20	0.77	0.19	0.83	0.20
Working only	1.93*	0.46	1.76*	0.41	1.76*	0.42
Not student or working	2.64*	0.73	2.28*	0.65	2.20*	0.65
Housing tenure at 26 (Ref: Homeowner)						
Rent from someone, not a relative	-	-	1.13	0.23	1.14	0.23
Rent from a relative	-	-	1.03	0.31	0.97	0.29
Live in residence without paying rent	-	-	1.77*	0.44	1.50	0.39
Wage at age 26	-	-	-	-	1.01	0.06
Maths ability	-	-	-	-	0.97	0.09
Drop out of university (Ref: No)						
Yes	-	-	-	-	1.07	0.28
Observations	**5,782**		**5,782**		**5,782**	

Notes:
1 The Socio-economic status index is a continuous variable constructed by the survey organisers using data on the respondents' father's education level, mother's education level, father's occupation, mother's occupation, and family income. The higher this index score (or the higher the quartile), the more privileged the child's background.
2 An odds ratio greater than 1 means a greater chance of non-response relative to the reference.
3 "Wage at 26" is a continuous variable that records how much the respondent was earning from employment when they were aged 26. The estimated odds ratio shows how non-response increases with a $10,000 increase in the wage that was earned.
4 "Missing" dummy variables are included when the respondent has not provided information on a covariates. I do not present their results for brevity.
5 "Maths ability" is a continuous variable based upon respondents' scores in a test of their cognitive mathematical ability taken at age 18. The estimated coefficient in the table above shows how a one standard deviation *increase* in maths ability influences the propensity to not respond.
6 * Indicates statistical significance at the 5% level
7 Source: Authors calculations from the NELS dataset. Sample size in all specifications was 5,782. Dependent variable was a binary variable, taking a value of 0 if the respondent did not report their income expectation and 1 if they did

It seems that respondents from the wealthiest backgrounds are the most likely to report their expected income. Similarly, young adults of American Indian descent are less likely to report their expected income than Whites. Of importance for my substantive research questions, it appears that 20 year old students are more likely to respond than their peers in the labour force. More reassuringly, there is no association between wages recorded at age 26 and missing expectations data. In other words, it is *not* the case that those who reported particularly high or low wages at age 26 are the individuals who did not report their salary expectations[66]. Likewise, it does not seem to be the case that missing expectations data is related to the respondents' maths ability or whether they are a student who drops out of university before they complete their degree[67].

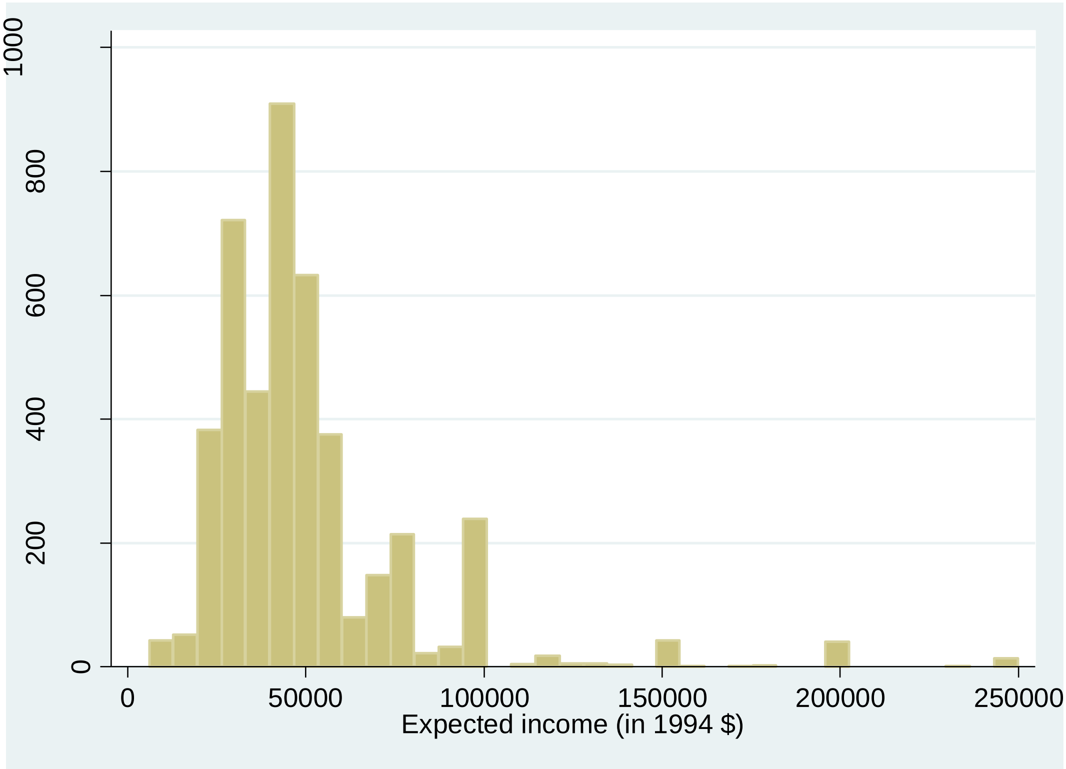
Nevertheless Table 3.1 does indicate some self selection into the study. If those who choose to take part are more (or less) realistic than those who do not, I would underestimate (or overestimate) the difference between average expectations and later realisations. Likewise, the fact that workers are more selective about taking part than students could introduce bias when I consider which of these groups are more realistic (research question 4). For instance, only the most optimistic workers may report their expected income. On the other hand both optimistic *and* cautious students may offer a response. In this scenario, my selection of workers would appear to be less realistic than a true random sample from the population. I have checked the robustness of the results I present in section 3.5 to this non-response by creating and applying a set of response weights, with estimates presented later in the chapter[68].

---

[66] I also tested for an interaction between wage at 26, and whether the respondent was a student or worker at age 20. The coefficient was neither big nor statistically significant.
[67] In a specification not presented, I entered students in groups depending on their subject area. I found no statistically significant differences
[68] Note that the effectiveness of such weights in correcting bias is dependent upon the explanatory power of the underlying non-response model. Table 3.1 indicates that few of the covariates included in the logistic regression are statistically significant. Consequently, one may expect the results to appear no different as model used to create the response weights is relatively weak.

For those individuals with complete expectations data, Figure 3.1 presents the distribution of expected age 30 income. Notice firstly the large, positive skew of the data, with a mean (median) of $50,312 ($40,000) and standard deviation of $30,051. This result is driven by a number of large observations; the top percentile expect to earn over $200,000 per year at age 30 (in 1994 prices). One may ask whether this variable is truly reflecting individuals expectations (what they *think* will happen) rather than aspirations (what they *hope* will happen). In this chapter, I go on to assume the former, but one can not rule out some individuals adhering to the latter. I have excluded some very large observations, where the figures maybe reflecting children's "aspirations" rather than their "expectations". In section 3.5, I present a set of robustness checks using quantile (median) regression to see how results differ when I do not exclude these data.

Figure 3.1 Distribution of expected age 30 income (in 1994 US $) for young US males



Notes: Diagram produced by using the NELS dataset. Sample size = 5,039.

A second feature of the distribution is the bunching of observations. Over half of all observations lie at five points: $30,000 (13%), $40,000 (14%), $50,000 (14%), $60,000 (7%) and $100,000 (3%). I describe a similar phenomenon for UK students in chapter 2. The general explanation is that respondents are rounding their responses to the nearest $5,000 or $10,000. This may reflect uncertainty about future income or, on the other hand, that individuals simply think in terms of round numbers. The implication is that there is some rounding error in individuals reported expectations. When considering expectations at the group or population level, it seems reasonable to assume that this rounding error will be on average zero (some individuals round up, others round down) [69]. On the other hand, considering expectations at the individual level, this is a potentially important source of measurement error.

The NELS also contains data on each respondents' wage history. As part of the survey at age 26, they were asked:

For your (current/most recent) job, about how much (do/did) you earn before taxes and other deductions?

Enter Amount $...

Interviewer instruction: Record the time scale of the amount (e.g. $30,000 per year)

1 = hourly, 2 = weekly, 3 = every two weeks, 4 = monthly, 5 = annually

How many hours per week, in a typical week (do/did) you (currently) work for pay in your job as a/an [F4BOCTV (verbatim job-title)]?

Enter hours per week:………………………Hours

This clearly asks for gross earnings in their current or most recent job. However the respondent could choose the timescale to report this figure, with a breakdown provided in Table 3.2[70].

[69] Obviously this assumes respondents round their expectations as would a mathematician, rather than another rule (for example, always rounding up to the next highest multiple of $5,000).

[70] Note the difficulties when recording salary details because of measurement error, with results based on surveys often different to that held in administrative records. Indeed this measurement error could vary by the unit of time respondents' answer in (see Cartenseen and Woltman 1979).

Table 3.2 Timescale used to report salary

Timescale	% of observations
Hourly	10.4
Weekly	12.3
Twice monthly	3.4
Monthly	8.6
Annually	65.3
Observations	3,475

Notes:
1 I have restricted this data to those who were working full-time year round at age 26.
2 Authors calculations from the NELS dataset.

For those who provided a weekly, fortnightly or monthly figure, I have scaled their pay up to the annual equivalent. All respondents were also asked how many hours they work in an average week. For those reporting an hourly wage, this was used to calculate their annual equivalent.

Wages from previous years were also collected retrospectively at age 26 (the final survey wave)[71]. Respondents were asked:

First, including all of the wages, salaries, and commissions you earned in (1997/1998/1999), about how much did you earn from employment before taxes and all other deductions?

Again gross wages are recorded, containing details on all forms of employment related income, including all commissions, tips and bonuses. Therefore, the NELS data has reported wages for respondents between the ages of 23 and 26[72,73]. Previously, I stated my assumption that respondents are providing their income expectation conditional on holding a full-time job. Thus I only consider realised wages when the respondent was working full-time[74]. Those with no history of full-time work have been excluded from the analysis. Table 3.3 shows that, by doing this, I

[71] Measurement error due to recall bias poses a possible difficulty in using this data. See Bound et al (2001) for further details on the difficulty of recording historical wages with retrospective questions.
[72] I deflate all information on actual wages and unearned income into 1994 prices using data from the Annual Wage Index, available at http://www.ssa.gov/OACT/COLA/awidevelop.html
[73] For 80% of respondents, data on full-time wages is available for at least 3 of these 4 years.
[74] Where gaps appear in individuals wage profiles (between 23 and 26), information from previous years (when they were working full-time) shall be extrapolated forward to estimate age 30 wages. Further details follow in section 3.4 and Appendix 3.2.

exclude a further 605 (12%) observations.

Table 3.3 Missing data on expected income and reported salary

	Observations remaining
All male respondents (Starting sample)	5,782
Individuals with missing expected income data dropped	5,149
Individuals with expected income below $6,000 dropped	5,110
Individuals with expected income over $250,000 dropped	5,039
Individuals with no full time wage observed between ages 23 and 26 dropped	4,434
Final sample	4,434

Notes:
1 Two item non response models are presented in Tables 3.1 and 3.4 that try to explain what factors are associated with missing data. Specifically, Table 3.1 investigates the drop in observations from 5,782 to 5,039 (missing or illogical expectations data). On the other hand, Table 3.4 looks at non-response to the actual salary data (i.e. the drop in observations from 5,039 to 4,434).
2 Authors calculations from the NELS dataset.

Of course, like the missing expectations data, this may introduce some selectivity into the sample. Individuals may choose not to work full-time between the ages 23 and 26. An obvious example is graduate students, many of whom remain in education throughout their early twenties. If these individuals are substantially more or less realistic than other groups, then some selection bias may be introduced into my results. Alternatively, there could be a direct relationship between income expectations and selection out of work. Those with unrealistically high income expectations may also have unrealistically high reservation wages. These individuals are less likely to receive a suitable wage offer, and therefore choose not to work. In this situation, I would be excluding the most unrealistic individuals from the analysis. In Table 3.4 I present a logistic regression that investigates this possible selectivity.

Table 3.4 Logistic regression of missing full-time wage history

	Specification 1		Specification 2		Specification 3	
	Odds ratio	**Standard error**	**Odds ratio**	**Standard error**	**Odds ratio**	**Standard error**
Socio-economic status index (Ref: Lowest quartile)						
Second quartile	1.06	0.31	1.05	0.31	1.06	0.36
Third quartile	0.93	0.24	0.88	0.22	0.96	0.28
Top quartile	1.09	0.34	1.02	0.33	0.84	0.30
Family income parents reported when respondent was age 18 (Ref: bottom quartile)						
Second quartile	0.77	0.16	0.79	0.16	0.80	0.19
Third quartile	0.59*	0.14	0.61*	0.14	0.63	0.17
Top quartile	0.32*	0.07	0.35*	0.07	0.37*	0.09
Race (Ref: white)						
American Indian or Alaska Native	0.55	0.57	0.72	0.74	0.91	0.88
Asian or Pacific Islander	2.86*	0.94	2.39*	0.73	1.59	0.44
Black, not Hispanic	1.03	0.33	1.07	0.33	0.82	0.22
Hispanic or Latino	0.85	0.24	0.86	0.25	0.79	0.24
More than one race	1.15	0.45	1.12	0.46	0.87	0.30
Reported health problems at 20 (Ref: Yes)						
No	0.37	0.16	0.41*	0.17	0.33*	0.22
Labour force status at 20 (Ref: Student who does not have a job)						
Student who also has a job	1.01	0.20	1.07	0.21	0.98	0.24
Working only	0.75	0.19	0.96	0.24	1.00	0.26
Not student or working	1.76	0.55	2.04*	0.63	1.84*	0.50
Housing tenure at 26 (Ref: Homeowner)						
Rent from someone, not a relative	-	-	2.25*	0.49	1.47	0.36
Rent from a relative	-	-	2.29*	0.87	1.52	0.70
Live in residence without paying rent	-	-	3.43*	0.82	1.82*	0.48
Expected income	-	-	1.01	0.01	1.01	0.01
Maths ability	-	-	1.27*	0.06	1.21*	0.07
Drop out of university (Ref: No)						
Yes	-	-	1.04	0.23	0.76	0.20
Working status at 26 (Ref: Working full-time)						
Work part time	-	-	-	-	4.57*	1.12
Study only	-	-	-	-	24.15*	6.27
Work full time & study	-	-	-	-	0.79	0.29
Work part time & study	-	-	-	-	15.88*	3.98
Neither work or study	-	-	-	-	18.93*	5.04
Observations	**5,039**		**5,039**		**5,039**	

Notes:
1 This table investigates the characteristics of the 605 young men who did not have a full time wage recorded at any point between the age 23 and 26
2 An odds ratio greater than 1 means a greater chance of non-response than the reference
3 See notes to Table 3.1 for details on the Socio-Economic Status Index and “Maths Ability” variables
4 “Expected Income” is how much a $10,000 increase in expected wage influences the chance of response.
5 * indicates statistical significance at the 5% level
6 Authors calculations from the NELS dataset. Sample size in all specifications was 5,039. Dependent variable was a binary variable, taking a value of 0 if the respondent did not any information available on their actual wage history and 1 if they did

Respondents who have parents in the top quartile of the income distribution are three times more likely to respond than their peers whose family income is in the bottom quartile. Similarly, those with reported health problems are more likely to be excluded than those without. On the other hand, it seems that respondents who were students at age 20 are just as likely to be excluded as those who were working. However, those who were unemployed at age 20 are relatively unlikely to have a full-time wage recorded between the ages of 23 and 26. Reassuringly, there is little evidence that those with the highest wage expectations were the least likely to be working full-time between the ages 23 and 26. Interestingly, specification 3 and 4 show that low ability respondents were *less* likely to be excluded from the analysis because of missing income data. It seems the brightest sample members tend to either not report their salary or have selected out of full-time work up to age 26 (e.g. to continue their education). If individuals of high ability are more efficient at processing labour market information and thus hold more realistic expectations, then their exclusion may have an influence on my results. The final specification shows that those who are still studying at age 26 are the most likely to be excluded. Further analysis not presented indicated that around half the excluded observations came from individuals who were studying full-time at age 26. It is likely that these individuals have never left higher education, and hence have no full-time wage history. The main implication seems to be that certain groups of students are likely to be excluded from the analysis; particularly those who continue onto graduate school and remain in education through their early 20's. If these individuals hold significantly more (or less) realistic expectations than other groups of students, then my results could again be influenced by their exclusion.

I further investigate for selection from missing data in Table 3.5 by presenting a set of summary statistics. The left hand column illustrates the characteristics of the initial 5,782 male observations in the complete NELS sample, while the column on the right shows the characteristics of the 4,434 individuals who are not missing any key information. Reassuringly, the distribution of observable characteristics remains reasonably similar.

Table 3.5 Summary statistics showing the NELS sample composition, before and after the exclusion of missing expectations and wage data

	Starting sample %	Final sample %
Labour force status at age 20		
Students who also have a job	26.6	26.0
Students who do not have a job	27.0	28.4
Working, not a student	35.0	35.8
Neither student or working	11.3	9.8
Highest qualification at age 26		
Less than high school	6.0	5.3
High school	55.6	56.2
Associates degree	7.1	6.9
Bachelors	28.2	28.6
MA/PhD	3.1	3.0
Race		
White	66.6	68.5
American Indian or Alaska Native	1.0	0.8
Asian or Pacific Islander	5.5	5.1
Black, not Hispanic	8.2	8.0
Hispanic or Latino	13.1	13.1
More than one race	3.0	2.6
Other	2.6	1.8
Family income student reported at age 18 ($1992)		
0-20000	18.5	18.1
20000-35000	19.8	19.8
35000-50000	17.7	18.4
50000-75000	16.0	16.5
75000+	12.3	12.4
Missing	15.7	14.9
University subject at 20 years old (If reported being a student)		
Agriculture	1.9	2.3
Accounting, Finance	6.1	6.2
Business Management	12.7	13.1
Journalism, Communication	3.3	3.6
Computer Science, Maths	4.8	5.4
Education	5.1	5.4
Engineering, Physical Sciences	16.9	17.4
Languages	1.8	1.7
Health	6.8	6.2
Law	4.2	3.9
Biological Science	7.4	6.6
Social Sciences, Humanities	9.1	9.1
Arts	5.0	4.6
Other	14.9	14.6
Working full-time At age 26		
Yes	74.0	84.0
No (e.g. unemployed, student, working part-time etc)	26.0	16.0
Observations	5,782	4,434

Notes:
1 “Starting sample” refers to all men in the age 26 sweep of the NELS. “Final Sample” refers to the sample I use in my analysis, once I have excluded missing data
2 Authors calculations from the NELS dataset.

Though the primary focus of this chapter is young men's expected income, I present some additional results referring to other aspects of their anticipated labour market success as a robustness check. In particular, I put forward the argument that if students are unrealistic about their future income, they are also likely to be unrealistic about other aspects of the labour market, like their future occupation. Analogous to finding excessive income expectations, individuals may expect to be in a professional occupation when they turn 30, but actually end up working in a relatively low paying job. As part of the NELS survey at age 20, individuals were asked what occupation they thought they would be working in at age 30[75]. In the final survey wave (age 26) individuals were asked what occupation they currently hold. Therefore I also compare expected and realised occupation to support my main analysis surrounding young adults' income expectations.

At this point, however, one should note that there are two significant problems with comparing expectations and realisations using the NELS data:

(a) At age 20, respondents were asked what they expect their income (and occupation) to be at age 30. However, data on labour market realisations is only collected between the ages 23 and 26.
(b) Respondents are asked about their expected *income.* Data on realisations focuses on *wages*.

I go on to discuss these points in section 3.4 and Appendix 3.2. Specifically, these sections cover how I use the information available to predict individuals age 30 income. To conclude this section, I simply ignore such issues and compare expectations of *income* at age 30 to realised *wages* at age 26.

Figure 3.2 presents the distribution for age 26 (actual) wages.

[75] The exact wording of the question was: "What job do you expect or plan to have when you are 30 years old?" Respondents were asked to write in an occupational description into an open text field.

Figure 3.2a Distribution of age 26 actual income (in 1994 US $) for young US males

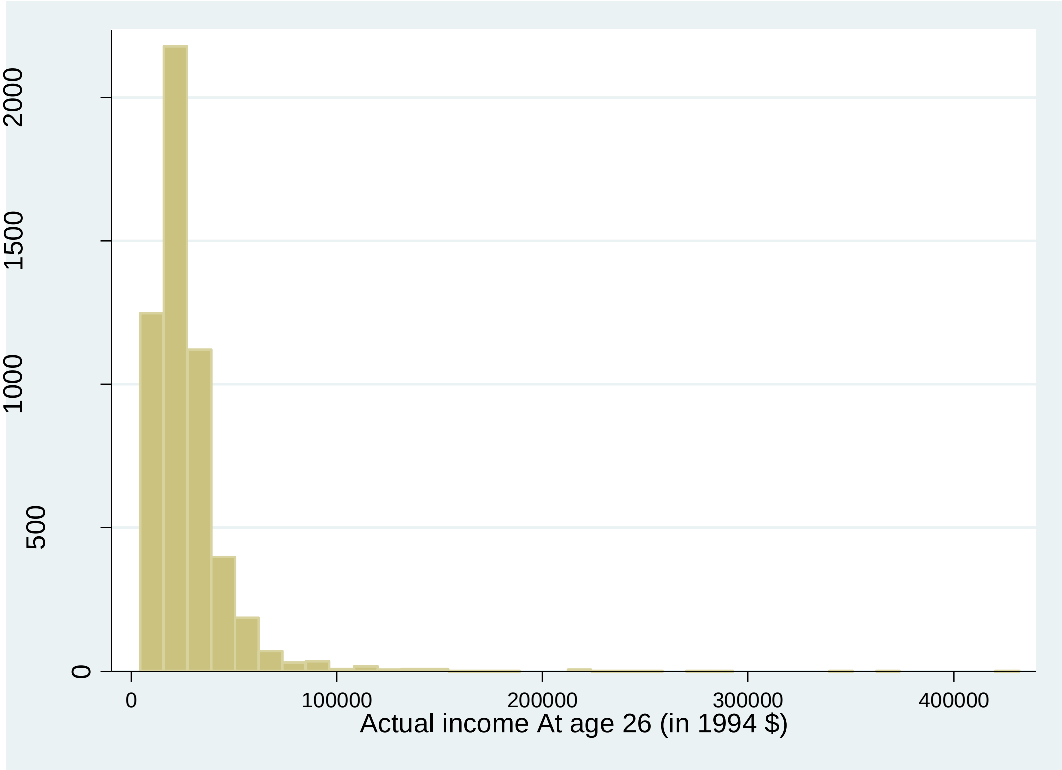
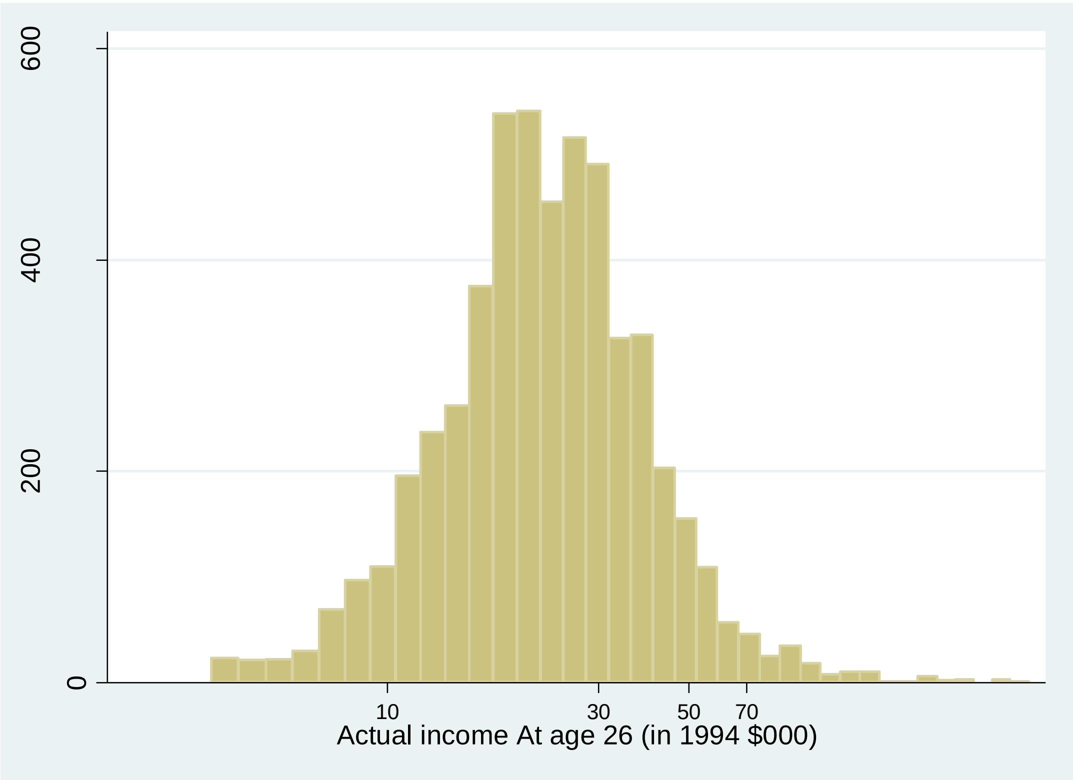


Figure 3.2b Distribution of <u>log</u> age 26 actual income (in 1994 US $) for young US males



Notes: 1 Source Authors calculations from the NELS dataset. Sample size = 4434

Comparing this to the expected income distribution at age 30 in Figure 3.1, there is significantly less bunching and positive skew. The standard deviation is much smaller ($16,479 for the actual age 26 wage distribution compared to $30,051 for the age 30 expected income distribution), though there is little difference in the decile ratio (3.4 compared to 3.2). Further insights come from investigating the ratio of the 10th to 50th percentile (p10/p50) and the 90th to 50th percentile (p90/p50) . The bottom halves of the distributions (p10/p50) are very similar (0.6 in the actual distribution compared to 0.625 in the expected). The difference is slightly bigger in the top half of the distribution, with the p90/50 for actual wage (1.8) below that in the expectations (2.0). There is also some initial evidence that young adults' expectations may be somewhat optimistic. The median (mean) expected income at age 30, in 1994 prices, is $40,000 ($50,312). Comparatively, the median (mean) actual wage of 26 year olds stands at $23,079 ($26,210). For expectations to be realistic on average by age 30, I would need to find that either:

(a) average annual real wage growth is around 15% between the ages 26 and 30.
or
(b) 30 year old men, on average, have large quantities of unearned income.

I turn to these two topics in the following section.

3.4 Prediction of age 30 income

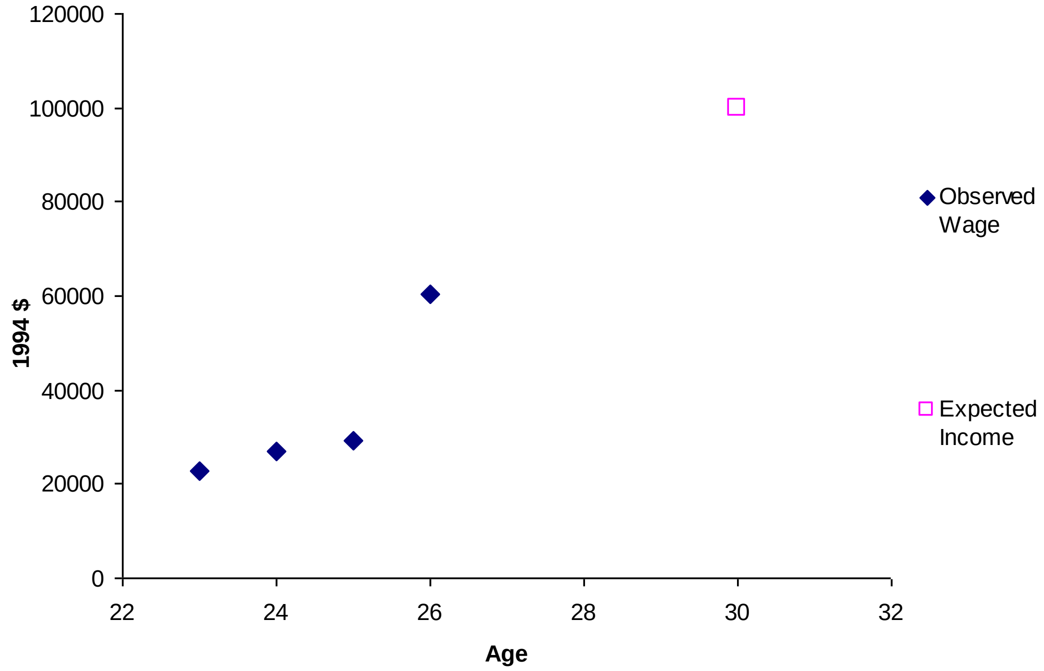
Respondents are asked what they expect their annual income to be when they turn 30, but realised wage data is only available between ages 23 and 26. In Appendix 3.2 I fully set out two methods of predicting age 30 income. In this section, I will provide the intuition behind these methods. Specifically, I separate this into two parts: (a) the estimation of wages, and (b) the estimation of unearned income.

Wages

Figure 3.3 illustrates the data observed for one particular individual in the NELS[76]. At age 26, this individual has a particularly large wage by his "historical" (age 23- 25) standards. This may be a *permanent* shift in his wage profile, for instance a change in career. In this case, previous earnings have little relevance for predicting future wages. In contrast this could be a *temporary* shock to his wage, for instance a salesman who has had a particularly good year. In future periods, his wage will revert to its historical average (i.e. the average of the previous 3 years). On the other hand, reality may lie somewhere in-between these two extremes, perhaps reflecting the fact that he happened to receive a large pay rise that year.

[76] For this particular individual, the income they expect is significantly higher than their wages recorded at age 26. This is not typical of all other respondents in the dataset. Rather I have chosen this individual as he is a good example of the substantial points I make throughout this section.

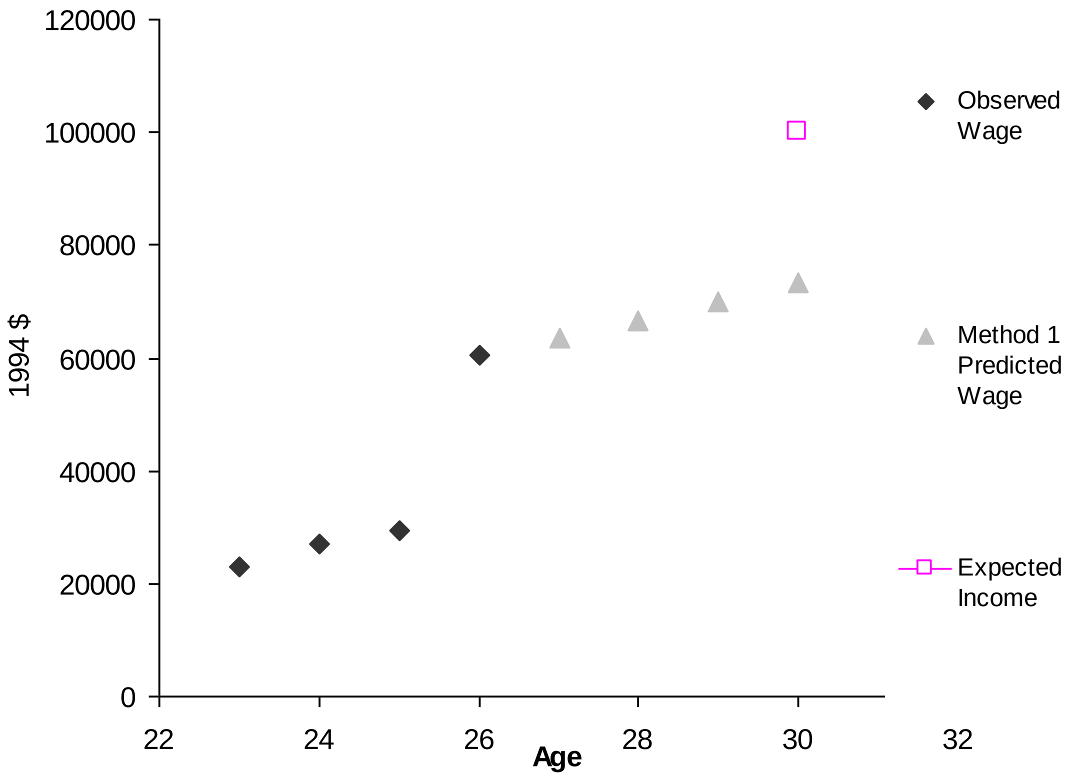
Figure 3.3 Data on expected income and actual wages that can be observed for one particular individual in the NELS



Note:
1 This individual reported zero unearned income at age 26; therefore his wages are equivalent to his income. Note that in these diagrams, I am simply trying to explain my extrapolation method for *wages*. This individual is not an example of a "typical" NELS respondent. Rather, I have chosen this observation as it provides a good example of the points I am tying to make.
2 Source: Authors calculations from the NELS dataset.

Given these possibilities, I use two methods to predict age 30 wages. The first method views large wage changes as a *permanent* shift in an individual's earnings profile. Under this method, I simply take the most recently observed wage (age 26) for each individual and extrapolate it forward (to age 30). Figure 3.4 presents a hypothetical example.

Figure 3.4 Illustration of wage prediction method 1 for ID 7286532 in the NELS



Note:
1 See notes to Figure 3.3
2 The above is a hypothetical example of extrapolation method 1.

To implement this method, I use estimates of the annual real wage growth for young workers provided by Rubinstein and Weiss (2007). Specifically they provide a table of average annual real wage growth rates broken down by labour market experience and educational attainment for three surveys; the Current Population Survey (CPS), Panel Survey of Income Dynamics (PSID) and National Longitudinal Survey of Youth 1979 (NLSY 79)[77]. The growth rates they calculate from the CPS, PSID and NLSY are provided in Table 3.6, with further details available on page 14 and Appendix 5 of Rubinstein and Weiss (2007).

[77] Rubinstein and Weiss restrict each of the above datasets to full-time, male, American workers, as I have done with the NELS. One should note, however, that these surveys all relate to different years. The CPS data relates to wages between 1998 and 2002, the PSID is for all years after 1968, while the NLSY draws its information between 1979 and 2000.

Table 3.6 Average, annual (real) wage growth rates for young workers: Rubenstein and Weiss estimates

		% Average (real) wage growth rate per annum by education level				
Number of years experience in the labour force	**Data source**	**Below high school**	**High school**	**Some college**	**College graduates**	**MA/PhD**
0-10	CPS	2.4	3.2	3.3	3.6	2.9
	PSID	2.8	3.0	3.8	3.9	3.2
	NLSY	2.4	3.4	4.6	5.2	5.5
11-15	CPS	1.6	2.2	-	-	-
	PSID	1.9	2.0	-	-	-
	NLSY	1.3	2.3	-	-	-

Notes:
1 Source: Table 1, page 14 of Rubinstein and Weiss (2007) Post Schooling Wage Growth: Investment, Search and Learning. Handbook of the Economics of Education, Volume 1

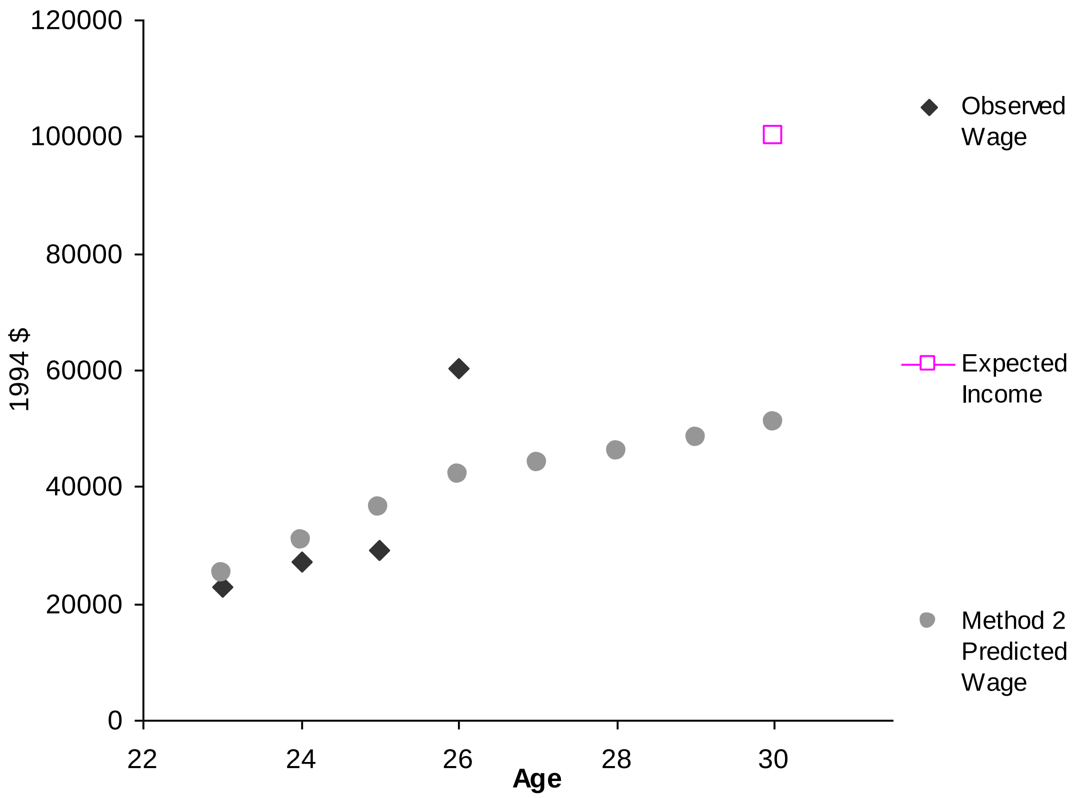
These growth rates are applied to each individual in the NELS, depending on their highest qualification achieved by age 26[78]. From this point on, I shall call this prediction "Method 1". Note that, for all groups, average real wage growth rates are under 6% per annum. This is well below the 15% per annum that I suggested NELS sample members needed in the unobserved period (i.e. between ages 26 and 30) for their expectations to be (on average) "realistic" (recall my brief discussion at the end of section 3.3).

On the other hand, prediction "Method 2" views large changes in earnings as a *temporary* shift in an individual's wage profile[79]. Thus individuals' wage *history*, rather than just the most recent observation, is now informative for estimation of future wages. A hypothetical example is shown in Figure 3.5.

[78] For example, an individual with college education, and who was earning $50,000 dollars at age 26, would be estimated to be earning $61,240 at age 30 (all in 1994 prices). This is calculated by $50,000 * ($1.052^4$), using the NLSY data and "College graduates" column in Table 3.6. In the event that wages go unobserved at age 26 (e.g. the individual was unemployed) I extrapolate from their last observed full-time wage. For example, if someone was earning $50,000 at age 25, and their wage was not recorded at age 26, I would predict their age 30 income to be $50,000 * ($1.052^5$) = $64,577

[79] I am using the term "temporary" in a slightly different manner here compared to page 102. Specifically, for the illustrative individual in Figure 3.3 I do not assume that their wage growth reverts to their age 23-25 trajectory (and thus that his wage at age 26 contains no useful information in predicting age 30 wages at all). Rather I allow the age 26 wage to have some permanent impact on my prediction of his age 30 wage, but for it to be tempered by what they were earning between ages 23 and 25.

Figure 3.5 Illustration of wage prediction method 2 for ID 7286532 in the NELS



Note:
1 See notes to Figure 3.3
2 The above is a hypothetical example of extrapolation Method 2.

To implement this method, I use a fixed effects regression model following the methodology of Carneiro and Heckman (2003). Appendix 3.2 describes this prediction method in detail, including model specifications and robustness checks. I also show in Appendix 3.2 that this produces *average* age 30 wage estimates that are very similar to those from "Method 1" (though wage estimates from method 2 suffer from less variability)[80]. The similarity of average wage predictions across methods is due to the "shocks" that are incorporated in method 1 being both positive and negative (hence cancelling each other out on average).

[80] This is because outlying observations are moderated in Method 2 by the influence of previous wages (it is a time mean). This does not occur in Method 1, where it is only the most recent observation that is used for prediction. Hence if there is a large shock to the most recent observation, this gets carried forward to the future prediction in Method 1, as opposed to being averaged out in Method 2.

In Table 3.7 I compare my predictions from these two methods to similar information recorded for 30 year olds in an external data source (the 2003-2005 CPS March Annual Supplements[81]).

Table 3.7 Predicted *mean* age 30 NELS wage compared to the *mean* age 30 Current Population Survey (CPS) wage

	% of observations in NELS	Predicted wage method 1 ($000)	Predicted wage method 2 ($000)	% of observations in CPS	CPS wage ($000)
Highest qualification at age 26					
Below high school	5.4	20.9	20.8	12.4	15.8
High school	56.4	27.0	25.5	47.8	24.7
Associates degree	7.0	29.9	30.5	8.2	28.1
Bachelors	28.0	37.7	37.8	24.6	38.1
Masters degree / PhD	2.8	42.4	43.6	6.8	44.5
Race					
White	69.8	31.4	31.3	60.6	31.5
American Indian	0.1	23.6	24.6	0.2	NA
Asian or Pacific Islander	5.1	38.5	36.4	6.2	35.4
Black (not Hispanic)	8.3	24.6	25.1	9.6	26.9
Hispanic	13.3	27.4	27.1	21.7	20.7
Other	4.6	27.1	26.6	1.7	NA
All respondents	100.0	30.4	29.6	100.0	28.9

Notes:
1 All observations in 1994 $
2 See notes to Appendix Table A3.8 for further details
3 The three left-hand most columns refers to authors calculations from the NELS dataset, the two columns on the right are based on data from the CPS taken from from 2002-2004 March Annual Supplement, restricted to men working full-time, all year round, at age 30 available from : http://www.census.gov/hhes/www/cpstc/cps_table_creator.html. Total sample size in the NELS is 4,434. In CPS, the total sample size is 1,412

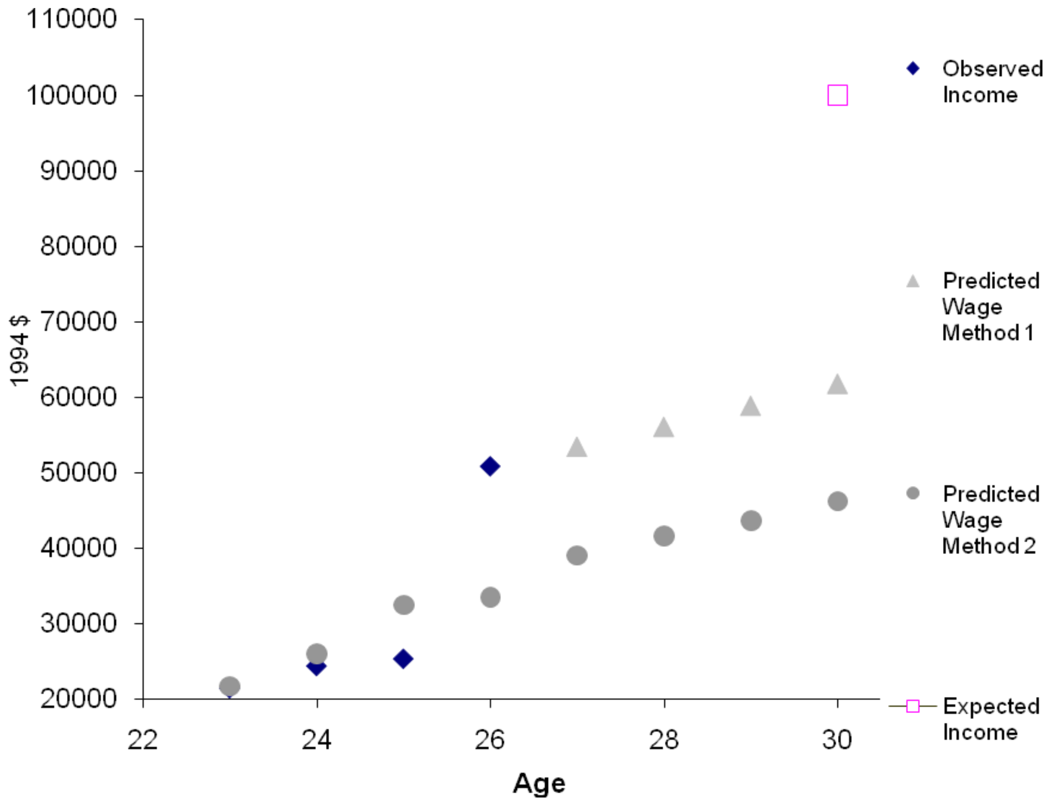
In general, average predicted age 30 wages appear similar to those in the external CPS data. I predict average wages to be $29,600 in the NELS, while in the CPS the equivalent figure is $28,900. Likewise, my predictions of average wages seem to be reasonably close to the CPS data for a number of sub-groups (e.g. those who are white or holding a bachelors degree), though there are instances where this is not the case (e.g. those with less than high school education and Hispanics)[82]. It is worth noting,

[81] The exact wording to collect income and wage data in the CPS is comparable to that used in the NELS.

[82] I predict those with less than high school education to earn around $21,000 while the CPS figure stands at just over $15,000. My definition of "less than high school education" is those who made it

however, that both these groups only form a small part of the overall sample. The general message is that my prediction methods seem to generate a reasonable estimate of *average* age 30 wages. Further evidence of this can be found in Appendix 3.2. However, in Figure 3.6, I show that my two wage predictions for the illustrative NELS respondent are $15,000 (30%) apart. Anywhere between the two predictions, or even a figure outside of this range, could be possible. Thus a comparison of expected and actual wages at the *individual* level does not seem a sensible approach with the NELS data. On the other hand, when dealing with group averages, over-estimates of wage growth for some individuals will be compensated by underestimates for others.

Figure 3.6 Comparison of wage prediction methods for ID 7286532 in the NELS



Note:
1 See notes to Figure 3.3, 3.4 and 3.5

---

into the final year of high school but did not graduate. On the other hand the CPS represents the whole US population, and defines less than high school education as everyone who did not graduate from high school, *including* those who dropped out *before* their senior year. This is probably the reason why, in the NELS comparing to the CPS, my predicted wage is higher and there are a smaller proportion of respondents with below high school education. In a similar manner, I predict average wages for Hispanics to be around $27,000, while the CPS figure is closer to $20,000.

Unearned Income

At age 26, respondents' were asked about their non-wage income at age 25, with 74% of individuals reporting no unearned income[83]. Unearned income may make up a more significant proportion of total income at age 30 than at age 26. To investigate this, I compare mean wages to the mean total income for 30 year old men in the 2003-2005 CPS March Annual Supplement[84]. Mean total income for this group is only $500 higher than mean wages. This suggests that "other" sources of income make up only a small fraction (roughly 2%) of 30 year old men's total income (on average). I also investigate the extent of unearned income reported in another American data source (the NLSY 1979), again finding that it has very little impact on the average individual (the median unearned income is zero)[85]. Therefore, to incorporate unearned income into my predictions, I simply use the value recorded at age 25 in the NELS. Given the minor contribution this makes to individuals total income, this should not introduce substantial bias at the group or population level (the same, however, is unlikely to be true if one were to try and make inferences at the individual level).

Summary

I have presented two methods to predict age 30 wages, both of which are comparable with external estimates from population level data. Moreover, even though age 30 unearned income is difficult to predict, this makes up only a small proportion of total average income at the group or population level. I am therefore confident that the substantive inferences in section 3.5 regarding population and group level averages are robust to the data issues discussed throughout this section. However, inferences made at the individual level are likely to suffer from what may be quite severe biases. Thus I choose not to conduct such analysis in this chapter. For a more detailed discussion of these issues, I encourage the reader to turn to Appendix 3.2, where I present a full description and justification of the methods used.

---

[83] The exact wording of this question can be found in Appendix 3.2, and asks respondents to include income from savings, stocks and bonds along with any child support, family or disability payments.

[84] Several questions about other (unearned) sources of income were asked in the CPS. This includes how much they received from benefits, welfare, assistance, dividends and interest. Hence the definition of "other income sources" seems largely comparable with that applied in the NELS (though an obvious difference is that the NELS asks for this information in a single question, compared to several component parts in the CPS). The data I use is drawn from the CPS "Table Creator", available from http://www.census.gov/hhes/www/cpstc/cps_table_creator.html I produce two values, one looking at men's average wages, the other their total income. I assume that the difference between these figures (average wages and average income) equals total income from unearned sources.

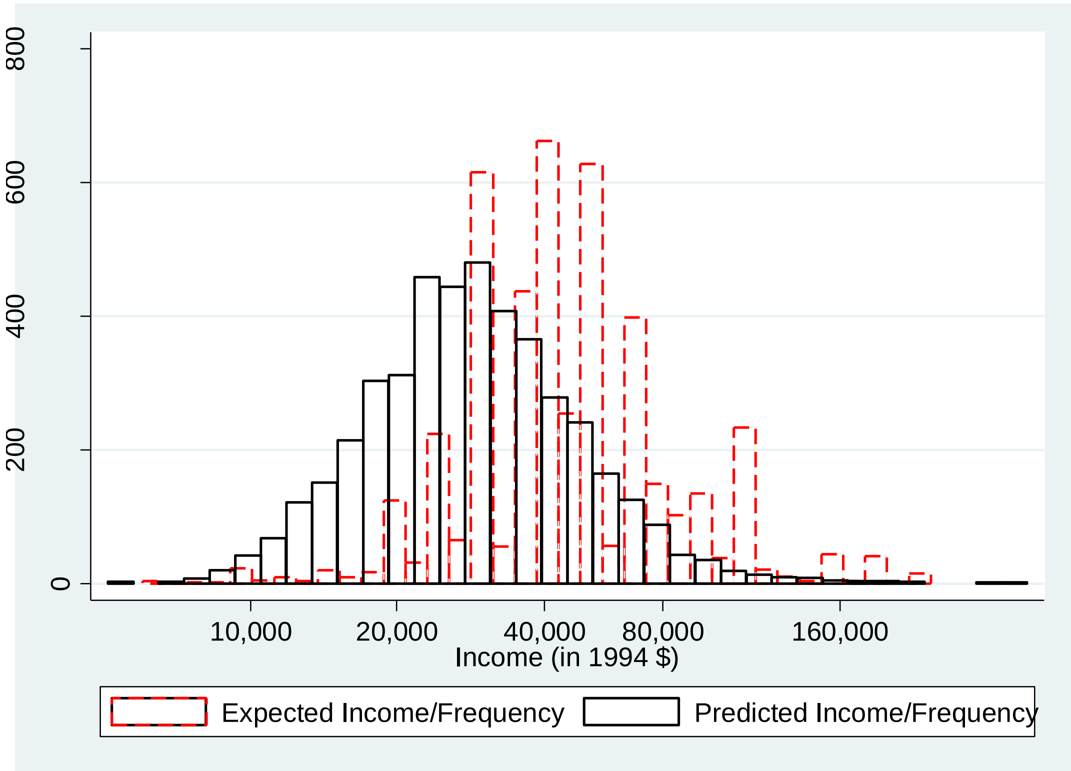
[85] Infact 62% of men report having no unearned income.

3.5 Results

In this section, I will compare students' expectations to my prediction of their income at age 30. Throughout this discussion, I will focus on the results using prediction "Method 2" (fixed effect extrapolation model) from the previous section. Results using "Method 1" are generally consistent to those presented, with a discussion in Appendix 3.3.

Before investigating the relative accuracy of students' and workers' expectations, Figure 3.7 presents the distributions for expected and predicted age 30 income across all individuals (i.e. both students and workers).

Figure 3.7 Distribution of log expected and log predicted income at age 30



Note:
1 All data in 1994 prices
2 Predicted Income refers to that estimated using prediction method 2.
3 Dashed bars refers to distribution of expected income at age 30, solid bars refer to my predictions of actual income at age 30
4 Source: Diagrams produced by the author using the NELS dataset. Sample size = 4,434

Expectations (dotted lines) are clearly to the right of the predicted age 30 income distribution (solid lines). Very few 20 year olds expect to earn less than $20,000 at age 30, though I predict that almost a quarter do. Conversely, there is quite a significant minority (3%) expecting to earn $100,000 or more, though in reality very few (1%) reach this milestone. Indeed, the median predicted income is $26,695 compared to expectations of $40,000, an average overestimation of around 50%[86].

I check the robustness of this result in Table 3.8, which illustrates the proportion of adults expecting to be in each occupation by age 30, and the actual proportion in each by age 26[87]. The last column gives the median wage for workers of all ages in each occupational group in 2004, drawn from CPS data, to give an idea of the financial status of each occupation (note Table 3.8 is ordered by this column).

[86] Note that here I am discussing the median. In Table 3.7, where I compared predicted age 30 wages to data from the CPS, I am discussing the mean.
[87] Of course, some young adults are still in education at age 26, who are likely to be working professionals by age 30. However, this group is only relatively small, and are contained within the 4.5% described as "not working/studying/homemaker". Even if I assigned this group to a professional category, I would still find large overestimation in the results.

Table 3.8 Proportion of 20 year olds expecting to enter each profession by age 30, and the proportion who have actually entered that profession by age 26

	% expecting to be in the occupation at age 30	% actually in each occupation at age 26	Difference between expected and actual (% points)	Average annual CPS wage for each occupation in 2004 (deflated to 1994 $000)
Homemaker, not working, studying	0.6	4.5	-3.9	0.0
Farmer	2.1	0.8	1.2	12.5
Labourer	2.0	9.3	-7.3	16.7
Service	1.1	3.1	-2.0	17.0
Skilled operative	3.1	7.7	-4.6	19.6
Clerical	1.5	6.2	-4.7	19.9
Craftsman	9.1	12.2	-3.1	20.6
Sales	2.3	6.3	-4.0	22.7
Protective services	7.7	3.7	4.0	24.9
Arts, Entertainment, Writing	6.8	1.7	5.1	29.3
Teacher	6.0	3.3	2.6	32.5
Professional Medicine (not Doctor)	5.5	1.2	4.3	35.0
Other Professional	11.8	6.0	5.9	35.6
Engineer	8.4	4.0	4.4	38.7
Computer technical	3.5	6.7	-3.1	39.2
Manager	11.4	12.7	-1.2	41.3
Legal	3.3	0.6	2.7	53.0
Doctor	3.4	1.1	2.3	63.7
Military	1.8	1.4	0.4	NA
Proprietor	8.5	7.3	1.2	NA
Observations	**4,218**	**4,368**		

Notes:
1 The difference column is the expectation % minus the actual %.

2 The number of observations differs due to missing data. In total, 4368 of the sample had an occupation recorded by age 26. Some of these individuals reported that they "did not know" what occupation they expected when asked at age 20 (hence a sample size of 4218)

3 The average CPS wage relates to the mean wage in each occupation for all workers above age 16 in 2004. This is the year the NELS sampled turned 30. I have included and ranked occupations by this information to give an objective measure of occupational status. Data is not available for military occupations and business owners Source: Table 39 http://www.bls.gov/cps/cpsa2004.pdf The weekly wages have been converted to annual equivalents and deflated to 1994 prices, using data from the US government social security office http://www.ssa.gov/OACT/COLA/awidevelop.html.

4 Source: Authors calculations from the NELS dataset and the CPS data (referred to above).

Although one can only make quite a crude comparison, as the data relate to expectations and realisations at different ages, it nevertheless illustrates that young adults also seem to be overly ambitious in their occupational expectations[88]. There are fewer individuals in the highest paying occupations (engineers, arts, doctors) and more in the less well paid (sales, services and clerical roles) than expected. Moreover, note that in the column labelled "Difference between expected and actual", negative figures tend to sit near the top of the table and positive numbers at the bottom. This also suggests that young adults occupy lower paying jobs than they previously expected.

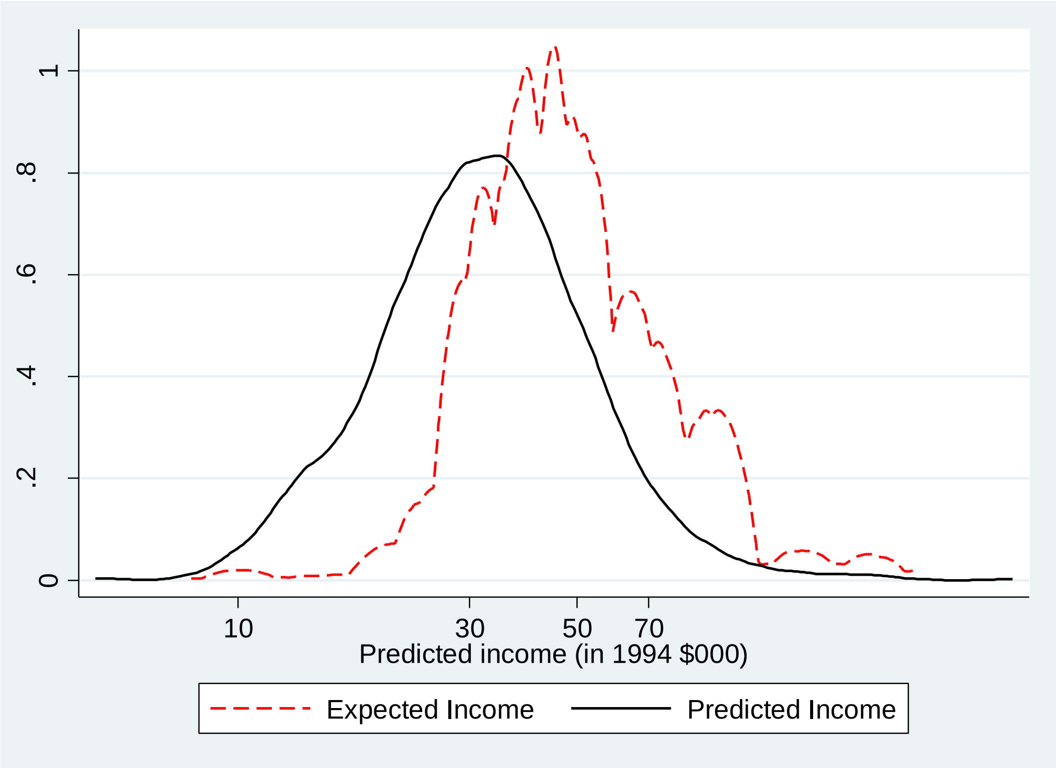
To summarise Table 3.8, I derive an "expected" and "predicted" income from this occupational data. Specifically, I use the reported proportions in each occupation as weights (i.e. column 2 as weights for expected income, column 3 for actual income), which I multiply by the occupation specific CPS wage (column 4). Using this method, I find that young adults expect an income of $29,683, but I predict them to actually obtain $24,538. Hence they overestimate their future income by 20%. Though this figure is significantly below the 50% found above, one should remember that this method captures just one aspect of the underlying issue. Even though a young man may be able to predict his occupation, he may overestimate the general pay that is received in that profession, or expect to be further up the career ladder than he actually achieves (e.g. expecting to become an army Sergeant by age 30, but only ending up a Private).

In general, this preliminary analysis strongly suggests that young men overestimate their future labour market success. They expect an average income of $40,000, but in reality I predict their annual earnings to be less than $30,000. Moreover, many young men expect professional work that they do not go on to achieve.

[88] Of course, there is an issue that I observe individuals at 26 rather than 30, and they could change occupation over the unobserved period. However, if respondents were asked what occupation they expected to have at age 26, would one really expect them to give a substantially different answer? I believe that expected occupation at 26 and expected occupation at age 30 would be highly correlated, and for this to be a reasonable proxy.

Figure 3.8 presents results, analogous to the above, for just those sample members who were still in education at age 20.

Figure 3.8 Distribution of log expected and log predicted income at age 30 for students



Note:
1 See notes Figure 3.7
2 Source: Diagrams produced by the author using the NELS dataset. Sample size = 2,412

Clearly, the results are very similar to those presented above. The median predicted income (using Method 2) is $30,187 compared to expectations of $45,000; students overestimate their future income by around 50%. Likewise, Table 3.9 illustrates how the career expectations of young men still in education match with their eventual occupational attainment by age 26.

Table 3.9 Proportion of 20 year old *students* expecting to enter each profession by age 30, and the proportion who have actually entered that profession by age 26

	% expecting to be in the occupation at age 30	% actually in each occupation at age 26	Difference between expected and actual (% points)	Average CPS wage for each occupation in 2004 (deflated to 1994 $000)
Homemaker, not working, studying	0.6	4.8	-4.2	0.0
Farmer	2.0	0.7	1.2	12.5
Labourer	0.3	5.1	-4.9	16.7
Service	0.7	3.6	-2.8	17.0
Skilled operative	0.7	3.4	-2.7	19.6
Clerical	1.0	7.1	-6.0	19.9
Craftsman	3.5	6.5	-3.0	20.6
Sales	2.4	8.0	-5.6	22.7
Protective services	6.0	3.7	2.3	24.9
Arts, Entertainment, Writing	8.5	2.5	5.9	29.3
Teacher	8.2	5.5	2.7	32.5
Professional Medicine (not Doctor)	7.1	1.3	5.8	35.0
Other Professional	14.5	8.0	6.5	35.6
Engineer	11.0	6.2	4.8	38.7
Computer technical	3.8	9.2	-5.4	39.2
Manager	12.8	13.7	-0.9	41.3
Legal	4.8	1.2	3.6	53.0
Doctor	4.9	1.6	3.4	63.7
Military	0.7	1.5	-0.8	NA
Proprietor	6.5	5.8	0.7	NA
Observations	**2,306**	**2,410**		

Notes:
1 See notes to Table 3.8
2 4 Source: Authors calculations from the NELS dataset and the CPS data (referred to Table 3.8).

As in Table 3.8, many students expect to work in professional careers, but end up in less prestigious jobs. For instance, whereas 8.5% expect to become artists or entertainers, only 2.5% work in these occupations by age 26. Likewise, only 6.2% are engineers at age 26, though around 11% thought they would be working in this profession. Again, when using the data in columns 2 and 3 to weight column 4 (as on page 113), I find students expect to earn, on average, $33,465 but their actual average income is $27,097; overestimation of around 25%.

There seems sufficient evidence to conclude that students overestimate their future success in the labour market. I find they overestimate their income at age 30 by, on average, 50%. Likewise, I find that many graduates are working in service, clerical and sales roles that as 20 year old students they did not anticipate doing for a career. These findings complement results from Smith and Powell's (1990) study of two mid-western universities. They found students overestimated their salary at age 30 by around 40%. It seems that this general result holds across the wider US student population.

I now turn to the results for young adults who were already in the labour market when asked for their income expectations. In section 3.2, I argued that:

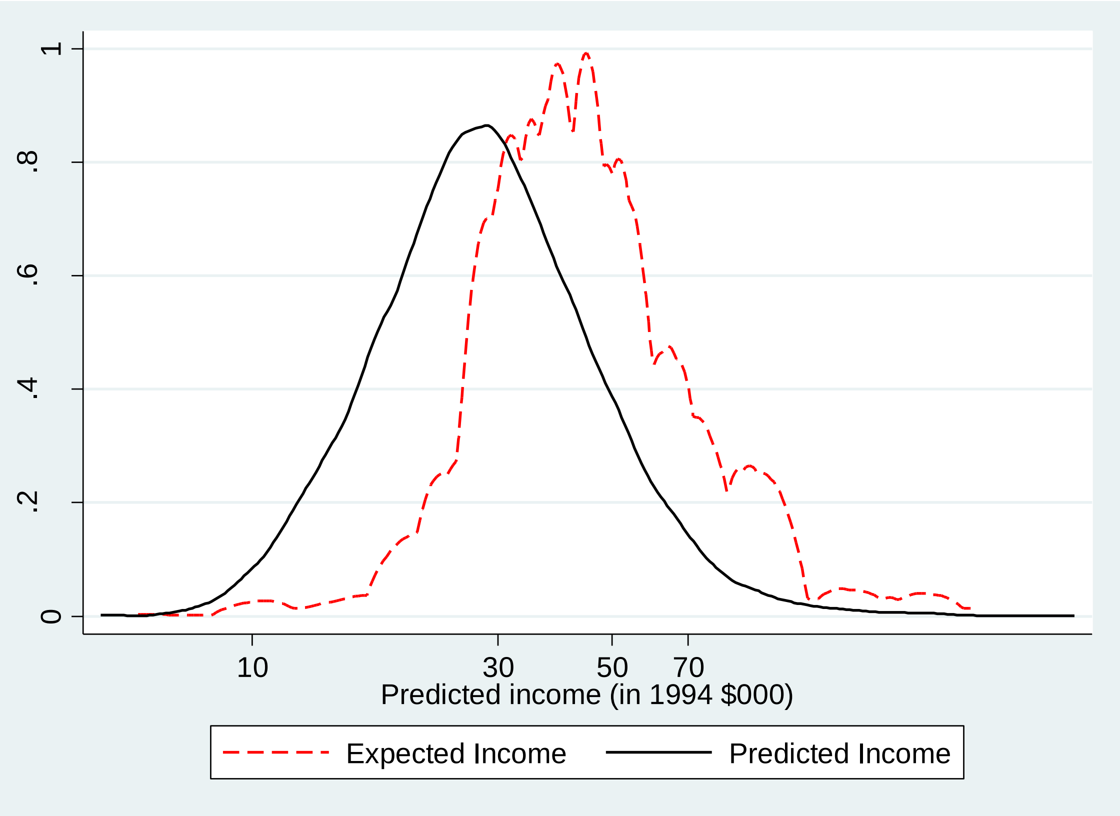
(a) Workers may make more accurate predictions of future income than their student peers

and

(b) That their expectations may, on average, be realistic.

Figure 3.9 shows little support for either of these hypotheses.

Figure 3.9 Distribution of log expected and log predicted income at age 30 for workers



Note:
1 See notes to Figure 3.7

2 Source: Diagrams produced by the author using the NELS dataset. Sample size = 1,587

Those who were working at age 20 expected an age 30 mean income of $40,000. In reality, I predict their mean income to be $24,789. Workers are overestimating their wage, on average, by around 60%. This is similar to the overestimation made by students, where I found a figure of 50%. These results are supported by my investigation of workers’ occupational expectations in Table 3.10.

Table 3.10 Proportion of 20 year old *workers* expecting to enter each profession by age 30, and the proportion who have actually entered that profession by age 26

	% expecting to be in the occupation at age 30	% actually in each occupation at age 26	Difference between expected and actual (% points)	Average CPS wage for each occupation in 2004 (deflated to 1994 $000)
Homemaker, not working, studying	0.6	3.9	-3.3	0.0
Farmer	2.3	0.9	1.3	12.5
Labourer	4.0	14.6	-10.6	16.7
Service	1.9	3.2	-1.3	17.0
Skilled operative	6.2	13.6	-7.4	19.6
Clerical	2.1	4.7	-2.6	19.9
Craftsman	16.6	19.4	-2.8	20.6
Sales	2.1	2.8	-0.6	22.7
Protective services	8.8	4.9	3.9	24.9
Arts, Entertainment, Writing	5.5	0.6	4.9	29.3
Teacher	3.2	0.6	2.5	32.5
Professional Medicine (not Doctor)	3.3	1.0	2.3	35.0
Other Professional	7.6	12.4	-4.7	35.6
Engineer	4.7	0.9	3.8	38.7
Computer technical	3.2	1.6	1.6	39.2
Manager	11.0	11.8	-0.8	41.3
Legal	1.6	1.2	0.4	53.0
Doctor	1.5	0.5	1.0	63.7
Military	1.4	1.2	0.2	NA
Proprietor	12.5	3.1	9.4	NA
Observations	**1,459**	**1,595**		

Notes:
1 See the notes to Table 3.8
2 Source: Authors calculations from the NELS dataset and the CPS data (referred to Table 3.8).

Whereas 4.7% of workers expected to become an engineer, less than 1% were working in this occupation at age 26. On the other hand, around 1 in 20 thought they would be working as a labourer by age 30. Yet around an eighth held this job at age 26. Calculating weighted average wages from this occupational data (see page 113), I find workers expect a wage of $29,263 but end up receiving $25,907; a difference of 15%. Once again, this may appear to be small when compared to the 60% overestimation in wages. But I remind the reader that this method captures just one aspect of the underlying issue (as discussed previously on page 113).

Overall, there is little evidence that workers hold realistic expectations. In fact, on average they are just as unrealistic as their student peers. Both groups tend to overestimate their future income and occupation; many believe they will receive the financial rewards on offer in professional careers, but will ultimately not be able to obtain this goal.

Comparing workers to different groups of students

The above analysis has treated those in higher education at age 20 as a homogenous group. In reality, students differ in all manner of characteristics, including the subject they study, prior academic achievement, whether they also hold a job, and, looking into the future, whether they eventually graduate. I extend the above analysis by trying to answer three questions. Firstly, though workers may not hold more realistic expectations than the "average student", they may make better predictions than particular groups. Do workers make better predictions than say Art and Humanities students, for instance, whom I found to be the least realistic over short time horizons in the UK? Secondly, are factors such as race, ability and social class influencing both enrolment in higher education and the accuracy of expectations? If so, is it these factors that are driving my results? Finally, can I provide any further evidence that experience in the labour market is unrelated to accuracy of expectations, as my findings so far suggest, by considering differences between students with and without a job?

I investigate these questions by estimating an Ordinary Least Squares regression model[89]. I specify the dependent variable as the natural logarithm of the expected income divided by the predicted age 30 income:

$$\log\left[\frac{Y_{ij}^{exp}}{Y_{ij}^{act}}\right]$$

[89] One might suggest that this looks like an individual level analysis that I ruled out in the previous section as inappropriate. On the other hand, another way of looking at this is that I am analysing conditional means (and thus that I am infact undertaking further analysis at the group level).

Webbink and Hartog (2004) use a similar specification in their analysis of Dutch students' wage expectations. This specification is assumed to satisfy the condition that errors are normally distributed with constant variance[90]. It also allows a distinction between respondents who over and under estimate their future income, unlike the specification preferred by Betts (1996) and Wolter (2000). Later in this section, I also present quantile regression estimates as an alternative to test the robustness of my results[91].

In the first specification, I compare working 20 year olds to students defined by the subject that they study[92]. In other words, workers enter the regression model as the reference group, with 14 dummy variables representing students in different disciplines. I then add in a term reflecting the respondents' cognitive ability in mathematics at age 18. Intelligent individuals are more likely to enter higher education. But they may also be particularly adept at processing the labour market information that they receive. Once I have controlled for students' superior ability, do I find that workers (who may hold a greater quantity of information) hold more realistic expectations?

Specification 3 controls for a series of other potentially confounding factors, including race and family background. Work from the sociological literature, for example Baird et al (2008), describes the importance of controlling for these characteristics when considering the accuracy of students' expectations. I also include an indicator for whether the respondent was a student who also held a part-time job while at university. Previously, I argued that workers will have more accurate expectations than students as they hold more labour market information. In a similar manner, one would expect students in paid employment to hold more labour market information than their university peers without a job. I use this analogy in the final specification as an additional test of whether labour market experience is related to accuracy of

[90] Analysis of the OLS residuals was carried out thoroughly after estimation of each regression model. There was little evidence that the normality and constant variance assumptions were violated.
[91] I choose to present the OLS results as I can easily take into account the complex survey design used in NELS. This is much trickier when using alternatives such as quantile regression.
[92] Under both prediction methods, students in all subjects are assumed to have the same average annual real wage growth rate between 26 and 30. I also experimented with a prediction model that allowed wage growth to vary between graduates from different subjects. Results were largely the same to those presented.

income expectations.

However, when making comparisons between students and workers, one should remember that some of those who enter higher education leave without obtaining a degree. In other words, even though these students were enrolled in a college programme at age 20, they may not have obtained a university level qualification by age 26. The OECD 1998 Education at a Glance report (OECD 1998) notes that the USA has a relatively low university completion rate (just over 60% of those who enter). This is reflected in the NELS data; around 30% of those who were students at age 20 had not obtained a degree by age 26[93]. I take this into account in my final specification by including a dummy variable that indicates whether the individual became a "college drop-out" (i.e. enrolled in university at age 20, but had not obtained a degree by age 26). Interpretation of the subject dummies will therefore change between the first three specifications and the fourth. In particular, the final specification will indicate the accuracy of students' expectations compared to workers, *conditional* on successful degree completion by age 26.

Formally, the final specification of the model is:

[93] Morgan (2005) finds a similar proportion when he uses a different sample selection of the NELS data.

$$\log\left[\frac{Y_{ij}^{exp}}{Y_{ij}^{act}}\right] = \alpha + \beta_0 R_i + \beta_1 F_i + \beta_2 S_i + \beta_3 T_i + \beta_4 D_i + \beta_5 W_i + \varepsilon_{ij}$$

With:

Y_{ij}^{exp} = Expected income at age 30

Y_{ij}^{act} = Predicted income at age 30, using Method 2

R = Race

F = Parental income when respondent was 18 years old

S = Subject of study, working or unemployed at age 20

T = Measure of individual ability at age 18

D = Whether the respondent was a university student at age 20, but had not obtained a degree by age 26

W = An indicator of whether the individual was a student who also held a job at age 20

ξ_{ij} = Error term. Individuals were initially sampled by school clusters at age 14, which is accounted for by adjusting the standard errors.

i = Individual i

j = School j, that the individual was initially sampled from at age 14. All standard errors have been adjusted to take into account the complex sampling design (clustering of children within schools)[94].

Results are presented in Table 3.11.

[94] I also experimented with a fixed effects regression model, including a dummy variable for each school that children were initially sampled from. Results were largely unchanged from those presented.

Table 3.11 Ordinary least squares regression results comparing the accuracy of students' income expectations to workers

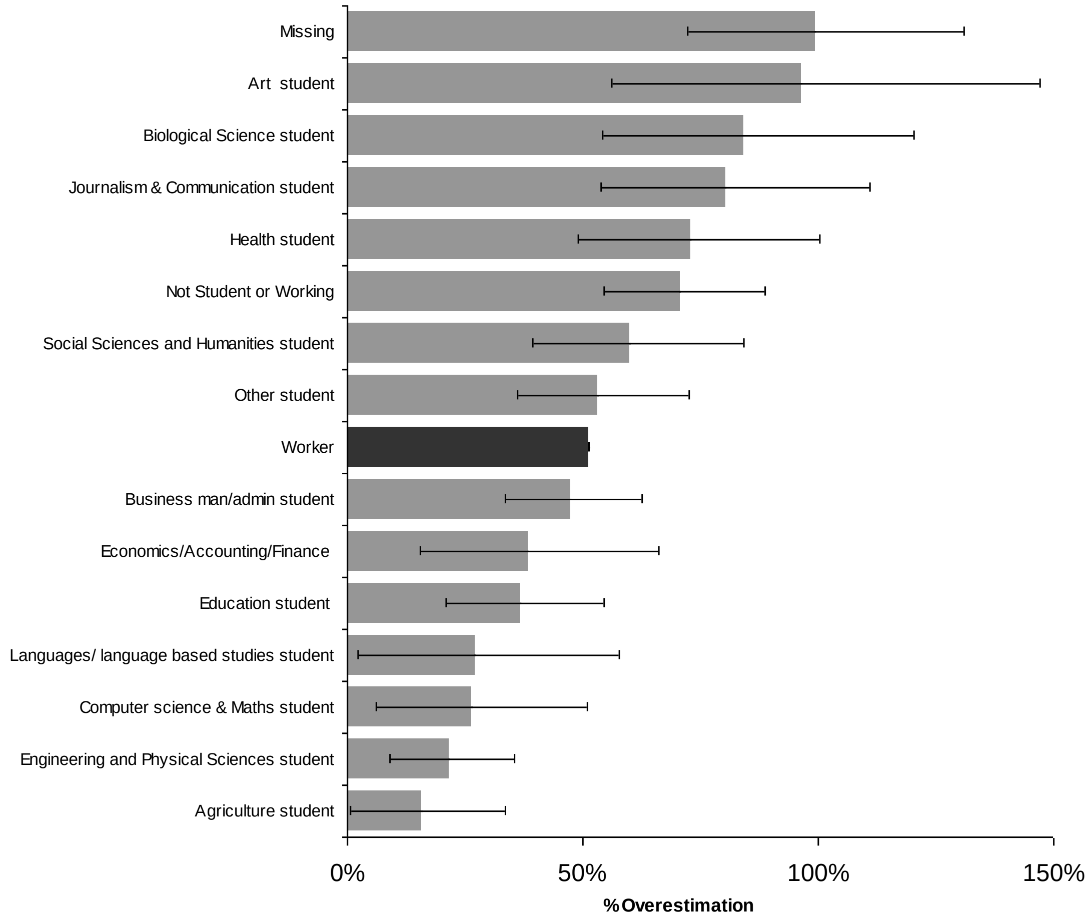
	Specification 1		Specification 2		Specification 3		Specification 4	
	Co	**SE**	**Co**	**SE**	**Co**	**SE**	**Co**	**SE**
Work-student status at age 20 (Ref: Working)								
Agriculture student	-0.27*	0.07	-0.20*	0.07	-0.21*	0.07	-0.29*	0.07
Economics, Finance student	-0.09	0.09	-0.02	0.09	-0.03	0.09	-0.14	0.09
Business, Management student	-0.03	0.05	0.02	0.05	0.00	0.06	-0.11*	0.06
Journalism, Communication student	0.18*	0.08	0.22*	0.08	0.21*	0.08	0.08	0.08
Computer Science, Maths student	-0.18*	0.09	-0.11	0.09	-0.15	0.10	-0.29*	0.09
Education student	-0.10	0.06	-0.05	0.06	-0.06	0.06	-0.15*	0.07
Engineering, Physical sciences student	-0.22*	0.05	-0.16*	0.06	-0.18*	0.06	-0.29*	0.06
Language student	-0.17	0.11	-0.10	0.11	-0.09	0.11	-0.22	0.12
Health student	0.13*	0.07	0.19*	0.07	0.17*	0.08	0.06	0.08
Law student	0.45*	0.19	0.45*	0.18	0.44*	0.18	0.26	0.17
Biological science student	0.20*	0.09	0.27*	0.09	0.27*	0.09	0.16	0.09
Social sciences, Humanities student	0.06	0.07	0.12	0.07	0.09	0.07	-0.03	0.07
Art student	0.26*	0.11	0.31*	0.11	0.31*	0.12	0.15	0.11
Other student	0.01	0.06	0.05	0.06	0.03	0.06	-0.10	0.06
Not student or working	0.12*	0.05	0.13*	0.05	0.11*	0.05	0.10*	0.05
Missing	0.28*	0.07	0.30*	0.07	0.26*	0.07	0.27*	0.07
Maths ability at age 18	-	-	-0.08*	0.02	-0.07*	0.02	-0.05*	0.02
Race (Ref: White)								
American Indian or Alaska Native	-	-	-	-	0.08	0.09	0.06	0.09
Asian or Pacific Islander	-	-	-	-	0.03	0.07	0.04	0.06
Black, not Hispanic	-	-	-	-	0.20*	0.05	0.18*	0.05
Hispanic or Latino	-	-	-	-	0.13*	0.05	0.11*	0.05
More than one race	-	-	-	-	0.09	0.09	0.07	0.09
Missing	-	-	-	-	0.07	0.08	0.03	0.07
Family income parents reported when respondent was age 18 (Ref: Bottom quintile)								
2nd quintile	-	-	-	-	-0.01	0.05	-0.01	0.05
3rd quintile	-	-	-	-	0.00	0.05	0.01	0.05
4th quintile	-	-	-	-	0.04	0.05	0.04	0.04
Top quintile	-	-	-	-	-0.01	0.06	0.02	0.06
Student at 20, who also held a part-time job (Ref: No)								
Yes	-	-	-	-	0.02	0.04	-0.03	0.04
College dropout (ref: No)								
Yes	-	-	-	-	-	-	0.31*	0.05
Constant	0.41*	0.03	0.38*	0.03	0.34*	0.05	0.36*	0.05

Notes:
1 The response variable is the natural logarithm of the ratio of expected to actual income
2 See notes to Table 3.1 for details on the "maths ability" variable
3 * indicates statistical significance at the 5% level
4 "Missing" dummy variables are included when the respondent has not provided information on any of the covariates. I do not present results for brevity.
5 Source: Authors calculations from the NELS dataset. Sample size = 4,434. Dependent variable is the natural logarithm of respondents expected age 30 income divided by their predicted actual age 30 income.

Model 1 enters just the indictor of whether the respondent was working or a student, defined by the subject they were studying, at age 20. The results show workers sit somewhere in the middle of this ranking; they make better predictions than some students, but worse than others. Engineering, Physical Sciences, Maths, Education and Agriculture students are all more realistic on average than workers at the 5% level. Similarly, Art, Law, Journalism and Biological Science students are all less realistic than workers at the 5% level. Figure 3.10 shows this in more detail, highlighting by how much each group overestimates their age 30 income (on average)[95].

[95] To calculate how much a person with given characteristics overestimates by, one must sum the relevant coefficients from Table 3.11, and then take the exponent of this value. For example, take a journalism student. I want to know how much they overestimate their future income by using specification 1. Firstly, I sum the relevant coefficients (0.41+ 0.18) to get 0.59. I then take the exponential of this value exp(0.59), to get a value of 1.80. If I then subtract 1 from this value (1 is where expectations equals realisation) and multiply by 100 (to get the value into percentage form), I arrive at the average % overestimation (80% shown in Figure 3.10).

Figure 3.10 Difference between expectations and realisations: Workers compared to students in different subjects



Notes:

1 Thick bars refer to average overestimation of age 30 income for each group. These figures have been calculated from model specification 1, which just contains the subject dummy coefficients and no other explanatory variables. The thin black line running through the centre of each bar is the 95% confidence interval of this estimate.
2 Source: Authors calculations from the NELS dataset. Sample size = 4,434.

It appears that young adults who are studying vocational, financial and mathematically based courses hold reasonably realistic expectations. For instance, those studying Computer Science and Maths overestimate their age 30 income by a (comparatively) small 20%. Students in Agriculture are even more realistic, their expectations are only just statistically different to their predicted realisations. Yet those studying artistic and writing based courses, with the exception of Language students, expect almost double what I predict them to earn.

This ranking of subjects is very similar to my results for UK students' starting salary expectations, presented in chapter 2. Thus across countries, over a long and short time horizon, it appears that vocational and mathematical students are the most realistic, while those in more creative subjects are the least. It is interesting to consider this result in light of studies that have investigated actual wage differentials by college major. Specifically, there are some quite striking consistencies with Black et al (2003). The authors of this study note that perspective students are provided with 'little concrete information' about the labour market success of graduates from the array of different US subject majors. Hence they illustrate the wages of graduates in around 40 disciplines, relative to the earnings of those who have left university with an Economics degree. They find that Engineers receive the highest wages, just as I find them to be the most realistic. Likewise, I find Art students are amongst the least realistic, while Black et al show that this group earn the lowest wages. Some other patterns seem to hold too; for instance when drawing comparisons between Biologists (lower earners and less realistic) and Physicists, Mathematicians and Chemists (higher earners, more realistic).

Some care does need to be taken, however, in drawing the above interpretation from my results. An alternative explanation for why I find students in certain subject to make better predictions is that most young people enrolled in higher education do not have any idea what they will be earning at age 30. Respondents may, consequently, be reporting only their quite vague view of what the average graduate salary is at this age, with little variation in this figure by the subject they study. When expectations are then compared to realisations, it seems that students in certain subjects (e.g. Maths, Engineering etc) make more accurate predictions than others. But the only reason for this is that they happen to be in a discipline that leads to high earning jobs (i.e. it is not so much to do with them having "more realistic" expectations, but rather that they just happen to be enrolled in a subject that leads to high later earnings). Hence the greater "accuracy" in their expectations is reflecting the fact that all students anchor their (quite vague) expectations around the same (rather hazy) point – but that some groups tend to earn more than others. Although descriptive statistics do suggest that there is at least some variation in income expectations between the different subject groups (i.e. they do not all anchor on exactly the same point) I can not rule this explanation out for why those in certain disciplines make better predictions than those in others[96]. The reader can see the extent of differential age 30 income expectations by discipline if they turn to Appendix Table A3.14.

In the second model I control for respondents' cognitive ability in maths on a test taken at age 18. Earlier, I suggested that those of higher ability maybe more realistic. This seems to be consistent with the data. A worker of average ability overestimates his future wage by 46%. However if their maths test score was two standard deviations above the mean, I predict they overestimate their future income by only 25%. Notice that the subject dummy coefficients have slightly decreased, for instance from -0.22 to -0.16 for Engineering students and -0.09 to -0.02 for Economics. Hence, although it seems that cognitive maths ability is related to accuracy of young adults' income expectations, there is little evidence to suggest this is why I find that some groups of students to make better predictions than workers.

In model 3, I add additional controls for Race and Family background. As stated by Baird et al (2008), race influences the accuracy of expectations. Blacks, Hispanics and Latinos all make worse predictions than Whites. On the other hand family background, measured by parental income quintile, is statistically insignificant. Notice, however, that none of the substantial results from model 2 change. Differences in family backgrounds and ethnicity do not explain why I find no difference between students and workers. I also include a variable that indicates whether the respondent was a student who also held some sort of formal employment at age 20[97]. If work experience provides young adults with valuable labour market information, which they process rationally, one would expect this group to make better predictions than their student peers who do not have a job. Again, this does not seem to be the case. The coefficient is small and statistically insignificant, with this result holding across several specifications and unconditional estimates not presented. This supports my finding that labour market experience is unrelated to the accuracy of young adults' long term income expectations. Not only do young adults working full-time make no better predictions than students, but those enrolled in higher education

[96] For instance, Appendix Table A3.14 suggests that Education students' expect their income at age 30 to be on average $39,749 (mean), compared to $62,446 for those studying finance or accounting. The equivalent figures for the mean are $35,500 and $52,500. This variation may tend to suggest that the explanation given above is not the sole reason why I find differences across subjects in this chapter.
[97] Around half of the students surveyed in the NELS fell into this category.

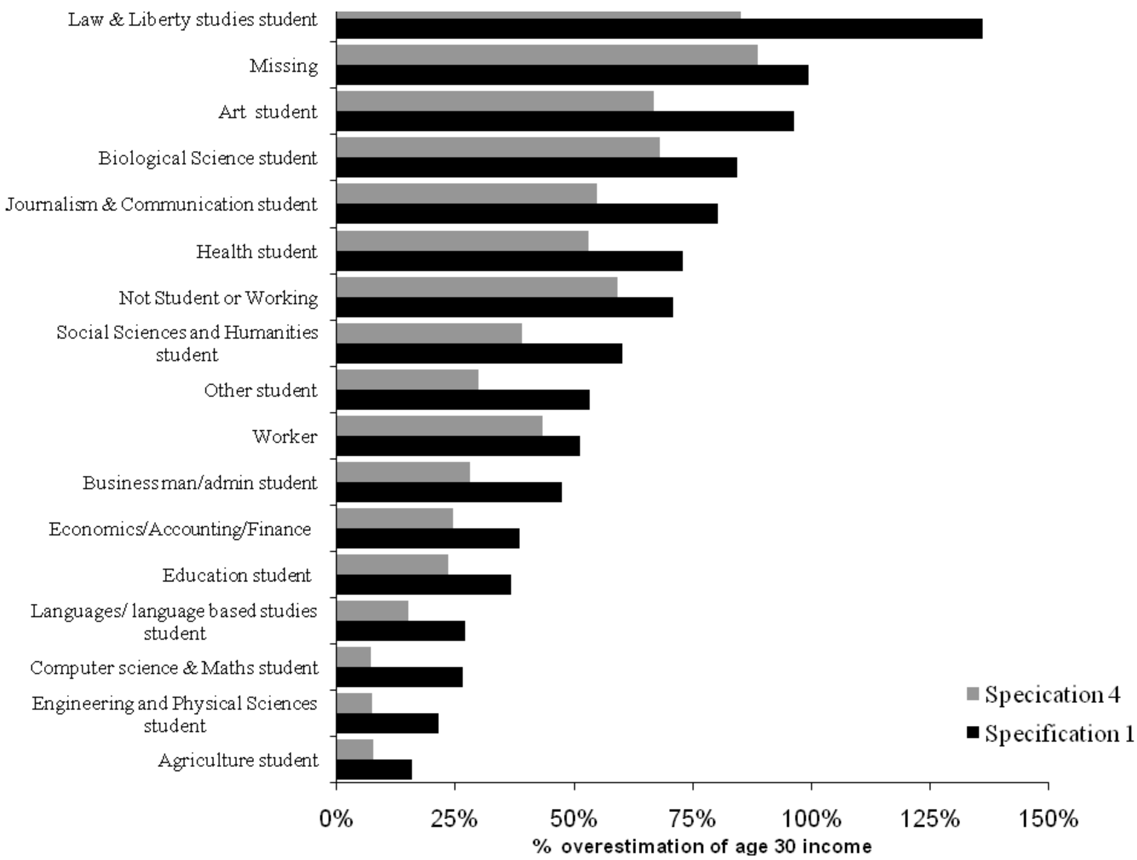
seem to receive little information about their long-term prospects in the labour market from holding a part-time job.

The final specification enters a dummy variable for individuals who were students at age 20, but had not obtained a degree by age 26[98]. In other words, this group dropped out of university *after* reporting their income expectations. Firstly notice the large (0.31) and highly significant ($t \approx 6$) coefficient. These students make particularly poor predictions of their future income, and are in fact significantly less realistic than their peers who entered the labour market straight from school[99]. For instance, workers overestimate their age 30 income by around 45%[100], compared to 130% for a 20 year old Art or Journalism student who failed to complete their degree. This result can be interpreted in two ways. One possibility is that these students stated their income expectation on the assumption they would obtain a certain level of human capital and a valuable labour market signal. However, they did not go on to actually receive the outcomes they initially anticipated from their human capital investments, thus causing their apriori expectations to be incorrect. Alternatively, these individuals could have dropped out of university *because* of their overly ambitious expectations. For instance, they may have gone to university thinking they would earn a high wage (i.e. their high expectation observed at age 20). But through their later experiences, they may have revised down their expectations substantially (i.e. If I were to observe expectations at age 21 say, they would be much lower). On the basis of this revision, they have decided that the benefits from obtaining a degree are not worth their continued investment, and hence leave university before the end of their course. Hence this variable is potentially endogenous; it could be these students unrealistically high expectations that is driving their decision to leave university, rather than their expectations being unmet because they drop out.

[98] This group makes up 17% of the 4,434 NELS sample members analysed in this paper.
[99] For the prediction of age 30 income in section 3.4, I have treated this group the same as those with a high school qualification who never been to university at all. One may argue that college drop outs may have some wage premium over this group. All the results from specification 2 onwards still hold even if I make different assumptions (e.g. that this group have the same wage growth rate as those with an associate degree).
[100] Note this is lower than before. Recall previously I focused on the median. Now I am using OLS regression, the measure of central tendency used is the mean.

Figure 3.11 illustrates how including this variable leads to a large change in the other parameter estimates. There is a particularly large effect on the subject dummy coefficients. Recall that these now compare the accuracy of students' expectations to workers, conditional on whether they complete university by age 26.

Figure 3.11 Difference between expectations and realisations: Specifications 3 and 4



Notes:
1 Black bars refer to model specification 3, where I control for ethnicity, family income and ability in mathematics. The light grey bars refer to model specification 4, which also includes an indicator of whether the respondent had graduated from university by age 26.
2 Source: Authors calculations from the NELS dataset. Sample size = 4,434.

The light grey bars in Figure 3.11 are the estimated overestimation of age 30 income for each group, calculated using specification 4. These results add further weight to my finding that students are no less realistic than workers. Consider, for instance, Business students. In specification 1-3, they were statistically indistinguishable from workers, overestimating their age 30 income by around 45%. Now I find they are substantially *more* realistic, conditional on them having graduated by age 26, overestimating their age 30 income by a comparatively small 30%. On the other hand, those who dropped out of Business school overestimate their future income by around 75%. Similarly, the once statistically significant difference between Art, Law, Health and Journalism students and workers has now disappeared. There is even less evidence that workers are more realistic than students, and in fact quite the opposite may even be true.

It is also interesting to note that, for some groups of students, the difference between expectations and later realisations is now statistically insignificant. This includes Agriculture, Computer Science, Maths, Engineering and Physical Science graduates, who overestimate their age 30 income by, on average, less than 10%. Hence, conditional on successful completion of an undergraduate degree, I find that some groups of students are actually quite realistic, even over a relatively long time horizon.

In Table 3.12 I present various robustness tests that refer to the third specification of the regression model described above[101]. Specifically, model A refers to when I do not extrapolate the actual wage data, and simply compare age 26 income to expectations at age 30. Model B presents results when using "method 1" to predict age 30 wages, as described in section 3.4. Model C adjusts for the item non-response described in Tables 3.1 and 3.4, via the application of response weights I have created from a logistic analysis of missing data, while model D refers to quantile (median) regression estimates.

[101] I have chosen this specification as it controls for the largest number of pre-determined factors, without including the potentially endogenous "college dropout" variable. In analysis not presented, I ran each of these robustness tests for all specifications and found largely consistent results.

The results generally support those presented in Table 3.11. Notice that Blacks and Hispanics always make worse estimates than Whites, while family income is never statistically significant. Likewise, higher ability is always associated with more realistic expectations. The estimated subject dummy coefficients are also similar to before; Art, Biological Sciences and Journalism students make relatively poor predictions compared to both workers and their university peers who are studying Agriculture, Engineering or Physical Sciences. Likewise, in analysis not presented, I found the “college dropout” variable to be positive, strong and highly significant, while its inclusion again caused a relatively steep decline in the other model coefficients.

To further test the robustness of my results, I estimate a binary logistic regression model of whether the respondent, at age 26, was working in the occupation he expected to be in at age 20. As argued previously, those who are the most realistic about their future occupation should also be the most realistic about their future income. Consequently, one would expect to see results from analysis of occupational data to be consistent with the above results regarding income. For instance, I should find that maths students make better predictions of their future occupation than those from creative subjects. Likewise, individuals who drop out of university should only rarely enter the occupation they expected at age 20.

Table 3.12 Robustness tests of accuracy of income expectations, using regression specification 3

	Test A		Test B		Test C		Test D	
	Co	**S.E**	**Co**	**S.E**	**Co**	**S.E**	**Co**	**S.E**
Work-student status at age 20 (Ref: Working)								
Agriculture student	-0.07	0.07	-0.11	0.07	-0.23*	0.07	-0.27	-
Economics, Finance student	-0.04	0.09	-0.08	0.11	-0.02	0.10	-0.03	-
Business, Management student	0.00	0.06	-0.01	0.06	-0.01	0.06	-0.02	-
Journalism, Communication student	0.24*	0.09	0.19*	0.08	0.20*	0.08	0.39	-
Computer Science, Maths student	-0.01	0.10	-0.13	0.1	-0.13	0.11	-0.10	-
Education student	0.10	0.07	-0.03	0.07	-0.06	0.07	-0.03	-
Engineering, Physical sciences student	-0.03	0.06	-0.17*	0.06	-0.19*	0.06	-0.14	-
Language student	0.02	0.12	-0.04	0.13	-0.08	0.11	-0.10	-
Health student	0.30*	0.07	0.19*	0.08	0.18*	0.07	0.17	-
Law student	0.51*	0.18	0.42*	0.20	0.48*	0.24	0.40	-
Biological science student	0.36*	0.08	0.25*	0.09	0.25*	0.09	0.43	-
Social sciences, Humanities student	0.23*	0.08	0.14*	0.07	0.11	0.07	0.17	-
Art student	0.39*	0.13	0.32*	0.14	0.29*	0.12	0.43	-
Other student	0.13*	0.06	0.03	0.06	0.02	0.06	0.13	-
Not student or working	0.16*	0.05	0.12*	0.05	0.13*	0.05	0.10	-
Missing	0.33*	0.07	0.27*	0.07	0.24*	0.07	0.42	-
Maths ability at age 18	-0.04*	0.02	-0.07*	0.02	-0.06*	0.02	-0.10	
Race (Ref: White)								
American Indian or Alaska Native	0.12	0.12	0.08	0.08	0.04	0.1	0.11	-
Asian or Pacific Islander	0.06	0.09	0.05	0.08	0.08	0.12	0.05	-
Black, not Hispanic	0.22*	0.06	0.22*	0.05	0.20*	0.05	0.16	-
Hispanic or Latino	0.13*	0.05	0.12*	0.05	0.14*	0.05	0.12	-
More than one race	0.09	0.07	0.16*	0.08	0.06	0.09	-0.06	-
Missing	0.16*	0.08	0.04	0.08	0.04	0.2	0.12	-
Family income parents reported when respondent was age 18 (Ref: Bottom quintile)								
2nd quintile	-0.09*	0.06	-0.03	0.05	-0.03	0.05	-0.01	-
3rd quintile	-0.02	0.05	-0.02	0.06	-0.02	0.05	-0.08	-
4th quintile	0.03	0.06	0.07	0.05	0.03	0.05	0.01	-
Top quintile	0.01	0.05	0.01	0.06	-0.03	0.06	0.03	-
Student at 20, who also held a part-time job (Ref: No)								
Yes	0.00	0.04	0.02	0.04	0.02	0.04	-0.04	-
Constant	0.51*	0.05	0.36*	0.05	0.37*	0.05	0.34	

Notes:
1 Test A refers to when I do not extrapolate the data, and simply compare age 26 income to expectations at age 30.
2 Test B refers to when I extrapolate the income data using prediction “Method 1” (see section 3.4).
3 In Test C, I re-weight the data to take into account the item non-response shown in Table 3.1 and 3.4. Robustness Test D presents the quantile (median) regression estimates. Note that standard errors have not been presented, due to the difficulties of providing accurate figures when using complex survey data (clustering and weighting) as in the NELS.
4 * Indicates statistical significance at the 5% level
5 Source: Authors calculations from the NELS dataset. Sample size = 4,434. Dependent variable is the natural logarithm of respondents expected age 30 income divided by their predicted actual age 30 income.

Specifically, I estimate a logistic regression using the binary indicator O as the response. This variable is assigned the value 1 if the respondent, at age 26, was working in the occupation he expected to be in at age 20. I enter the same covariates as in the model described on page 122 and Table 3.11. Formally, this model is specified:

$$\log\left(\frac{\Pi(O_i)}{1-\Pi(O_i)}\right) = \alpha + \beta_0 \mathrm{R_i} + \beta_1 \mathrm{F_i} + \beta_2 \mathrm{S_i} + \beta_3 \mathrm{A_i} + \beta_4 \mathrm{D_i} + \beta_5 \mathrm{W_i} + \varepsilon_{\mathrm{ij}}$$

Where:

$\Pi(O_i)$ = Probability of respondent i entering the occupation they expected at age 20 (by age 26)

R = Race

F = Parental income when respondent was 18 years old

S = Subject of study, working or unemployed at age 20

A = Measure of individual ability at age 18

D = Whether the respondent was a university student at age 20, but had not obtained a degree by age 26

W = An indicator of whether the individual was a student who also held a job at age 20

ξ_{ij} = Error term. Individuals were initially sampled by school clusters at age 14, which is accounted for by adjusting the standard errors.

i = Individual i

j = School j, that the individual was initially sampled from at age 14

Results are presented as odds ratios in Table 3.13. A coefficient greater than 1 indicates a higher probability of the respondents' occupational expectations being correct.

Table 3.13 Logistic regression results comparing how realistic students' occupational expectations are to workers

	Specification 1		Specification 2		Specification 3		Specification 4	
	OR	SE	OR	SE	OR	SE	OR	SE
Work-student status at age 20 (Ref: Working)								
Agriculture student	0.61	0.34	0.52	0.28	0.59	0.35	0.74	0.45
Economics, Finance student	1.26	0.37	1.03	0.32	1.35	0.43	1.90*	0.60
Business, Management student	0.80	0.22	0.71	0.20	0.92	0.27	1.25	0.38
Journalism, Communication student	1.02	0.34	0.91	0.30	1.17	0.40	1.71	0.61
Computer Science, Maths student	1.53	0.63	1.27	0.50	1.72	0.67	2.82*	1.25
Education student	2.18*	0.57	1.92*	0.51	2.52*	0.73	3.41*	1.04
Engineering, Physical sciences student	1.86	0.44	1.59*	0.39	2.07*	0.55	2.83*	0.77
Language student	0.78	0.38	0.65	0.33	0.82	0.41	1.18	0.65
Health student	0.47┼	0.22	0.39*	0.19	0.56	0.27	0.78	0.38
Law student	1.01	0.41	1.01	0.40	1.40	0.55	2.56*	1.05
Biological science student	0.74	0.36	0.61	0.31	0.79	0.42	1.07	0.56
Social sciences, Humanities student	0.67	0.22	0.56┼	0.19	0.77	0.27	1.07	0.37
Art student	0.98	0.38	0.86	0.35	1.07	0.49	1.75	0.71
Other student	1.33	0.32	1.23	0.31	1.66*	0.43	2.51*	0.68
Not student or working	0.75	0.20	0.74	0.19	0.77	0.20	0.78	0.21
Missing	1.31	0.38	1.23	0.35	1.85	0.56	1.78*	0.55
Maths ability at age 18	-	-	1.20*	0.06	1.17*	0.06	1.11	0.07
Race (Ref: White)								
American Indian or Alaska Native	-	-	-	-	0.63	0.33	0.69	0.39
Asian or Pacific Islander	-	-	-	-	0.76	0.24	0.72	0.23
Black, not Hispanic	-	-	-	-	0.61	0.21	0.64	0.22
Hispanic or Latino	-	-	-	-	0.90	0.21	0.97	0.23
More than one race	-	-	-	-	0.85	0.29	0.89	0.32
Missing	-	-	-	-	0.60	0.24	0.70	0.27
Family income parents reported when respondent was age 18 (Ref: Bottom quintile)								
2nd quintile	-	-	-	-	0.90	0.21	0.93	0.21
3rd quintile	-	-	-	-	0.94	0.22	0.92	0.21
4th quintile	-	-	-	-	1.24	0.30	1.21	0.29
Top quintile	-	-	-	-	0.96	0.26	0.87	0.24
Student at 20, who also held a part-time job (Ref: No)								
Yes	-	-	-	-	0.58*	0.10	0.65*	0.11
College dropout (ref: No)								
Yes	-	-	-	-	-	-	0.30*	0.06

Notes:
1 * Indicates statistical significance at the 5 % level, ┼ Indicates statistical significance at the 10% level. 2 A higher odds ratio indicates more realistic occupational expectations.

2 Source: Authors calculations from the NELS dataset. Sample size = 4,218. Dependent variable is the whether the respondent was working at age 26 in the job they expected at age 20 (taking a value of 1 if they were and 0 otherwise).

There is reasonable agreement between these results and those presented in Table 3.11. For instance, notice that mathematical and vocational students tend to make better predictions of their future occupation than workers, just as Table 3.11 showed they made better predictions of their income. However, it is also worth pointing out that, in some specifications, the coefficients only begin to approach traditional levels of statistically significance (though, qualitatively, they have the same sign). It is also interesting to note how the inclusion of the "college drop out" variable in specification 2 has the same effect in Tables 3.11, 3.12 and 3.13 (the subject coefficients all tend to increase, while the expectations of university drop-outs are particularly unlikely to become true).

Qualitatively, other results from Table 3.11 also hold. Lower ability respondents are less likely to enter the occupation they expected at age 20, though this does not quite reach statistical significance, even at the 10% level. Likewise, Black respondents are less likely to enter their chosen occupation than Whites, but again this is not statistically significant. One interesting difference is that the coefficient on students holding a part-time job is now statistically significant at the 5% level. However, it is the opposite sign to what one would expect; those with a part-time job are *less* likely to enter the job they expected.

Nevertheless, the directions of the parameter estimates are generally consistent with the results presented in Table 3.11. It seems that (qualitatively) these results support my substantial conclusions from the income expectations data. This gives me further confidence that my results are not being driven by my prediction methods for age 30 income or assumptions I make about the income expectation data.

3.6. Discussion and conclusion

The small US and European literature on young adults' expectations has typically focused on how well university students can predict their first salary upon graduation. Though they sometimes deal with longer time horizons, results are normally shown to hold for only a very specific proportion of the US student population. Furthermore, existing work rarely compares the accuracy of different groups. There is little or no comparison of students versus workers, those studying for a mathematical degree versus a more creative subject, or those who successfully graduate versus those who do not. I try to resolve these issues by using rich, longitudinal data that has been drawn from across the US population of high school seniors. Hence I not only make a better attempt at representing the expectations of the US student population, but also tackle a set of new and interesting hypotheses that have not been previously considered in the literature.

However, one should not ignore the difficulties I have encountered with the NELS data. Missing data, particularly the fact that age 30 income is not directly observed, is a notable problem. A second issue is whether the questions asked are accurately capturing young adults' expectations (what they realistically believe will happen) rather than their aspirations (their hopes and dreams). The ordering and wording of the questions (given in section 3.3) should have guided respondents towards making a realistic assessment of their future income. Yet certain groups may have interpreted this question quite literally (e.g. Maths students report their expectations) while others have not (e.g. Art and Journalism students state their aspirations). This issue is not specific to this chapter, but rather the more general practice of collecting expectation data in economic research. As such, this seems an area that is ripe for future work.

Noting these caveats, my results suggest that, on average, students at age 20 have unrealistic expectations of their income at age 30. Yet this broad result needs qualification. Certain groups of students, conditional on successful completion of their degree, are actually quite realistic. For instance, Maths, Education and Engineering students overestimate their age 30 income by less than 10%. One may wish to view these results in light of Black et al (2003), who note that US students are provided with 'little concrete information' about the success of graduates in different subjects. Their paper tries to provide this information, illustrating the wages of graduates in around 40 disciplines. They find that Engineers receive the highest wages, just as I find them to be the most realistic. Likewise, Art students are amongst the lowest earners and least realistic. Even some quite specific patterns seem to hold. Biologists earn less than Physicists, Mathematicians and Chemists, though they do not seem to realise this when undertaking their human capital investments. Consequently, my results suggest there is certainly a need to provide prospective students with the type of information presented by Black and his co-authors (noting the caveat I placed on my interpretation of this result – that it could be reflecting students' not having a clue what their wage will be at age 30 and hence those in all subject groups anchoring their expectation at some rough average point).

On the other hand, there is substantial evidence that those young adults who are working at age 20 make quite poor predictions of their future income; overestimation is, on average, 50%. It is also interesting to consider again why workers seem to make no better predictions than students. As stated at the start of this chapter, it maybe that young workers are not focused on a particular career and hence suffer from a lack of direction in the labour market. Alternatively, it might be that young workers are myopic and choose to collect information from those who are closer to them in terms of age and the next rungs on the career ladder. Another possibility is that workers have both "accurate" and "inaccurate" sources of labour market information that they struggle to distinguish between. For instance, a manager may be keen to retain a particular staff member who is considering employment elsewhere. Thus the manager may overstate the chances of pay and progression within the firm. If the worker can not tell that this is "bad" information, it may lead him to raise his future income expectations. Indeed, in situations where workers only receive relatively poor quality information, one would expect them to be no more (and possibly even less) realistic than their student peers.

Finally, some young adults may not realise the value of the information that they hold, or how it applies to them and their future; they may discard (or give less weight) to some important information as they see it as irrelevant. For example, a young worker may know what a 30 year old employee in his organization is paid. But he (perhaps unrealistically) views his current job as a stop-gap solution, and believes he will have entered an entirely different industry in a few years time. He therefore does not fully incorporate the information he holds on the wages of 30 year olds into his income expectations. Indeed, this interpretation seems to be consistent with the findings of Smith and Powell (1990). They find that students can accurately estimate average graduate wages, but expect their own salaries to be a lot higher. Hence, although they are well informed about average wages (i.e. hold relatively good information), they do not necessarily incorporate this into their expectations (i.e. make good predictions of their own future salary).

Linking these points back to my opening paragraphs, simplistic assumptions that young adults hold a combination of “full-information” and “rational expectations” may be based on a rocky foundation. It seems that young adults may be missing some important labour market information, making further research in the spirit of Black et al (2003) ever the more important. Yet economists must also develop a better understanding of how young adults use the labour market information that they hold. In many ways, it is difficult to believe they will give it the appropriate weight when making schooling decisions as is often assumed in a rational expectations framework; indeed, as the quote from Becker suggested at the start of this chapter, young adults probably do not realize their own limitations and tend to over-estimate their chances of good fortune[102]. Thus understanding exactly what information young adults hold, how they use it, and the effect this has on their schooling decisions should become an important area of future economic research.

[102] See Chevalier et al (2009) for some empirical evidence on this topic.

Chapter 4

Disadvantaged children's "low" expectations: Is the UK really so different to other industrialized nations?

"Children who grow up in inferior environments may expect less of themselves and may not fully develop their academic potential because they see little hope for ever being able to complete college or use their schooling in any effective way"

Cameron and Heckman (1999), Financing College Tuition: Government Policies & Educational Priorities, page 76-124

Educational attainment has risen dramatically across the developed world over the past 15 to 20 years, with particularly strong growth in university participation. For instance, the proportion of young adults obtaining a bachelors degree in Sweden has increased from 24% in 1995 to around 40% today (OECD 2009). Similar increases have been seen in Switzerland (9% to 30%), Japan (25% to 39%) and New Zealand (33% to 52%), amongst others. Despite this rising trend, access to tertiary education remains unequal. Children with well-educated, affluent parents are still over-represented in higher education, with relatively limited opportunities for those from disadvantaged backgrounds. For instance, low socio-economic status children make up less than a quarter of university entrants in Sweden, despite representing over a third of the total population (Department for Business, Innovation and Skills 2009). Likewise, the university entrance rate for disadvantaged children in the Republic of Ireland is half that of the national average. This issue has taken on particular prominence either side of the Atlantic, where worries mount over equality of educational opportunity and social mobility (see Blanden and Machin (2004), Machin and Vignoles (2004)). This is illustrated by the fact that only one in ten low income children from the US earn a degree by age 24 compared to almost three-quarters from wealthier families (Department for Business, Innovation and Skills Research Paper 2009). Likewise, Chowdry et al (2008) show that around one in three of all UK children enter higher education, compared to only 10% of those with the least educated parents.

Consequently, several governments across the world have introduced policies to increase the number of disadvantaged children entering higher education. Ensuring children hold high expectations and aim for university from an early age is seen as a crucial step towards reaching this goal. In other words, there is a belief that future educational plans made during adolescence have a significant impact on later university attainment, and a concern that poor children's low expectations may be limiting their opportunities to succeed. For instance, Khoo and Ainley (2005) suggest that children's socio-economic background is linked to their educational intentions, which in turn influences their educational attainment. Likewise, Cowan (2009) finds that low educational expectations increase the probability of children engaging in "risky" behaviour. "Widening access" policies have thus emerged across the developed world that encourage children to hold high expectations for the future.

Such schemes are particularly prevalent in the US, with examples including the "Gear-up" and "I have a dream" programmes. Both try to encourage disadvantaged children to apply for university and aim towards a professional career. Similar initiatives have emerged during the last decade in the UK. This includes the "Aim Higher" and "Gifted and Talented" schemes, which give disadvantaged children near the end of compulsory schooling (around age 16) the chance to interact with professional workers and to be "mentored" by current university students. Hence the educational expectations of teenagers in the US and UK seem to be an on-going and prominent concern of academics and policymakers alike.

The aim of this chapter is to gain a better understanding of the link between family background and children's educational plans across the OECD. Specific research questions are set out in the following section, including whether advantaged children are more likely to expect entry into university than their disadvantaged peers, and if this is just reflecting differences in the schools they attend and test scores achieved during adolescence. I also consider how family background influences the educational expectations of the most academically able 15 year olds within each country; in particular, I investigate disadvantaged children who score high marks on an international assessment are more or less likely to plan for higher education than an affluent pupil who achieves a mark around the national average. In turn, I contemplate whether widening access schemes towards the end of compulsory schooling (age 15-16) could play an important role in university access for the poor.

This research is undertaken in a cross-nationally comparative framework, with a particular focus on countries within the UK (England, Scotland, Northern Ireland) and North America (US and Canada). In doing so, I make an important addition to the existing literature by illustrating whether the socio-economic gap in children's educational expectations is particularly 'big' in the aforementioned countries, and thus if these countries have a particular policy need to encourage disadvantaged children into considering higher education. It also allows me to explore whether age 15 test scores and school level factors explain more of the socio-economic gap in the UK than other developed countries. Thus I add a comparative depth to studies such as Reynolds and Pemberton (2001), Chowdry et al (2009), Chevalier et al (2009) and Emmerson et al (2005) who all investigate adolescents' educational expectations within a single national setting (either the US or UK).

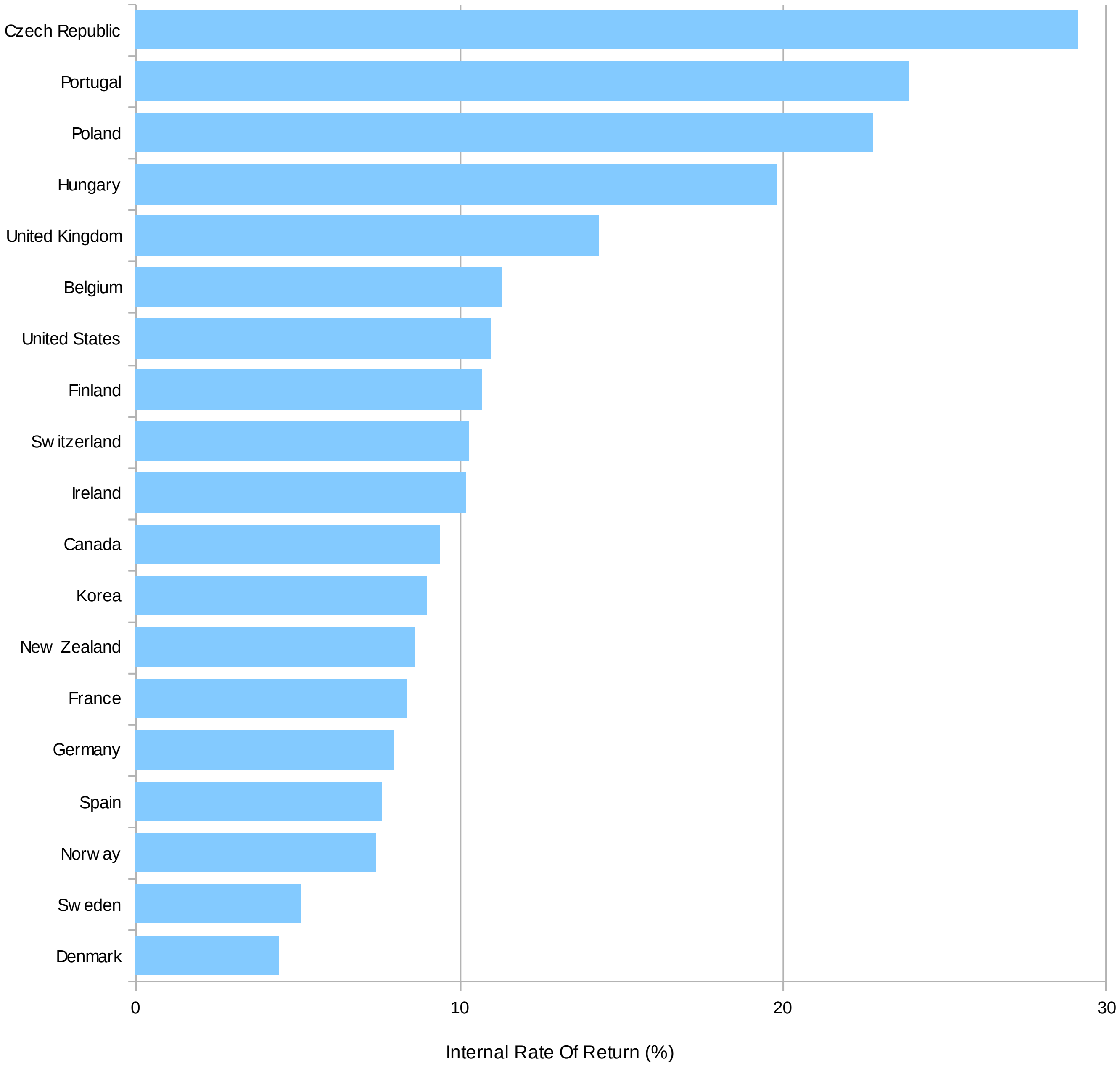
I begin in section 4.2 by reviewing the relevant literature. This includes a discussion of how children's expectations differ to their aspirations, and how this concept is linked to their eventual educational attainment. Section 4.3 describes the Programme for International Student Assessment (PISA) data on 15 year old children that I analyse and my empirical methodology. During sections 4.4 and 4.5 I discuss the results. I conclude in sections 4.6 and 4.7 with a discussion of how my findings may inform educational policy in countries that encourage disadvantaged children to 'aim higher', like the UK.

4.2. Motivation

Obtaining a tertiary qualification offers substantial economic rewards, particularly in the US and UK. In the latter, the government often stresses that the average graduate earns £100,000 more over their lifetime than if they were to enter the labour market without a tertiary qualification (see Department for Business, Innovation and Skills 2008). Figure 4.1 illustrates that such returns are high by international standards; obtaining a tertiary level qualification seems to be a particularly good investment for young adults either side of the Atlantic.

Friedman and Friedman (1980), amongst others, have argued that all young adults who can benefit from university should have access to the returns it offers, regardless of their family background. One reason is that this may lead to a more equitable society. Yet it is also important for economic efficiency. Labour is a scarce resource that needs to be allocated appropriately, but the brightest children may be excluded from the best jobs if they are unable to 'fully develop their academic potential'.

Figure 4.1. The (private) internal rate of return to obtaining a bachelors degree across a selection of OECD countries



Notes:
1 Figures refer to the 'private internal rate of return', which the OECD describes as an 'investment approach' to calculating the benefit of holding a degree.
2 OECD (2008) states that the internal rate of return (IRR) is the discount rate required to equalise the financial benefits of a degree (mainly the lifetime wage premium) to the financial costs (mainly the opportunity cost of foregone wages). Thus the higher the IRR, the greater the returns to university as a human capital investment.
3 Source: Data drawn from OECD Education At A Glance Report (2008), Table A10.2, page 196

As noted in the introduction, the fact that disadvantaged children are under-represented amongst the undergraduate population has become a particularly topical issue in the US and UK. Some argue, like Chowdry et al (2008), that this mainly reflects the underachievement of disadvantaged children at secondary school, and hence policymakers should concentrate on raising this group's academic attainment[103]. Yet various governments have introduced "widening access" schemes that tackle this issue more directly. This includes promoting university to "first generation" students, providing financial incentives to young adults and their institutions, distance or e-learning opportunities and setting targets for the proportion of adolescents who enter higher education (the current target being 50% in the UK).

Part of this policy is to address the concern that disadvantaged children do not see university as a realistic possibility; that it is 'not for the likes of them' (Chowdry et al (2008), Shields and Mohan (2008)). As suggested by Cameron and Heckman (1999) at the start of this chapter, some disadvantaged children may perceive there to be a lack of opportunity to complete higher education, which stops them from applying[104]. Consequently, policymakers in the US and UK have introduced a series of programmes to raise disadvantaged children's expectations of being able to obtain a tertiary level qualification.

At this point, it is important to draw a distinction between children's "expectations" and their "aspirations". The former implies a realistic assessment of future outcomes, while the latter reflects children's hopes and dreams (Gutman and Akerman (2008)). So if a child *expects* to obtain a university qualification, they truly believe that they will go on to complete this level of education. It is this concept that I attempt to explore in this chapter. However, one must consider whether 15 year old children (the age of children that I study) are able to make such realistic assessments of the future. Drawing from the developmental literature, Gottfredson (2002) notes that, around age 14, children are beginning to recognize the need for compromise in their educational and occupational goals. Likewise, Gutman and Akerman (2008) suggest that at this

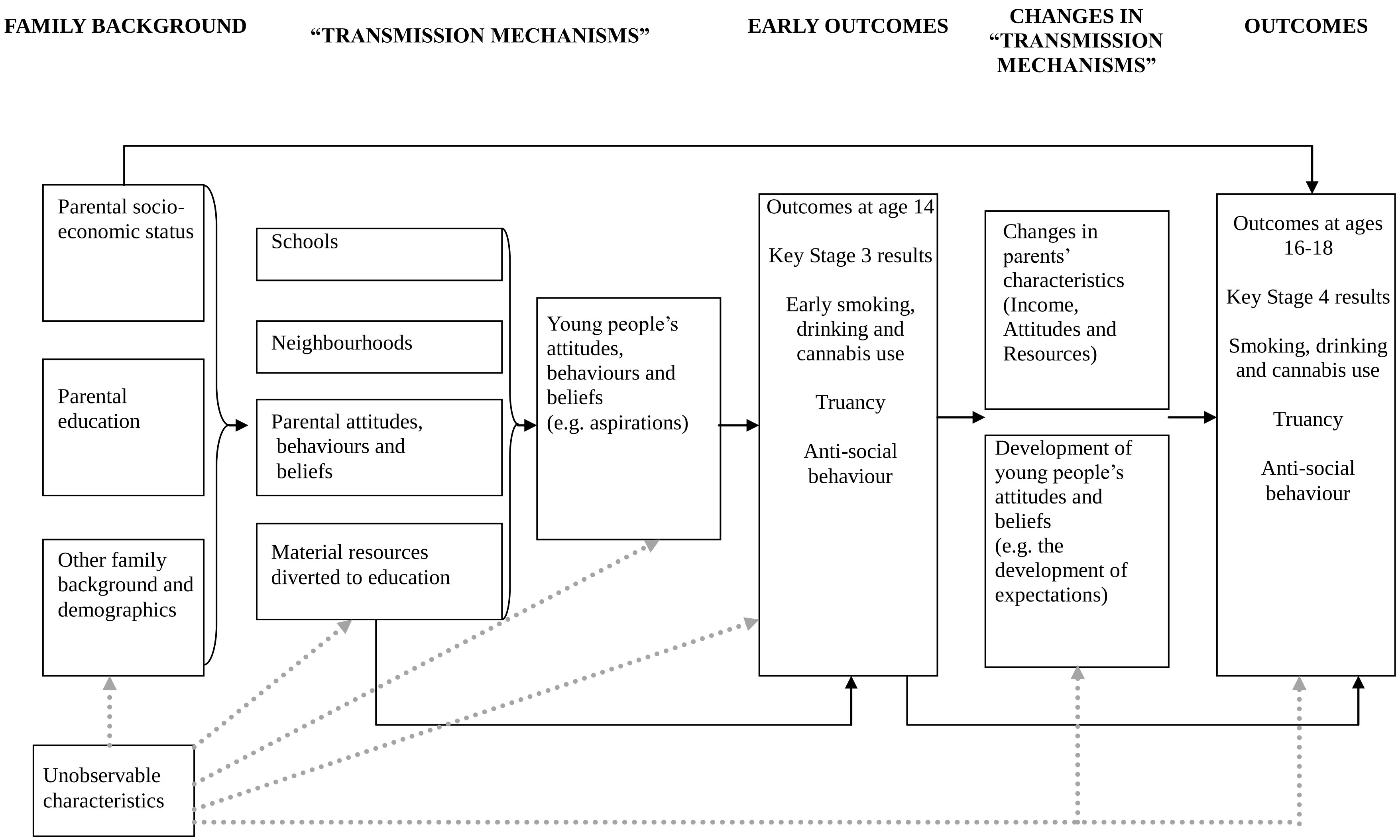
[103] The authors do note, however, that aspiration and expectation raising activity may encourage hard work in school, and thus help to close this gap in attainment.

[104] For instance a report by the Sutton Trust (2008), a UK based charity, states: 'exam grades on their own will not necessarily lead to university if young people do not have a high level of expectation and make ill-informed decisions'

age young people 'relinquish their most preferred choices and settle for more acceptable, available choices', recognising the external constraints that they face. From a different prospective, Morgan (1998) finds that adolescents' educational expectations are not 'irrational fantasies'; rather, they are grounded in logical thinking, and vary with the marginal costs and benefits associated with such continued schooling. Hence there is evidence which suggests that young adults are able to distinguish between their aspirations and expectations. Accurately capturing such details in a social survey is, however, another matter. I shall further elaborate on this point when discussing the PISA data in the following section.

It is also important to make clear that the value of any scheme that attempts to 'raise disadvantaged children's expectations' is based on the assumption that this will have a causal influence on their later behaviour and attainment. A conceptual model to illustrate this relationship is set out in Figure 4.2, drawing upon the work of Chowdry et al (2009).

Figure 4.2. A model linking family background to children's aspirations, expectations and outcomes, based upon Chowdry et al (2009)



This framework recognizes the multi-dimensional nature of family background, based around measures of parental education, occupation (socio-economic status) and income. The authors then specify four “transition mechanisms” (schools, neighbourhoods, parental attitudes, material resources) by which family background influences children’s attitudes, behavior, beliefs (including their aspirations for the future) and outcomes at age 14. The main focus of this chapter is, however, on the next stage of the model – the transition from adolescence to young adulthood (i.e. from age 14 to 18). During this period, children may change their attitudes, behaviours and beliefs about the future, which, in turn, alters their academic trajectory[105]. Based on the work of Gottfredson (2002), I propose that one key development between these ages is that children begin to recognize the external constraints that they face, and thus start to develop *expectations* about their future (regarding, in particular, higher education). These expectations then become the key behavioral “transmission mechanism” that encourage greater effort and investment in school and less “risky” behaviour (drinking, drug use and early sexual activity) between ages 14 and 18 which, in turn, leads to higher educational attainment.

It is important to recognise, however, that this is not a static relationship; children will continually revise these expectations, based on their on-going attainment. Indeed, it is likely that higher expectations lead to higher attainment, which leads to continued high expectations, and so forth. Yet, as one can not identify the exact age at which such feedback begins, it has proven to be methodologically challenging to estimate the extent to which one factor is driving the other. Nevertheless, several authors have explored the association between these variables, with some attempt to address the direction of causality. For instance, Khoo and Ainley (2005) investigate the educational plans and achievement of a sample of Australian teenagers. Estimating a structural equation model, they show that children's expectations are strongly associated with their later outcomes, even after controlling for a host of potentially confounding factors. In a similar manner, Reynolds and Pemberton (2001) find that expecting to go to university at age 15 is almost a prequisite for actual later attendance in the US; they show that less than 3% of children who do not expect to go to university actually obtain a degree by the time they turn 30. Likewise, Morgan

[105] This framework also recognises that family circumstances and parental characteristics will continue to play a role.

(2004) uses a regression based path analysis to investigate whether educational expectations held during the mid-teens determines entry into post-compulsory schooling in the US. In turn, he finds evidence of a strong and statistically significant association. Of course, economists may express concerns about the potential endogeneity of expectations in any regression based set-up, particularly due to omitted variable bias. Consequently, Morgan (2004) shows that expectations remain a highly significant predictor of later outcomes using an instrumental variable analysis. However he also recognizes the difficulties of identifying such models in this set-up. Brown et al (2004) use similar methods to Morgan, and find a strong and highly significant relationship between children's expectations and later attainment in the UK. Using panel data, with measurement of young adults' educational expectations at several ages, Morgan (2005) finds university plans are serially correlated across time. He suggests that this is consistent with an underlying dynamic causal relationship between expectations and attainment as described above. Similarly, Chowdry et al (2009) find that a number of disadvantaged children in England stop believing that they will enter university between ages 14 and 16, and that these teenagers subsequently make less academic progress than their peers who maintain high expectations. In a wider context, Cowan (2009) investigates the relationship between American teenagers' educational expectations and their chances of engaging in risky behaviour. Using an instrumental variable analysis, he finds that 'anticipated schooling has an effect on behaviour above and beyond the effect of realized schooling' and thus that raising children's expectations of completing university may prove to be an inexpensive way of reducing their tobacco, marijuana and alcohol consumption.

Given the above, any difference between advantaged and disadvantaged children's educational expectations will lead to a division in their behaviour, attainment and eventual graduation rates. Indeed, the framework set out in Figure 4.2 suggests that such a divergence in beliefs may well occur; expectations are assumed to have six primary determinants (schools, neighbourhoods, parental attitudes, family resources, childhood attitudes and prior attainment) all of which are associated with family background. For instance, advantaged children will tend to go to better schools, where teachers may build their children’s academic confidence and emphasise their ability to complete this level of study. Similarly, it may be that only well-educated parents stress the wider benefits of learning (meeting new people, broadening horizons, growing up) to their offspring, who become driven towards higher education as a result. Availability of resources will also determine children's expectations; those from less fortunate households may believe they are credit constrained and thus do not have the necessary finance to complete higher education. There may also be peer and role model influences, both in school and the wider community, where disadvantaged children do not see university as a realistic goal because they do not know any adult who has completed higher education and have few friends who believe they can achieve the same. Attitudes may also be transferred between generations, such as ambition and work ethic, which could influence children's educational plans via the extent they are willing to stretch themselves in the future. Finally, as expectations involve the recognition of external constraints, they will be tempered by children's pre-existing skill, with large socio-economic differences in academic ability already evident at age 14 (see Hanushek and Woessmann (2010) for a survey of the international evidence).

The analysis I undertake in this chapter is motivated by the theoretical framework and empirical analysis described above, which suggests that children's expectations have an important influence on their later academic attainment, and that there are likely to be large differences in these expectations between socio-economic groups. But is this difference bigger in some countries than in others?

In my first research question I consider the size of the socio-economic gap in children's educational expectations across the OECD, focusing on results for the US and UK. Although concern has been expressed about the difference between advantaged and disadvantaged children's expectations in these countries, I have never seen it put into a comparative context. It is therefore difficult to know if the socio-economic gap in expectations is especially 'big' within these countries, and if they have a *particular* need to encourage disadvantaged children into considering higher education. In making such comparisons, I emphasize similarities and differences between the US, Canada and constituent parts of the UK. Such countries share a number of common features (language, geographical location, broad political and labour market systems) but also contrast in many ways (schooling systems, levels of poverty, inequality and social mobility). Identifying differences between these countries may therefore provide some insight into why the socio-economic gap in children's educational expectations varies across developed countries. In summary:

Research Question 1. What is the absolute size of the gap between advantaged and disadvantaged children's expectations of completing university? Is this gap particularly large in the US and UK compared to other members of the OECD?

Of course, one may argue that differences in test scores could be entirely responsible for the socio-economic gap in children's educational expectations; the only reason why advantaged 15 year olds are more likely to expect entry into university than their disadvantaged peers is that they perform better on school assessments around this age (e.g. "outcomes at age 14" in Figure 4.2)[106]. On the other hand, sizeable differences may remain even after controlling for test performance at this point in time (as per the arguments made above). That is to say that disadvantaged 15 year olds may be less likely to expect a university education than their wealthy peers, even if they score equally well on assessments nearing the end of compulsory schooling. This may suggest that raising disadvantaged children's educational expectations is an important part of widening access policy. I again consider whether this is a specific problem to the US and UK, or if the situation here is comparable to other parts of the developed world. My second research question is therefore:

Research Question 2. Do the higher educational expectations of advantaged children only reflect their better test performance at age 15? After controlling for scores on this assessment, is the socio-economic gap in the UK and US particularly large in comparison to other members of the OECD?

In answering the question above, I am not able to make a causal statement about the relationship between children's test scores, socio-economic status and their expectations. As laid out in Figure 4.2, children are assumed to begin making firm educational plans between ages 14 and 16, yet the exact point in time is almost impossible to identify. It could be that children start thinking seriously about

[106] Consequently, if ones aim is to encourage disadvantaged children to make early university plans and raise their higher education participation, then the optimal policy would probably be to develop their earlier cognitive skills. Encouraging children to enter university at ages 15/16 would be of comparatively little use in closing the social class gap in university participation.

university from a younger age than I can measure (e.g. 14), which has already had an impact upon their motivation at school, and is thus reflected in their scores on the PISA test (taken at age 15)[107]. In other words, the process of educational expectations influencing motivation and behaviour has already begun, causing test scores to be endogenous in the models that I estimate. This may be a particularly big issue in countries like England where children have to make educational decisions at a young age, and who receive regular information of their performance in national assessments.

Likewise, one may argue that the measure of "ability" that I use (scores on a cognitive test at age 15) is likely to be endogenous to social background. In other words, the PISA data does not contain an indicator of children inert talent, but rather scores on a test that have probably been influenced by social background in themselves. Whenever I refer to this test scores as children's "ability" it is imply for convenience. It is, however, vital for the reader to understand that this is measured at age 15 and hence highly likely to have already been affected by socio-economic background. I, nevertheless, argue that results controlling for this term is of interest despite this potential endogeniety with respect to family background. In particular, it will show the difference in expectations for advantaged and disadvantaged children who have managed to reach the same level on the PISA (although careful interpretation of what this result means is needed). I shall return to this point in the following sections.

If socio-economic differences in children's expectations remain even when they score the same mark on this age 15 cognitive test, what other factors might be playing a role? As noted above, good schools are a valuable resource that affluent parents may provide their children, and which probably have an impact on adolescents' educational plans[108]. In my third research question, I explore whether differences in school level factors can explain the socio-economic gap in children's expectations, above and beyond the influence they have on children's age 15 test performance.

[107] In other words expectations at prior time points (that I am unable to control for) are confounding the relationships that I estimate.
[108] For instance, advantaged children may receive more support and encouragement from teachers to apply to university, have more information about the benefits of higher education and application procedures, and be the subject of strong and positive peer effects.

Research Question 3. Do school level factors help to explain the socio-economic gap in teenagers' educational plans (after accounting for differences in age 15 test performance?) Are such factors particularly important within the UK?

Finally, an issue I have not considered thus far is that if childhood expectations really do help determine later attainment, then should government policy concentrate on raising all children’s expectations as high as possible? Such a policy faces two possible difficulties. Firstly, higher expectations have a greater chance of being unmet. Hence one must trade off the possible boost to attainment against the potential disappointment children may suffer. Secondly, a widening access policy that encourages all youngsters into considering university may lead to applications from some children who are not suited to this level of study. Thus one may actually *harm* some children’s long-run outcomes if policies to promote the benefits of higher education actually lead them into making inappropriate investment decisions.

The above argument suggests that widening access schemes should be properly targeted, and that we should be particularly concerned if disadvantaged children who perform well in assessments at the end of compulsory schooling think ‘university is not for the likes of them’. In other words, it seems of obvious benefit that the most talented children in a country make early plans to enter higher education, regardless of their family background. Again, it is important to realize that the measure of “academic talent” that I use is children’s test scores at age 15 (which will have already been influenced by socio-economic background as discussed above). Nevertheless, I again argue that this remains an interesting question – do the few disadvantaged children who manage to overcome adversity up to age 15 and score highly on the PISA tests believe they have the means to obtain a university degree? I particularly focus on whether this is the case for disadvantaged children in the US and UK, where initiatives like the 'gifted and talented' programme specifically target the educational expectations of this academically able group[109]. Thus, my final research question is:

[109] A recent report by the Office for Fair Access (201) in the UK suggests that academically talented pupils should be identified no later than the end of year 9 (age 14). This is a slightly earlier age than the expectations measured within PISA (where children are typically aged 15).

Research Question 4. Are high scoring children on the PISA math's assessment from disadvantaged backgrounds more or less likely to expect entry into university than their advantaged peers who receive a mark around the national average? Is this of particularly great concern in the US and UK compared to other developed countries?

In answering these four research questions I make a number of contributions to the existing literature and current policy debate. Firstly, I do not know of any other study that has considered whether the socio-economic gap in British and American children's educational expectations is "big" or atypical in comparison to the other OECD countries. Secondly, no other study, to my knowledge, has considered whether advantaged children's higher expectations are reflecting anything other than this group's higher test scores at age 15. Finally, I am unaware of any other paper that considers whether children who manage to overcome their disadvantaged background and score highly on cognitive tests at age 15 are more or less likely to expect entry into university than advantaged children who do not perform as well on such assessments.

4.3. Data

The data I use are drawn from the 2003 round of the Programme for International Student Assessment (PISA); a study of 15 year-olds' cognitive skills held every three years. Although 46 countries took part, I restrict my analysis to 33 industralised nations[110]. In each country, a minimum of 150 schools were included in the sample, selected with probability proportional to size. Thirty students were then randomly selected from within. Average response rates of both schools (90%) and pupils (90%) were high, though this varies moderately between countries[111] . Further details are available in the PISA 2003 technical report (OECD 2004b). A set of sampling weights are also provided by the survey organisers that tries to correct for the unit non-

[110] Here I treat the constituent parts of the United Kingdom (England, Scotland and Northern Ireland) as separate countries. Likewise, I separate Flemish from French Belgium.
[111] The lowest of which was England, at 64% for schools and 77% for pupils . Micklewright et al (2010) investigate this non-response, and create an alternative set of responses weights (as opposed to those provided in the dataset by the survey organisers) to try and correct for bias in the estimates. They show that the UK only moves one place in the PISA ranking of children's test scores once these weights have been applied.

response. The achieved sample size, across the 33 countries I consider, is 224,094.

As part of the study, children were asked to complete a questionnaire. This included the question:

"Which of the following do you **expect** to complete" [emphasis in original question]

Lower secondary education (Middle or junior high school)
Upper Secondary education (High school)
Post-secondary non-tertiary (Vocational/technical certificate after high school)
Tertiary "Type b" education (Associate's degree)
Tertiary "Type a" education of higher (Bachelors degree or higher)

Country specific options were provided in the questionnaire. The phrases in brackets illustrate these for the US. The primary outcome I analyse in this chapter is whether the child ticked the top category (bachelors degree or higher). Response rates to this question were very high. Table 4.1 shows that almost 99% of children responded, from a low of 93% in France to a high of 100% in Poland. Consequently, I exclude the few (1%) observations where educational expectations are missing[112].

[112] These observations are not a random selection from the population. Rather they tend to be children of lower ability, who also do not have complete information on family background.

Table 4.1. Sample sizes and missing expectations data across the OECD countries

Country	Starting sample size	Complete educational expectation data	% Missing expectations data
Poland (POL)	4,383	4,381	0.0
Finland (FIN)	5,796	5,793	0.1
Italy (ITA)	11,639	11,631	0.1
Japan (JAP)	4,707	4,700	0.1
Korea (KOR)	5,444	5,440	0.1
Spain (ESP)	10,791	10,776	0.1
Turkey (TURK)	4,855	4,852	0.1
Hungary (HUN)	4,765	4,756	0.2
Slovakia (SLOV)	7,346	7,328	0.2
Greece (GRE)	4,627	4,613	0.3
Portugal (PORT)	4,608	4,594	0.3
Switzerland (SWZ)	8,420	8,393	0.3
Sweden (SWE)	4,624	4,605	0.4
Australia (AUS)	12,551	12,492	0.5
Mexico (MEX)	29,983	29,845	0.5
Denmark (DEN)	4,218	4,191	0.6
Scotland (SCO)	2,723	2,707	0.6
USA (USA)	5,456	5,419	0.7
Austria (AUT)	4,597	4,558	0.8
Iceland (ICE)	3,350	3,324	0.8
Ireland (IRE)	3,880	3,848	0.8
Luxembourg (LUX)	3,923	3,892	0.8
Northern Ireland (NI)	2,853	2,829	0.8
Belgium(French) (BELFREN)	2,958	2,931	0.9
Norway (NOR)	4,064	4,023	1.0
New Zealand (NZ)	4,511	4,447	1.4
Netherlands (NLD)	3,992	3,902	2.3
Belgium(Flemish) (BELFLEM)	5,838	5,696	2.4
England (ENG)	3,959	3,817	3.6
Czech Republic (CZE)	6,320	6,076	3.9
Germany (GER)	4,660	4,457	4.4
Canada (CAN)	27,953	26,707	4.5
France (FRA)	4,300	3,997	7.0
TOTAL	**224,094**	**221,020**	**1.4**

Notes:
1 Missing data refers to item non-response only. Details on unit non-response can be found in the OECD (2004) Technical Report.
2 Data sorted by the percentage of missing observations
3 Source: Calculations by author using the PISA 2003 dataset

Recall that my concern in this chapter is children's expectations (realistic assessments of their future), *not* their aspirations (idealistic goals). As noted in section 4.2, the developmental literature suggests that by the time of the PISA study (approaching age 16), children typically separate one concept from the other. Indeed, there has been work in the sociological literature that compares children's expectations to their aspirations around this age (see Patton and Creed (2007)). Such studies usually distinguish between the two concepts by altering and emphasizing the operative word (e.g. asking children what they would "<u>like</u>" to do, and then what they "<u>expect</u>"). Yet, to my knowledge, there has been little work on the validation of such questions in quantitative surveys. In particular, there seems scant evidence of whether such subtle phrases are able to elicit the appropriate information from respondents. Unfortunately, the question asked in PISA shares much of the same criticism. It emphasizes the word **"<u>expect</u>"** using bold, underlined letters, yet provides children with no further instruction. Hence this is the only guide they have towards reporting their expectations rather than their aspirations. Whether such subtle wording can be adequately translated into other languages, as required in this cross-national analysis, is a further concern.

If this question is actually capturing children's *aspirations*, then the proportion reporting that they *"expect"* to complete university will be significantly higher than current graduation rates[113]. If this only occurs in certain countries, then these nations will out-lie from the rest. Indeed, if it is a translational issue that is causing this problem, language will be a common theme amongst these outlying nations. I search for such patterns in Figure 4.3. Specifically, in each country I compare the proportion of children who expect to obtain a degree (that I have calculated from the PISA data) with actual graduation rates drawn from OECD (2009). The 45 degree line is where the proportion of children expecting to complete university equals actual graduation rates.

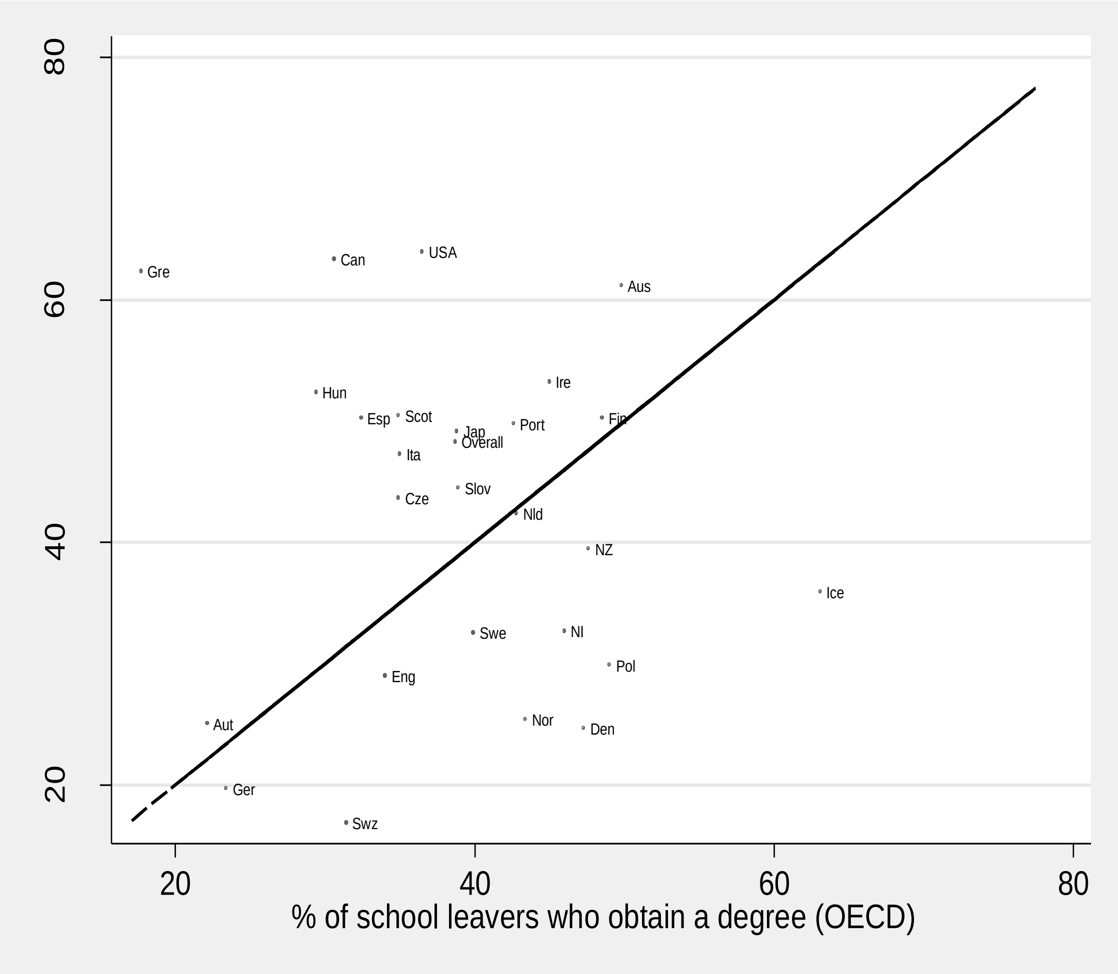
[113] Of course, such a finding may just reflect that children are not very good at predicting the future (the question does capture children's expectations, it is just that these expectations are inaccurate). Nevertheless, if this pattern occurs consistently across all nations then one may question whether this is actually capturing adolescents' expectations

Generally, Figure 4.3 suggests that responses are not out of touch with reality; 45% of OECD children expect to complete university against actual graduation rates of 40%. Indeed, several countries, including England, sit below the 45 degree line; the proportion of children expecting to enter university is *below* actual completion rates. I do note, however, that there are some countries where one may have concerns. For instance the proportion of Canadian and American children "expecting" to obtain a bachelor's degree is significantly higher than actual graduation rates. However, Reynolds and Pemberton (2001) point out that there are high drop out rates from university (at least in the US), and as such the proportion of US children expecting to complete university are at least in-line with current *entrance* rates[114]. It is also interesting to see that the proportion of children expecting to complete university varies quite substantially across the English speaking countries, suggesting that this cross-national variation is not simply due to a difference in language.

Thus, despite concerns with some countries, the overall pattern of response is quite encouraging, and generally seems to be consistent with a measure of children's expectations. Hence these data do seem to be of value in answering the research questions I set out in section 4.2. Yet I am unable to investigate (and thus rule out) other potential problems regarding measurement error. For instance, it might be that advantaged children have a tendency to report their expectations and disadvantaged children their aspirations, and that this particular response pattern varies across OECD countries. Likewise, I advise caution in interpreting results for less developed members of the OECD like Turkey, Greece and Mexico, where "expectations" often seem to be out of touch with reality.

[114] Reynolds and Pemberton (2001) noted a similar finding when using the American NLSY sample. OECD (2009) suggests that 65% of US school leavers enter university, very similar to the number I find expecting to obtain a degree. Unfortunately this information is not available for Canada.

Figure 4.3. Proportion of children expecting to obtain a degree versus actual graduation rates



Notes:
1 Data on the % of school leavers who obtain a degree (x-axis) has been drawn from OECD Education At A Glance Report (2009), Table A3.2, page 74. This refers to net graduation rates (*i.e.* as the sum of age-specific graduation rates). See Annex 1 of OECD (2009) for further details. Information on Mexico, Luxemburg, Korea, France and Belgium not available in this data. Data on the proportion of children who expect to obtain a degree (y-axis) is based on my calculations from the PISA 2003 data.

2 Data is only available for the UK as a whole (not separately for England, Scotland and Northern Ireland) in the OECD (2009) report. Hence I use data on higher education participation for these countries Data taken from: http://www.dcsf.gov.uk/rsgateway/DB/SFR/s000716/SFR10_2007v1.pdf for England
http://www.scotland.gov.uk/Publications/2009/11/20112425/4 for Scotland
http://www.delni.gov.uk/he-api0607.pdf for Northern Ireland

3 To calculate statistical significance I have compared the proportion of children expecting to complete university (drawn from the PISA data) to the OECD 'Education At A Glance' figures of actual graduation rates. I assume the latter refer to the population, hence conduct a one sample test of the PISA figures against these values.

4 Distance from the 45 degree line (where average expectations = actual graduation rates) are statistically significant in all countries at the 1% level except Germany, Austria, Netherlands, Slovakia, Czech Republic and Finland

5 Source: Calculations by author using the PISA 2003 dataset of the y-axis and OECD Education at a Glance on y-axis. For country by country sample sizes see Table 4.1

I now turn my attention to the variables that I use to distinguish between children from “advantaged” and “disadvantaged” backgrounds. The first such measure is parental education. As described in section 4.2, parents with more schooling might place greater emphasis on their children's education, or instill a taste for learning in their off-spring. Likewise, parents may be able to provide more information and encouragement about going to university if they hold a tertiary qualification themselves, and perhaps act as educational role models. Parental education will also be a key factor driving household income and children’s cognitive development. This is therefore a key distinguishing feature between children from advantaged and disadvantaged backgrounds.

Information on parental education was captured through the sampled children as part of the PISA background questionnaire. Specifically, children were asked to report the level of education their mother and father completed at school and what type of tertiary qualifications they hold[115]. Schlutz (2005) and Jerrim and Micklewright (forthcoming) investigate possible measurement error in such reports using the PISA 2006 wave, where parents and children were asked separately to report mother’s and father’s level of education[116]. They gave the same category in two thirds of cases, though this was higher (around 86% of occasions) when the parents held a degree.

These responses were then recoded by the survey organisers into ISCED levels of education, a measure designed by UNESCO to aid cross-national analysis (though some differences in definitions across countries may remain - Steedman (2001)). The highest ISCED level achieved by either parent is then used to create the “highest parental education” variable. Table 4.2 shows how this is distributed across each of the OECD nations, including a “missing” category where this information is unavailable (typically 5-10% of cases)[117].

---

[115] Note that children were instructed to report this information for their *mother and father like figures.* Consequently, children living in a household with a complex family structure, for instance with a step-mother or step-father, may not be reporting the education of their biological parents. I have experimented with including a variable that captures this in my analysis. However, I have chosen to exclude it as the estimated effect was usually small and statistically insignificant.

[116] The parental questionnaire that contains this information was an “international option” in 2006. This information is therefore only available for 11 countries, and has relatively high rates of non-response. I could not use the 2006 data for this analysis as it did not contain a question on children’s educational expectations.

[117] These children are not a random sample from the population. Rather, they disproportionately come from children who performed poorly on the PISA test and come from less well-off families. One may

Table 4.2. Distribution of highest parental education across OECD countries

	% None	% ISCED 1	% ISCED 2	%ISCED 3B or 3C	% ISCED 3A	% ISCED 4	% ISCED 5B	% ISCED 5A +	% Missing
Turkey	4	31	20	1	23	0	7	13	0
Austria	0	1	5	34	9	6	29	14	2
Poland	1	0	2	21	42	12	7	15	0
Northern Ireland	0	1	11	24	4	19	17	17	7
Portugal	19	20	16	3	15	0	7	17	2
Mexico	8	18	25	3	13	0	14	19	1
Switzerland	2	2	18	23	7	6	21	19	3
Ireland	1	5	10	0	17	27	20	19	2
England	1	1	6	22	5	20	18	20	9
Denmark	1	0	8	8	12	10	36	20	4
Italy	0	2	22	5	16	19	13	20	1
New Zealand	3	1	5	13	9	18	20	20	12
Luxembourg	4	9	2	6	8	13	24	20	13
Germany	5	1	8	19	5	15	15	23	10
Norway	0	0	3	4	6	22	36	24	4
Hungary	0	0	7	20	16	24	7	24	1
Slovakia	1	0	2	14	38	17	3	25	1
France	2	2	12	22	21	0	11	25	6
Iceland	0	2	10	9	11	26	14	26	2
Spain	3	18	7	2	16	10	13	27	5
Greece	0	8	12	4	18	16	13	27	0
Czech Republic	0	0	1	21	37	7	2	28	4
Korea	2	5	14	11	31	0	7	30	1
Scotland	3	0	5	17	14	0	23	30	8
Belgium(Flemish)	1	2	4	4	15	17	20	30	7
USA	1	1	4	0	29	16	13	34	3
Canada	0	1	4	0	19	15	21	34	6
Finland	0	3	7	0	21	3	30	36	1
Belgium(French)	2	3	4	4	12	10	21	36	8
Sweden	2	1	7	7	21	0	21	37	5
Australia	1	1	11	2	16	13	13	39	3
Japan	0	3	3	6	30	0	17	41	0
Netherlands	1	4	10	0	6	27	0	45	7
OECD	**3**	**6**	**10**	**8**	**17**	**11**	**16**	**26**	**4**

Notes: 1 Data sorted by the percentage of children who reported either parent as holding an ISCED level 5A+ qualification. 2 Figures refer to row percentages.
3 ISCED level 0 refers to no formal school, level 1 is equivalent to primary education only, level 2 is lower secondary education, level 3B/3C refers to basic vocational education, 3A is upper secondary education, level 4 is post secondary education (either short vocational courses of preparation for tertiary education), level 5B is specialised vocational education, level 5A is a university education

argue that this could be driving some of the cross-national differences observed. Given the relatively small non-response in the majority of countries, I do not attempt any correction for this issue. However, I do I include a "missing" dummy variable in all subsequent regressions to ensure these children are not dropped from the analysis.

(bachelors degree), while level 6 refers to doctorates.
4 Source: PISA 2003 dataset. For country by country sample sizes see Table 4.1

As sample sizes become small for certain groups, I recode this information into three broad categories (low, medium and high) following a similar re-categorisation in the Luxemburg Income Study (a well known dataset often used in cross-national research). I define high as holding a tertiary qualification (ISCED 5B or 5A+), medium as post-secondary schooling but no experience of higher education (ISCED 4) and low as completed secondary schooling or less (ISCED 0-3). This broad categorisation also helps to ensure that I have a sufficient number of observations within the “advantaged” and “disadvantaged” groups that I define later in this section.

The second measure of family background that I use to distinguish between advantaged and disadvantaged children is parental occupation. This variable is probably the best proxy available for household income and financial resources (which are unfortunately not collected as part of the PISA study) that play an important role in the development of children's educational expectations as laid out in section 4.2. Parental occupation will also pick up relevant aspects of social class, such as the societies, cultures and communities the child has grown up in.

As with parental education, information on mother's and father's occupation was collected directly from the sampled children. Specifically, they were asked the title of their mother's and father's main job and a description of the type of work this involves. Responses were coded by the survey organisers into four digit ISCO codes (the International Labour Organisation's occupational classification), which assigns the reported occupation to one of over 300 categories. Schulz (2005) investigates the potential measurement error in this data using the 2006 PISA field trial. He found that parents and children reported the same occupational group (defined in this fine level of aggregation) on roughly six out of ten occasions. My experimentations with the final 2006 PISA sample revealed similar results.

Children's responses were then coded by the PISA survey organisers into the quasi-continuous ISEI index of occupational status, designed by Ganzeboom et al (1992). This index assigns each occupation a score between 16 and 90, depending upon the relevant "inputs" (educational level required) and "outputs" (the salary commanded) from that particular job[118]. Hence this is an objective occupational scale which is designed to be correlated with income. Moreover, Ganzeboom et al (1992) specifically designed this scale to aid the type of cross-national analysis I undertake in this chapter, and have thus attempted to validate it as a measure of socio-economic status across a large number of developed countries (although some still question aspects of its validity – see Bukodi, Dex and Goldthorpe (forthcoming)). Nevertheless, the ISEI index remains an attractive measure against the possible alternatives (such as an aggregation of the 4 digit ISCO codes into the 9 major occupational groups). Summary statistics for the distribution of this variable across the OECD countries can be found in Table 4.3. It is interesting to note that the distribution of the ISEI index generally seems to be quite similar across countries, with very little cross-national variation in the 10th and 90th percentiles, and only slightly more at the 25th and 75th percentiles.

[118] The OECD describes: "The index captures the attributes of occupations that convert parents' education into income. The index was derived by the optimal scaling of occupation groups to maximise the indirect effect of education on income through occupation and to minimise the direct effect of education on income, net of occupation (both effects being net of age)."

Table 4.3. Distribution of highest parental occupation (ISEI index) across countries

	Percentile							
	10th	25th	50th	75th	90th	mean	SD	% Missing
Mexico	24	28	33	54	69	42	19	5
Turkey	23	29	45	49	66	42	15	12
Portugal	26	30	39	51	69	43	16	3
Poland	23	33	43	53	67	45	15	2
Spain	25	30	43	54	70	45	17	4
Greece	26	31	46	56	69	46	17	6
Korea	29	37	45	51	69	46	13	3
Austria	27	34	45	56	69	47	16	4
Hungary	30	38	45	56	69	48	15	6
Ireland	29	34	49	57	69	48	16	4
Italy	29	34	49	56	70	48	16	2
Luxembourg	29	34	50	56	69	48	17	4
Switzerland	29	34	48	55	69	48	16	3
Northern Ireland	29	34	48	59	69	48	17	6
Denmark	29	38	51	57	69	49	15	3
France	29	34	51	59	70	49	17	4
Belgium(French)	29	37	51	67	70	50	17	6
Germany	30	38	51	59	70	50	16	9
Japan	33	38	45	55	69	50	15	11
Slovakia	30	37	50	64	69	50	16	4
England	29	35	51	66	70	50	17	7
Belgium(Flemish)	29	35	51	66	70	51	17	5
Canada	29	38	51	65	69	51	16	7
Finland	29	34	51	67	71	51	17	1
Sweden	30	38	51	66	70	51	16	3
Scotland	30	40	51	66	70	51	16	4
Czech Republic	33	42	51	64	69	52	15	4
Netherlands	30	39	51	67	70	52	16	7
New Zealand	29	40	51	66	69	52	16	16
Australia	30	43	52	69	69	53	16	5
Iceland	29	43	53	67	71	54	17	2
USA	30	40	56	67	71	54	16	6
Norway	34	43	53	69	71	55	15	3
OECD	**28**	**34**	**49**	**59**	**69**	**48**	**17**	**5**

Notes:
1 Data refers to points on the ISEI scale of occupational status, as described in this section. On this scale, higher values indicate a more prestigious occupation.
2 Countries sorted by mean ISEI score
3 Source: Author’s calculations using PISA 2003 data. Country by country sample sizes in Table 4.1

The third variable that I use to measure family background is children's reports of the number of books at home. It has been argued that this is correlated with a number of aspects of family background including parental education, household income and social origin (Ammermueller and Pischke 2009, Schuetz et al 2008). Yet the same authors also suggest that it picks up factors like the value parents place on their children’s education and the encouragement they provide with regards to schooling. Likewise, the PISA survey organisers argue it is a measure of the 'home educational resources' available to the child. Hence this picks up such residual aspects of family background that are not fully captured within my measures of parental education and occupation, but are nevertheless likely to be important in the development of children's expectations.

This information was also reported by the participating children in the background questionnaire. Specifically, they were asked 'how many books are there in your home' (excluding magazines, newspapers and textbooks) with six possible options. However, the two bottom and two top categories contain a rather sparse number of observations. Thus I combine the bottom (0-10, 11-25, 26-100), and top (101-200, 201-500, above 500) three fields to form low and high groups, along with a 'missing' category, following a strategy similar to that used by the survey organisers (see OECD 2004c page 283). As with my re-categorisation of parental education, this also helps to ensure that I have sufficient observations within the “advantaged” and “disadvantaged” groups that I define later in this section. The distribution of this variable across the OECD countries can be found in Table 4.4.

Table 4.4. Distribution of the number of books in the home across OECD countries

	% 0-100 books	% Over 100 books	% Missing
Mexico	86	10	4
Turkey	79	18	3
Portugal	68	31	2
Greece	65	34	2
Belgium(Flemish)	59	38	4
Scotland	58	40	2
Northern Ireland	58	40	2
USA	58	40	2
Ireland	58	40	2
France	57	41	2
Poland	58	41	1
Switzerland	56	41	2
Netherlands	55	42	3
Austria	56	42	2
Luxembourg	54	44	2
Italy	54	45	1
Japan	54	45	1
Denmark	52	45	3
Slovakia	54	45	1
England	50	45	5
Belgium(French)	51	46	4
Finland	52	47	1
Germany	46	48	6
Canada	43	49	8
Korea	51	49	0
New Zealand	47	51	3
Spain	47	52	1
Australia	41	56	2
Sweden	40	58	2
Hungary	41	58	1
Norway	36	61	2
Iceland	36	63	2
Czech Republic	33	63	4
OECD	**55**	**42**	**3**

Notes:
1 Data refers to row percentages
2 Data sorted by % over 100 books
3 Source: Author’s calculations using PISA 2003 data. Country by country sample sizes in Table 4.1

Another key covariate in this chapter, which forms an integral part of my second, third and fourth research questions, is children's “academic performance”. As part of the PISA 2003 study, children sat a two hour test. The PISA consortia claim that this measures children's 'functional ability' (how well they can use the concepts examined in 'real life' situations) in three domains (reading, maths and science). In 2003, maths was assigned as the major domain, where the vast majority of questions children were asked were on this topic. All test questions were explicitly designed with cross-national comparability in mind. Answers were summarized by the survey organizers into a single score for each of the three domains using an ‘item-response model’; the intuition being that true ability in each subject is unobserved, and must be estimated from the answers to the test. Consequently, five ‘plausible values’ are generated for each pupil, estimating their true proficiency in each subject. These scores were scaled by the survey organizers to have a mean (across all OECD countries) of 500 points and standard deviation of 100. Throughout my analysis, I use the first of these plausible values for the maths domain to control for children's cognitive skills at age 15[119]. In doing so, I once again remind the reader that this variable requires careful interpretation – as it is measured at age 15, it does not reflect children’s inert ability or some sort of “natural” talent. Rather, it will reflect all the inputs into children’s education production function up to age 15 (e.g. time/goods input from parents, schooling resources, quality of peers) which will have been influenced by their socio-economic background. Through the remainder of the paper, when I use the term “ability” I am doing so as a short-hand for their cognitive skills in the maths domain as measured at age 15 (and that this will in part reflect socio-economic background). Moreover, in the regression models that follow I only control for maths scores at age 15 as the correlation between test scales is high (r≈0.8 between maths and reading, and the same between maths and science), with little change in my substantive results when I also include the reading and science scales[120]. The distribution of children's maths test scores across all the countries I consider can be found in OECD (2004a).

[119] I experimented using the other plausible values, and by running five separate models and averaging the estimated coefficients and standard errors. Results are very similar to those presented.

[120] Note that only around half the children within each country actually answer questions in each of “minor” PISA domains (reading and science). Scores are estimated by the study organisers for the remaining children using a Rasch modelling approach.

Finally, measurement of school level factors has important implications regarding my penultimate research question. In this chapter, I simply use a fixed effect specification to eliminate all the between school variation. This will show whether all school level factors combined are able to explain a sizeable proportion of the advantaged-disadvantaged gap. This will thus capture a broad array of differences between children at different schools, including the information, advice and guidance provided by staff about higher education, and the influence of peers on their choices. Yet using a fixed effect specification does limit my ability to identify what exactly at the school level is driving the results; for instance whether it is something that educational institutions actually provide (e.g. where teachers encourage and assist university applications) or some wider factor (e.g. the community in which the child has grown up).

In the following section, I use the aforementioned variables in a logistic regression model of children's educational expectations. In all models, I also control for gender and whether the child was a first or second generation immigrant (as this group may be under different pressure from their family to complete higher education)[121]. Likewise, in all estimations I include 'missing' categories (dummy variables) to ensure children are not dropped from the analysis when pieces of information are unavailable. Thus the final form of this model is:

[121] Children had to answer three questions regarding whether they, their mother or their father was born outside the country that they are taking the test in. I define a child as an “immigrant” if they answer yes to any of these three questions.

$$\log\left(\frac{\Pi(\boldsymbol{E}_{ij})}{1-\Pi(\boldsymbol{E}_{ij})}\right)=\boldsymbol{\alpha}\ +\boldsymbol{\beta}_1.\boldsymbol{Sex}_i+\boldsymbol{\beta}_2.\boldsymbol{I}_i+\boldsymbol{\beta}_3.\boldsymbol{SES}_i+\boldsymbol{\beta}_4.\boldsymbol{I}_i*\boldsymbol{SES}_i+\boldsymbol{\beta}_5.\boldsymbol{Ab}_i+\boldsymbol{\beta}_6.\boldsymbol{Sch}_j$$

Where:

$\Pi(E_{ij})$ = Probability of the child expecting to graduate from university, where

E= 1 if the child expects to complete university, 0 otherwise

Sex = A binary indicator of the child's gender (0 = female, 1 =male).

I= Whether the child is a first or second generation immigrant (0 = Native , 1 = Immigrant)

SES = A vector of variables capturing the child's socio-economic background. This includes:

- Highest parental education – A set of two dummy variables, one referring to "some post-secondary education but no tertiary" (medium) the other "tertiary and above" (high). (Ref: "compulsory schooling or less" – i.e. low)
- Number of books in the home - A single dummy variable referring to whether there are "more than 100 books" (high) in the family home (Ref: "Less than 100 books" - low)
- Highest parental occupation measured on the ISEI scale - Entered as a piecewise linear term with knots at the 10th, 25th, 50th, 75th and 90th percentile of the national ISEI distribution

Ab = A vector reflecting children's cognitive skills at age 15 as measured by their scores on the PISA maths test. These measures are entered as piece-wise linear components, with knots at the 10th, 25^{h}, 75th and 90th percentile of the national test score distribution.

Sch = A school level fixed effect. In specifications where this is not included, I allow for the complex survey design (clustering of children within schools) by making the appropriate adjustment to the estimated standard errors.

i = Child i

j = School j

I then use this model to generate predictions of how likely a hypothetical child with given characteristics is to expect to complete university. Specifically, I create these predictions for:

1. *An "advantaged" child*

Defined as:

- Either of their parents holds a tertiary qualification ("high" parental education)
- There are over 100 books in the family home ("high" books)
- The highest occupation of their parents sits at 75th percentile of the national ISEI distribution
- Country native
- Female

2. *A "disadvantaged" child*

Who I define as:

- Neither parent has completed any post-compulsory schooling ("low" parental education)
- There are less than 100 books in the family home ("low" books)
- The highest occupation of their parents sits at 25th percentile of the national ISEI distribution
- Country native
- Female

I then calculate the difference between these two predictions in order to compare the expectations of "advantaged" and "disadvantaged" groups.

Three model specifications (and sets of predictions) are estimated to answer the first three research questions set out in section 4.2 (I return to my final research question, comparing the expectations of disadvantaged children who score highly on the PISA test to lower performing children from affluent backgrounds, in the following section). Specifically, I use a sequential approach in building this model, first illustrating the "overall" expectations gap between advantaged and disadvantaged groups, then adding in terms (age 15 test scores and a school level fixed effect) to highlight the extent of the expectations gap conditional on these factors. Hence in the first model I exclude the PISA test score and school measures (i.e. β_5 and β_6 are constrained to zero). This allows me to identify the absolute difference between advantaged and disadvantaged children's educational expectations, and whether this gap is particularly big in the US and UK compared to other developed nations:

Model 1: $$\log\left(\frac{\Pi(\boldsymbol{E_{ij}})}{1-\Pi(\boldsymbol{E_{ij}})}\right) = \alpha_1 + \beta_1.\boldsymbol{Sex_i} + \beta_2.\boldsymbol{I_i} + \boldsymbol{\beta_3.SES_i} + \boldsymbol{\beta_4.I_i * SES_i}$$

In the second specification I include children's scores on the PISA maths test, but do not control for school level factors (i.e. I now obtain estimates of the vector β_5, but β_6 is still constrained to 0). This allows me to address my second research question; is the difference between advantaged and disadvantaged children's expectations just reflecting differences in test performance at age 15 (again comparing the US and UK to other developed countries)?

Model 2: $$\log\left(\frac{\Pi(\boldsymbol{E_{ij}})}{1-\Pi(\boldsymbol{E_{ij}})}\right) = \alpha_1 + \beta_1.\boldsymbol{Sex_i} + \beta_2.\boldsymbol{I_i} + \boldsymbol{\beta_3.SES_i} + \boldsymbol{\beta_4.I_i * SES_i} + \boldsymbol{\beta_5.Ab_i}$$

The third model is the full specification as outlined above. This now also includes a school level fixed effect (β_6) that will indicate the extent to which school level factors can explain the advantaged-disadvantaged expectation gap.

Model 3: $$\log\left(\frac{\Pi(\boldsymbol{E_{ij}})}{1-\Pi(\boldsymbol{E_{ij}})}\right) = \alpha_1 + \beta_1.\boldsymbol{Sex_i} + \beta_2.\boldsymbol{I_i} + \boldsymbol{\beta_3.SES_i} + \boldsymbol{\beta_4.I_i * SES_i} + \boldsymbol{\beta_5.Ab_i} + \boldsymbol{\beta_6.Sch_j}$$

In doing so, I note the work of Hsiao (1986) who states that fixed effect logistic regression models result in inconsistent parameter estimates when the number of observations within a higher level unit (e.g. observations in a panel or, in this case, children clustered within schools) is small. However, as the PISA study draws a sample of 30 children from each institution, this should not cause substantial bias in my results[122].

[122] Hsiao (1986) notes that the bias in fixed effect logistic regression models become small when the number of observations clustered within a higher level unit (e.g. children within schools) is

I present all results mainly in terms of log-odds. This measure is more attractive than alternatives like the odds ratio and marginal effect (predicted probabilities) as they are not sensitive to the point on the logistic distribution on which they are estimated, and are therefore not influenced by differences between countries in the absolute proportion of children who expect to complete higher education. I illustrate this point in Table 4.5. The second and third columns present the proportion of children expecting to complete university depending on whether either of their parents holds a degree. Column 4 provides the marginal effect (the percentage point difference between columns 2 and 3) while column 5 illustrates the difference in terms of the log odds.

Comparisons between countries can look very different depending on which measure is used. Take England, one of my countries of interest, and Korea. The difference between the second and third column is quite similar in terms of the log-odds (1.69 in Korea to 1.73 in England), but very different when considering the marginal effect (22 percentage points compared to 39)[123]. This is being driven by the fact that, across the population, Korean children are generally more optimistic about their prospects of completing university than those in England (77% expect to obtain a degree in the former, compared to 29% in the latter). As my concern in this chapter is the expectations of disadvantaged children *relative* to their advantaged peers, I prefer the log-odds as it abstracts from the *absolute* proportion of the population believing that they will complete higher education. However, appreciating that this metric is rather cumbersome to interpret, I also occasionally present predicted probabilities to assist the reader's understanding.

---

approximately greater than 20.

[123] Across all countries, the estimated correlation between the marginal effect and log odds is 0.78.

Table 4.5. Children's expectations of completing university, depending on whether either of their parents' holds a bachelor's degree

	% Expecting to complete university if *either* their mother or father holds a degree	% Expecting to complete university if *neither* their mother or father holds a degree	Marginal Effect	Difference in log-odds
Mexico	69	53	16	0.68
Netherlands	56	31	25	1.04
Finland	67	41	25	1.07
Belgium(French)	48	23	25	1.13
France	56	28	28	1.19
Portugal	73	45	28	1.20
Sweden	50	23	27	1.21
Canada	80	54	26	1.23
Ireland	76	48	29	1.23
Greece	81	55	25	1.25
Norway	45	19	26	1.25
Italy	72	41	31	1.31
USA	82	55	28	1.32
Denmark	47	19	28	1.33
Scotland	73	41	33	1.36
Japan	69	36	34	1.38
Spain	74	41	33	1.41
Australia	80	49	30	1.43
Turkey	93	75	18	1.49
Belgium(Flemish)	59	24	35	1.52
Switzerland	39	12	27	1.55
New Zealand	70	32	38	1.60
Luxembourg	71	33	39	1.60
Slovakia	73	35	38	1.61
Northern Ireland	64	26	37	1.62
Iceland	64	26	38	1.62
Korea	93	71	22	1.69
England	60	21	39	1.73
Austria	59	20	40	1.75
Germany	44	12	32	1.75
Czech Republic	73	31	42	1.79
Poland	67	23	44	1.92
Hungary	87	41	45	2.26
OECD average	**70**	**40**	**30**	**1.25**

Notes:
1 The column labeled 'marginal effect' illustrates the percentage point difference between children's expectations. This is the difference between the first two columns. Conversely, the final column illustrates the difference between the same figures, but in terms of the log-odds.
2 The final row, labeled OECD average, refers to when one combines data across all 33 OECD countries considered.
3 Countries sorted by the difference in terms of log-odds
4 Source: Author's calculations using PISA 2003 data. Country by country sample sizes in Table 4.1

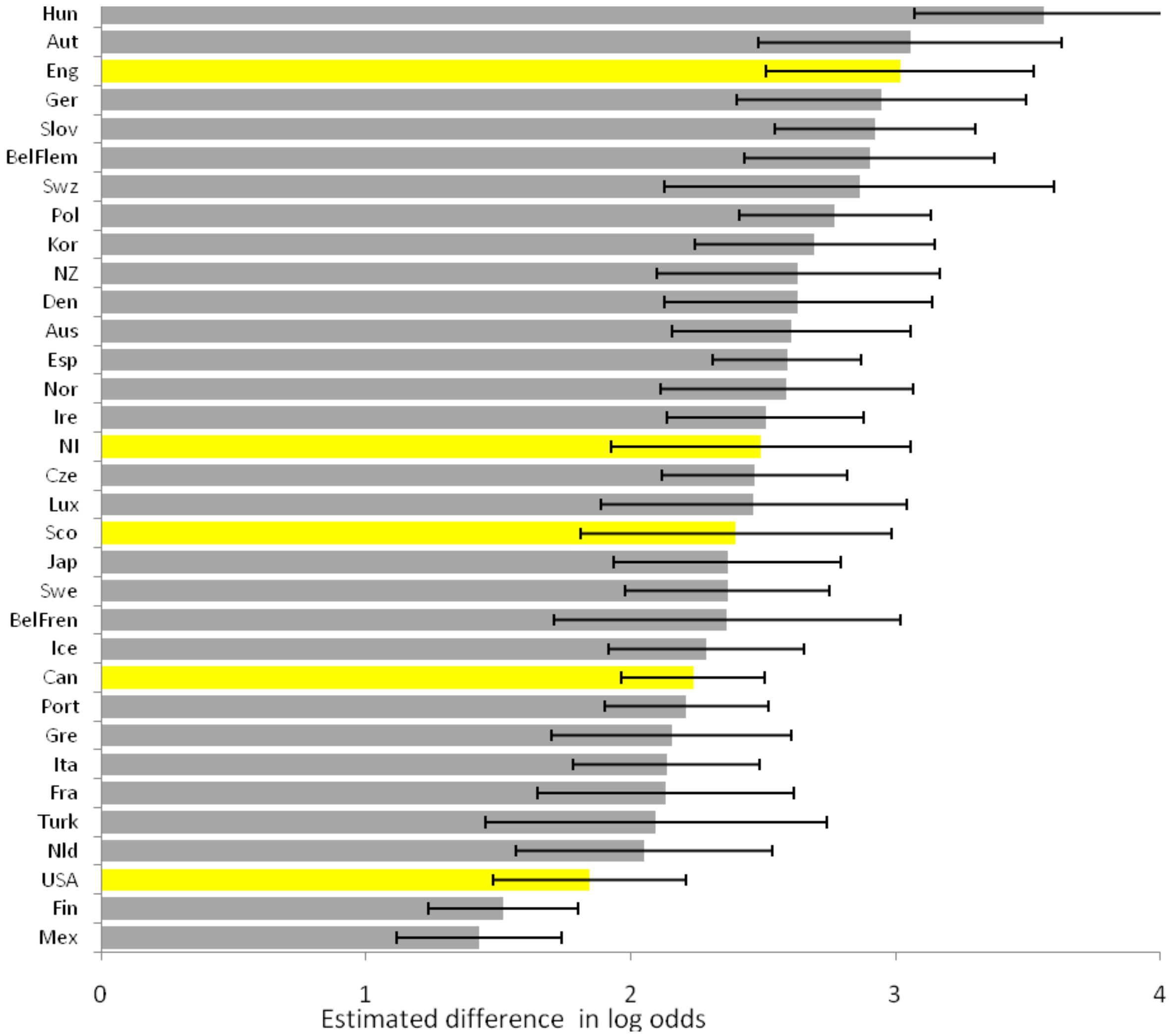
4.4. Results

I shall now present results for the first three research questions set out in section 4.2. At times, I shall use phrases like 'effect' and 'influence'. I do not intend to imply causality when doing so; I am merely using this shorthand for convenience. A full set of parameter estimates can be found in Appendix 4.1.

Figure 4.4 illustrates the absolute gap between advantaged and disadvantaged children's expectations (i.e. the results from ‘Model 1’ as described in Section 4.3)[124]. In all countries, disadvantaged children are less likely to expect entry into university than their more affluent peers. This gap is generally big, around two and a half log odds, and is always significantly different from zero at the one percent level. To put this into perspective, if a hypothetical disadvantaged child had a 50% chance of expecting to complete university, the probability for their identical (but advantaged) peer would be closer to 90%[125]. Yet as I suggested when motivating this research, the US and parts of the UK stand out from the rest of the OECD - but for quite contrasting reasons.

[124] A list of country abbreviations can be found in the left hand column of Table 4.1.
[125] These probabilities were calculated using the formula: probability = exp(log[odds]) /(1+exp(log[odds])). Log odds of 0 correspond to a probability of 50%. Log odds of 2.5correspond to a probability of 92%. Hence, in this hypothetical example, a difference of 2.5 log odds (i.e. the difference between advantaged and disadvantaged groups) leads to a 42% difference in the probability.

Figure 4.4. Estimated difference between advantaged and disadvantaged children's plans to complete higher education (based on model 1)



Notes:

1 Results are based upon predictions from the regression "model 1" that I describe on page 170. These predictions are based upon measures of highest parental education, highest parental occupation and number of books in the home. Other controls include gender, immigrant status and an interaction between immigrant status and the four measures of advantage listed above. Other controls include gender, immigrant status and an interaction between immigrant status and the three measures of advantage listed above. Note that these results are drawn from the first model specification and so do not include information on the children's PISA test scores.

2 The thick, solid bars represent the difference between advantaged and disadvantaged children's expectations of completing university, as measured in log-odds. The thin black line at the ends of these bars illustrates the 95% confidence interval of this estimate.

3 Country names corresponding to abbreviations can be found in the first column of Table 4.1

4 Source: Author's calculations using PISA 2003 data. Country by country sample sizes in Table 4.1 and Appendix 4.1. Dependent variable in regression was whether respondent expected to obtain a degree (coded 0 if they did not and 1 if they did)

Starting with England, the difference in log odds between advantaged and disadvantaged children's expectations is roughly 3.0; the third largest estimate in the OECD. Comparable gaps can be found in Germany, Switzerland and Austria; countries where access to university is restricted to children on the appropriate educational “track”[126]. Dustman (2004) shows that this “track” is strongly associated with family background, hence one would anticipate there to be large differences between advantaged and disadvantaged children's expectations in these countries. For the gap to be just as big in England, where such tracking does not take place, is quite a striking result. Likewise, England sits amongst a number of Eastern European countries (Hungary, Slovakia, Poland) that are well known for their educational polarisation (see Shavit and Blossfeld 1993).

England is also ranked noticeably higher than the other Anglo-Saxon countries. This includes Scotland and Northern Ireland (the other constituent parts of the UK) where the estimated difference in log odds is closer to the OECD average of around two and a half. However the 95% confidence interval (the thin black line running through the centre of each bar) suggests that caution is required when interpreting this result. One can only reject the null hypothesis that England is significantly different to Scotland at an unconventional threshold (15% level). A similar story emerges when making comparisons to other English speaking countries; the estimated socio-economic gap may be weaker in Australia, Ireland and New Zealand than in England, but one cannot rule out that this is just a reflection of sampling variation. Nevertheless, one should not lose sight of the comparison between England and the broader array of OECD countries; the association between family background and children's educational plans is stronger here than 12 other countries at the 5% and a further two at each of the 10% and 15% levels. Hence these initial estimates suggest that the difference between advantaged and disadvantaged children's expectations of completing university do seem quite big in *parts* of the UK (England) compared to other members of the OECD, though this cannot be generalised to the other constituent nations (Scotland and Northern Ireland).

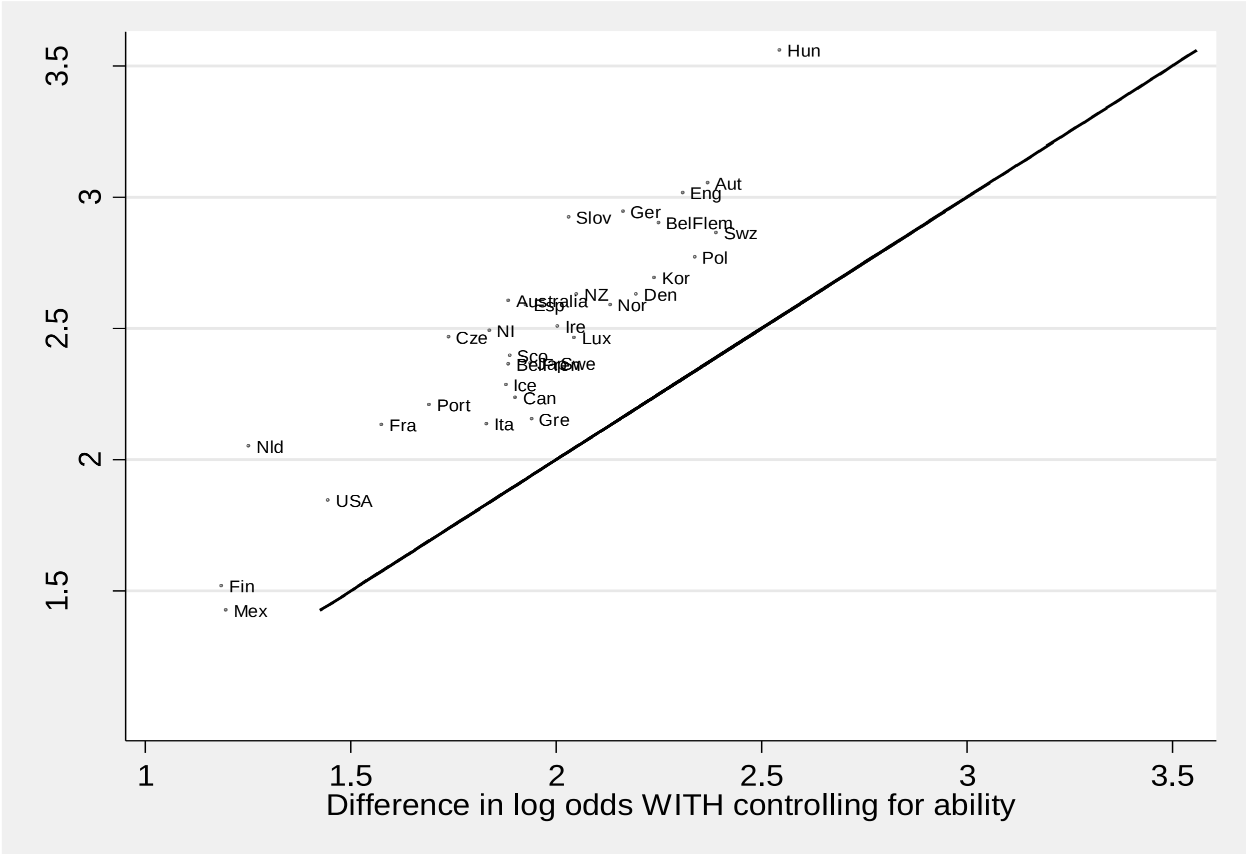
[126] In these countries, children are sorted into different schools by their level of ability at a young age (known as “tracking”).

The US also stands out in Figure 4.4 - though in quite the opposite direction. The socio-economic gap here on the log-odds scale is under two; the third *smallest* out of the 33 countries considered. Moreover, this result is not just a matter of sampling variation; results for the US are statistically different to 16 other nations at the 5% level and a further 6 at the 10% level. This includes England, where there is obviously a stark contrast in the size of the socio-economic gap (a difference which is statistically significant at the 1% level), and Canada (which is itself relatively low down the international ranking). Hence it seems that the absolute gap between advantaged and disadvantaged children's educational expectations is actually quite “small” in the US – at least when compared to other members of the OECD.

It is important to stress that these findings are *unconditional* on children's age 15 test score. Figure 4.5 illustrates how results change once I control for respondents’ scores on the PISA maths assessment[127]. Specifically, the estimates of the differences shown in Figure 4.4 (‘model 1’) are presented on the y-axis, with analogous figures after taking into account children's test scores (‘model 2’) along the x-axis. The distance each point is from the 45 degree line illustrates the difference between these two sets of results.

[127] Once again, any claim of causality faces the issue that ability is (potentially) endogenous.

Figure 4.5. Estimated difference between advantaged and disadvantaged children's plans to complete higher education, before and after controlling for differences in age 15 test scores



Notes:
1 Figures on the y-axis correspond to the difference between advantaged and disadvantaged children's educational expectations when I do not control for children's scores on the PISA math's tests (these are the results from Figure 4.4). Analogous figures, once I have included these measures of children's age 15 test performance, can be found along the x-axis. The 45 degree line illustrates where this leads to no change in results. Thus the further points are from this line, the more age 15 test performance 'explains' (in a statistical sense) the difference between children's expectations.

2 Results are based upon predictions from regression model 1 and 2 that I describe on page 170. These predictions are based upon measures of highest parental education, highest parental occupation and number of books in the home. Other controls include gender, immigrant status and an interaction between immigrant status and the four measures of advantage listed above. Other controls include gender, immigrant status and an interaction between immigrant status and the three measures of advantage listed above.

3 Country names corresponding to abbreviations can be found in the first column of Table 4.1.

4 Source: Author's calculations using PISA 2003 data. Country by country sample sizes in Table 4.1 and Appendix 4.1. Dependent variable in regression was whether respondent expected to obtain a degree (coded 0 if they did not and 1 if they did)

It is important to make clear the careful interpretation that needs to be placed on these results. Firstly, the measure of "ability" included is children's test scores at age 15 which, as discussed previously, is likely to be endogenous to social background (i.e. these scores are likely to have been heavily influenced by the circumstances the child has been born into). Hence when I describe "change" in the following results, it does ***not*** mean how much of the expectations gap is due to differences in "natural talent" or inert "ability" (as this is not what the PISA test captures). Rather, I am referring to "change" in a statistical sense only (i.e. how do the parameter estimates change when I include this extra term). In a similar manner, the conditional association between family background and expectations that I am left with after controlling for this term should ***not*** be considered the difference in educational plans for advantaged and disadvantaged children with the same "natural talent"[128]. Rather it is the difference in educational expectations between advantaged and disadvantaged children who have managed to score the same on the age 15 PISA maths assessment.

It seems that age 15 test scores "explain" (in a statistical sense) only a modest amount of the difference between advantaged and disadvantaged children's expectations; in most countries, the difference in log-odds has dropped by 20 to 30 percent. For instance, the estimated difference in log odds has dropped from 1.8 to 1.5 in the US and 2.2 to 1.9 in Canada – a pattern that is reflected across most of the OECD. The absolute change in England, however, is slightly greater – the log odds have fallen from 3.0 to 2.3. Nevertheless, in no country does it seem to be the case that advantaged children's higher expectations only reflect their superior performance on tests taken at age 15. Although this certainly plays a role, the majority of the socio-economic gap in children's educational expectations remains unexplained. For instance, even after controlling for differences in age 15 test scores, an advantaged child from England is still over 30 percentage points more likely to expect completion of university than their disadvantaged (but equally able) peers[129].

[128] The endogeniety of the PISA test scores with respect to socio-economic background probably leads to an underestimation of differences in expectations between advantaged and disadvantaged children of the same "natural" ability. Specifically, it is likely to be the case that advantaged children may have received more investment so end up scoring higher on cognitive assessments in the teenage years (compared to disadvantaged children who would have probably received less investment and hence will not score so highly on such tests). See Cunha et al (2006) for a discussion of life-cycle skill formation and further explanation of why disadvantaged children are likely to score lower on cognitive assessments taken in the teenage years.

[129] I calculate this figure of 30 percentage points for England by converting the estimated difference in

Do England and the US still stand out in the international ranking once I have taken children's test scores into account? Table 4.6 suggests that cross-national patterns are indeed similar to before.

The US is again placed towards the top of the table, while England is near the bottom. However, the socio-economic gap in the latter is now significantly stronger than in only six other OECD countries at the five percent level, and a further two at the ten percent level. Once again, this does not include Scotland and Northern Ireland, where the log of the odds ratio remains roughly around the OECD average of 1.9. Consequently, after controlling for age 15 test scores, there is only rather limited evidence that the difference between the educational expectations of advantaged and disadvantaged children is unusually large in the UK. On the other hand, the association between family background and children's expectations of completing university remains significantly weaker in the US than in other OECD countries (14 at the 5% level, with a further 4 at the 10% level). This, once more, includes Canada and England. Hence, while there is now only quite limited evidence that the socio-economic gap is particularly big in the UK compared to other OECD countries (as in statistically significant differences can only be detected in comparison to a few countries – though I note the issue of potential Type II errors), there remains a strong suggestion that it is relatively small in the US.

terms of log odds into an estimated marginal effect (in a similar manner to what I have shown in Table 4.5)

Table 4.6. Difference between the expectations of advantaged and disadvantaged children after controlling for differences in age 15 test performance

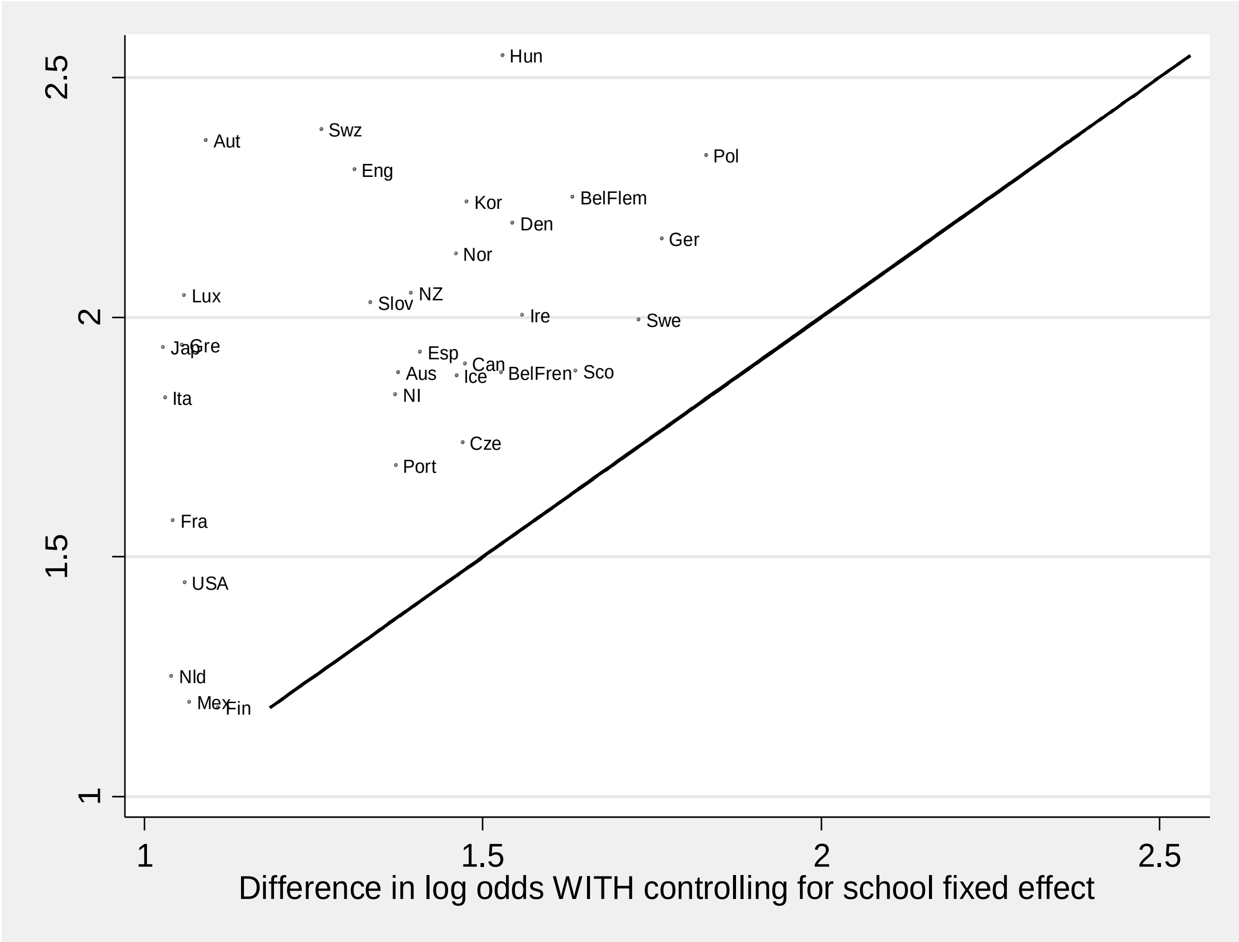
	Difference between advantaged and disadvantaged (in log odds)	SE	Significantly different from England?	Significantly different USA?
Finland	1.19	0.15	***	-
Mexico	1.20	0.16	***	-
Netherlands	1.25	0.26	***	-
Turkey	1.41	0.31	***	-
USA	1.45	0.19	***	
France	1.58	0.26	**	-
Portugal	1.69	0.17	*	-
Czech Republic	1.74	0.19	*	-
Italy	1.83	0.17	-	-
Northern Ireland	1.84	0.32	-	-
Iceland	1.88	0.19	-	-
Belgium (French)	1.88	0.35	-	-
Australia	1.89	0.25	-	-
Scotland	1.89	0.31	-	-
Canada	1.90	0.15	-	*
Spain	1.93	0.15	-	**
Japan	1.94	0.22	-	*
Greece	1.94	0.24	-	-
Sweden	1.99	0.20	-	**
Ireland	2.00	0.20	-	**
Slovakia	2.03	0.19	-	**
Luxemburg	2.04	0.31	-	*
New Zealand	2.05	0.27	-	*
Norway	2.13	0.25	-	**
Germany	2.16	0.30	-	**
Denmark	2.20	0.26	-	**
Korea	2.24	0.24	-	***
Belgium (Flemish)	2.25	0.25	-	***
England	2.31	0.27	-	***
Poland	2.34	0.19	-	***
Austria	2.37	0.28	-	***
Switzerland	2.39	0.38	-	**
Hungary	2.55	0.25	-	***

Notes:
1 The final two columns illustrate whether the estimated difference in log odds are significantly different to those in England and the US. *, ** and *** indicate a statistically significant difference at the 10%, 5% and 1% level.
2 Countries sorted by the difference in expectations of advantaged and disadvantaged groups
3 Statistical significance calculated using a two sample t-test assuming independent samples are drawn between countries
4 Source: Author's calculations using PISA 2003 data. Country by country sample sizes in Table 4.1 and Appendix 4.1. Dependent variable in regression was whether respondent expected to obtain a degree (coded 0 if they did not and 1 if they did)

In Figure 4.6 I investigate whether this also remains the case when I include a school level fixed effect (results from 'model 3')[130]. Again, the distance each country sits above the 45 degree line illustrates how much of the remaining expectations gap is "explained" (in a statistical sense) by differences in school level factors (over and above the role schools play in developing children's age 15 test scores). Caution is, however, require when interpreting these results. In particular, I am unable to identify whether the reduction in the expectations gap is being driven by something that educational institutions actually provide (e.g. advantaged children attending good schools where teachers encourage and assist university applications) or if this is capturing some wider factor (e.g. peer effects or the community in which the child has grown up).

[130] Note that one faces certain difficulties when estimating such models. In particular, I must drop 4% of observations where all the children (or none of the children) in a particular school expect to complete university. This varies dramatically across countries. It is particularly high in countries where access to higher education is only available to children on a certain educational track. For instance, 25% of all observations are dropped in Germany, 20% in Austria and 19% in Switzerland compared to less than 1% in England, Scotland, Northern Ireland, Canada and the US.

Figure 4.6. Estimated difference between advantaged and disadvantaged children's plans to complete higher education, controlling for differences in age 15 test performance, before and after including a school level fixed effect



Notes:

1 Figures on the y-axis correspond to the difference between advantaged and disadvantaged children's educational expectations when I do not include a school level fixed effect (i.e these are the results from "Model 2" that were presented on the x-axis in Figure 4.5). Analogous figures, once I have included the school level fixed effect, can be found along the x-axis (i.e these are the results from model 3). Note that I have controlled for children's scores on the PISA maths tests in BOTH sets of results. The 45 degree line illustrates where including the fixed effect leads to no change in results.

2 Results are based upon predictions from the regression model that I describe on page 170. These are based upon measures of highest parental education, highest parental occupation and number of books in the home. Other controls include gender, immigrant status and an interaction between immigrant status and the four measures of advantage listed above.

3 Country names corresponding to abbreviations can be found in the first column of Table 4.1

4 Source: Author's calculations using PISA 2003 data. Country by country sample sizes in Table 4.1 and Appendix 4.1. Dependent variable in regression was whether respondent expected to obtain a degree (coded 0 if they did not and 1 if they did)

The estimated difference in log-odds has declined dramatically in some countries, but not in others. As suggested earlier, school level factors play a particularly prominent role when access to university is restricted to only those children on a certain educational track (e.g. Switzerland, Italy and Austria)[131]. Yet there is also a non-trivial change in England – the difference in log-odds has declined by over 40% (from 2.3 to 1.3). On the other hand, including the school level fixed effect has had only a more modest impact in Scotland, Northern Ireland and the US (a decline of less than 25%). Hence there is a tentative suggestion that schools, or at least something that is determined at the school level, may be a bigger part of the socio-economic gap in England than other Anglo-Saxon countries. However it is important to once again stress that, even after taking into account children's age 15 test performance and school level factors, large differences remain between advantaged and disadvantaged groups. For instance, in a hypothetical country where 50% of disadvantaged children believe they will complete university, roughly 70% of advantaged children would expect to attend (after accounting for differences in their test performance and schools). It is also worth noting that, after controlling for marks on the PISA assessment and school level factors, England sits *below* the OECD average. Moreover, the estimated difference in log-odds is now not significantly larger here than in any other country.

In summary, initial estimates suggested that the socio-economic gap may be slightly bigger in England (but not Scotland and Northern Ireland) than some other OECD countries. However, once differences in age 15 test scores have been taken into account, the evidence in support of this finding is reduced. Yet it is important to remember that children from disadvantaged backgrounds remain 30 percentage points less likely to expect entry into university than their advantaged peers who score the same on the aforementioned assessment (based on predictions from my model). Schools, or rather something defined at the school level, can explain a much more substantial proportion of this remaining difference in England (almost a half) than in Scotland, Northern Ireland or the US.

---

[131] However, one should be cautious when interpreting results for these nations because of the large number of observations that have been dropped from the analysis (due to the outcome being "completely determined" by the school). For instance, Austria and Switzerland stand out in Figure 4.7 because of their distance from the 45 degree line – yet I have dropped almost 20% of observations in these countries between these two models.

4.5. Results – The expectations of disadvantaged children who score highly on the PISA test compared to their advantaged, but lower scoring, peers.

I now turn to my final research question – do disadvantaged children who score highly on the age 15 PISA assessment hold higher expectations than their advantaged peers who do not perform so well? I face a problem, however, in that very few disadvantaged children (as per my definition in section 4.3) actually reach the upper tail of the age 15 maths test distribution. Hence I am in danger of addressing this question using an unreasonably small number of observations (i.e. there is an issue of "common support").

To try and overcome this difficulty, I use a broader definition of advantage based upon the PISA "Economic, Social and Cultural Status" (ESCS) index; a continuous measure of family background (scaled to be mean 0 and standard deviation 1 across the OECD countries) that is contained within the provided dataset. Specifically, the survey organisers have produced a weighted average, via a principal component analysis, of three variables (highest level of parental education, parental occupation, and availability of items in the family home) to generate a measure of children's socio-economic status. The first two of these variables (parental education and occupation) are as described in the earlier data section. The "availability of home possessions" is itself an index (from another principal components analysis) based upon children's reports of whether they have various items (e.g. computers, works of art, number of books) in their family home (details of which can be found in Appendix 4.2). According to OECD (2004), this provides an approximate measure of household wealth. Further details on the construction of the ESCS index are available in Appendix 4.2.

This index has several attractions as an alternative measure of family background. Firstly, it continues to capture the multi-dimensional nature (education, occupation, income/wealth) of "advantage". Secondly, as this variable is continuous, I can easily widen the proportion of children contained within my definition of advantaged and disadvantaged groups (to, for instance, the top and bottom quartile). Also note that, by using this measure, I can ensure that the same relative proportion of the population is defined as advantaged and disadvantaged in each of the OECD nations. Indeed, it is implicitly for these reasons that other authors, such as Chowdry et al (2009), have constructed similar measures via the same technique. Yet this variable also has a number of limitations. As it is created via a principal components analysis, it is somewhat difficult to interpret. There is also likely to be some information loss from suppressing various measures into one, all-encompassing, continuous index. One may also have some doubts over the validity of using household items as a measure of family wealth. Whether a child grows up in a home with a dishwasher or works of art will to some extent reflect parental preferences, and thus may provide little insight into whether they truly come from an advantaged or disadvantaged background. Similarly, one may question the cross-national comparability of such measures[132]. Yet, despite these limitations, this remains an attractive alternative measure of socio-economic status due to its flexible nature. The distribution of this index across OECD countries can be found in Table 4.7.

[132]Indeed, it is for these reasons that I did not use this variable or include such information in my initial definition of advantaged and disadvantaged groups (described in section 4.3).

Table 4.7. Distribution of ESCS measure of family background across countries

	Percentile							
	10th	**25th**	**50^{th}**	**75th**	**90^{th}**	**mean**	**SD**	**% Missing**
Mexico	-2.38	-1.78	-1.07	-0.09	0.67	-0.96	1.15	1
Turkey	-2.22	-1.76	-1.07	-0.29	0.59	-0.96	1.08	0
Portugal	-2.30	-1.63	-0.72	0.23	1.18	-0.65	1.27	1
Poland	-1.17	-0.79	-0.31	0.25	1.02	-0.21	0.82	0
Greece	-1.45	-0.91	-0.25	0.50	1.20	-0.19	1.00	0
Spain	-1.53	-0.88	-0.19	0.51	1.13	-0.19	0.98	1
Korea	-1.24	-0.66	-0.08	0.47	0.96	-0.12	0.85	0
Switzerland	-1.15	-0.64	-0.11	0.45	0.99	-0.10	0.83	1
Hungary	-1.12	-0.71	-0.18	0.51	1.18	-0.09	0.89	0
Japan	-0.98	-0.59	-0.10	0.43	0.94	-0.08	0.73	1
Ireland	-1.17	-0.68	-0.09	0.52	1.09	-0.07	0.88	1
France	-1.18	-0.69	-0.05	0.59	1.13	-0.05	0.92	1
Italy	-1.29	-0.73	-0.03	0.63	1.32	-0.02	0.97	0
Slovakia	-0.96	-0.60	-0.12	0.54	1.13	-0.02	0.81	0
Northern Ireland	-1.09	-0.69	-0.08	0.65	1.28	0.01	0.91	3
Austria	-0.95	-0.48	0.02	0.65	1.26	0.08	0.85	1
Scotland	-1.05	-0.50	0.10	0.76	1.26	0.09	0.91	2
Netherlands	-0.94	-0.47	0.10	0.76	1.20	0.11	0.86	3
Belgium(French)	-1.13	-0.54	0.17	0.84	1.38	0.13	0.98	2
England	-0.95	-0.47	0.09	0.79	1.38	0.14	0.89	5
Belgium(Flemish)	-0.92	-0.46	0.17	0.86	1.35	0.18	0.91	3
Germany	-0.94	-0.42	0.16	0.86	1.50	0.18	0.99	6
Denmark	-0.85	-0.39	0.19	0.79	1.31	0.19	0.85	1
Luxembourg	-1.40	-0.48	0.32	0.98	1.51	0.19	1.09	1
Australia	-0.83	-0.34	0.24	0.83	1.26	0.23	0.83	1
New Zealand	-0.91	-0.33	0.27	0.86	1.35	0.23	0.91	2
Czech Republic	-0.76	-0.33	0.17	0.87	1.39	0.26	0.81	3
Finland	-0.80	-0.34	0.29	0.87	1.32	0.26	0.83	0
Sweden	-0.80	-0.31	0.26	0.90	1.37	0.26	0.87	1
USA	-0.87	-0.33	0.29	0.93	1.41	0.28	0.90	1
Canada	-0.72	-0.25	0.33	0.97	1.50	0.35	0.84	5
Norway	-0.36	0.06	0.59	1.18	1.64	0.61	0.79	1
Iceland	-0.40	0.13	0.74	1.34	1.77	0.69	0.88	1
OECD	**-1.44**	**-0.71**	**-0.02**	**0.68**	**1.26**	**0**	**1**	**2**

Notes:
1 Data refers to points on the ESCS scale of family background, as described earlier in this section. On this scale, higher values means a more advantaged family background.
2 Countries sorted by mean values.
3 Source: PISA 2003 data. Country by country sample sizes in Table 4.1 and Appendix 4.1.

I proceed by dividing this variable in four equal groups (separately for each country) and defining:

'Advantaged' = top quartile group of the national ESCS distribution
'Disadvantaged' = bottom quartile group of the national ESCS distribution

I then use this information in a fourth regression model as a set of dummy variables, with the bottom quartile ('disadvantaged') as the reference group. Formally, this model is specified:

Model 4: $$\log\left(\frac{\Pi(\boldsymbol{E}_{ij})}{1-\Pi(\boldsymbol{E}_{ij})}\right)=\boldsymbol{\alpha}_1+\boldsymbol{\beta}_1.\boldsymbol{Sex}_i+\boldsymbol{\beta}_2.\boldsymbol{I}_i+\boldsymbol{\beta}_3.\boldsymbol{SES}_i+\boldsymbol{\beta}_4.\boldsymbol{I}_i*\boldsymbol{SES}_i+\boldsymbol{\beta}_5.\boldsymbol{Ab}_i$$

where:
Adv = A vector of three dummy variables reflecting advantage, based upon quartiles the ESCS measure of family background described above (reference = bottom quartile)

Ability = A vector of four dummy variables based upon quintiles of the (national) age 15 maths test distribution (reference = bottom quintile group)

All other variables are as described in section 4.3[133].

Note that "ability" (defined as performance on the PISA test at age 15) also enters this model as a set of dummy variables, reflecting quintiles of the (national) PISA maths test distribution (with the lowest quintile group as the reference group). I did this to allow for a straightforward interaction between advantage and test performance while retaining non-linearity in the latter. However, I chose not to include these interaction terms in the final specification as a likelihood ratio test (for the inclusion of these extra parameters) suggested they were statistically *insignificant* at the 10% level in 29 of the 33 countries[134,135]. In Appendix 4.3 I check the robustness of my results when

[133] Note that I do not control for school level factors in this specification as my interest is in the raw differences between groups, and not why such patterns occur. I thus allow for the complex survey design used in PISA by making appropriate adjustments to the estimated standard errors.
[134] The countries where the likelihood ratio test was significant were Mexico, Canada, Northern Ireland

using an alternative specification, in which I include PISA test scores as a continuous term and interact this with an alternative measure of family background.

From this model, I compare the predicted log odds of a high scoring, disadvantaged child (inside the top 20% of the maths test distribution and within the bottom 25% of the ESCS distribution) expecting to complete university to the predicted log odds for an advantaged, child scoring around the national average (inside the 40^{th} – 60^{th} percentile of the maths distribution and within the top 25% of ESCS distribution). Table 4.8 illustrates that there are at least 50 children in every country fitting each of these criterion, suggesting there is adequate data to support such estimations. It also provides further motivation for this research; in most countries less than one in ten disadvantaged children reach the top quintile of the age 15 maths test distribution, compared to over one third of those from the most advantaged homes. To broaden the social composition of universities, the few disadvantaged teenagers who manage to overcome adversity up to age 15 must believe higher education is a realistic and obtainable goal.

---

and France. However, one would expect to find a significant difference (at the 10% level) in three countries purely by chance.

[135] Wald tests of individual parameters of interest (e.g. the interaction of the top ability quintile with the top ESCS quartile) were also largely insignificant. The coefficients on these interactions were positive in roughly half of the countries and negative in the others.

Table 4.8. Percentage of 15 year olds from the top and bottom quartile group of the ESCS distribution who are in the top quintile group of the national PISA math's test distribution

	Number of observations defined as 'advantaged'/ 'disadvantaged'	**Number of disadvantaged children who reach top quintile group of maths test distribution**	**% of disadvantaged children who reach top quintile group of maths test distribution**	**% of advantaged children who reach top quintile group of maths test distribution**
Belgium(French)	740	46	6	38
Northern Ireland	715	46	6	37
Hungary	1,192	49	4	43
Scotland	681	50	7	36
England	999	58	6	39
New Zealand	1,129	61	5	35
Ireland	972	66	7	34
Luxembourg	981	66	7	36
France	1,075	68	6	36
Denmark	1,057	70	7	35
Norway	1,016	70	7	35
Poland	1,099	70	6	38
Germany	1,166	72	6	33
Netherlands	1,007	81	8	34
Czech Republic	1,582	82	5	36
Greece	1,159	83	7	38
Japan	1,201	86	7	34
Portugal	1,152	86	7	41
Iceland	844	88	10	31
USA	1,364	90	7	37
Sweden	1,165	91	8	37
Austria	1,198	97	8	33
Slovakia	1,845	101	5	39
Korea	1,366	105	8	36
Turkey	1,214	107	9	40
Belgium(Flemish)	1,460	108	7	34
Finland	1,452	132	9	33
Switzerland	2,109	159	8	34
Australia	3,145	229	7	35
Spain	2,698	254	9	35
Italy	2,918	296	10	29
Mexico	7,526	657	9	34
Canada	6,989	687	10	31
OECD	**56,216**	**4,411**	**8**	**35**

Notes:

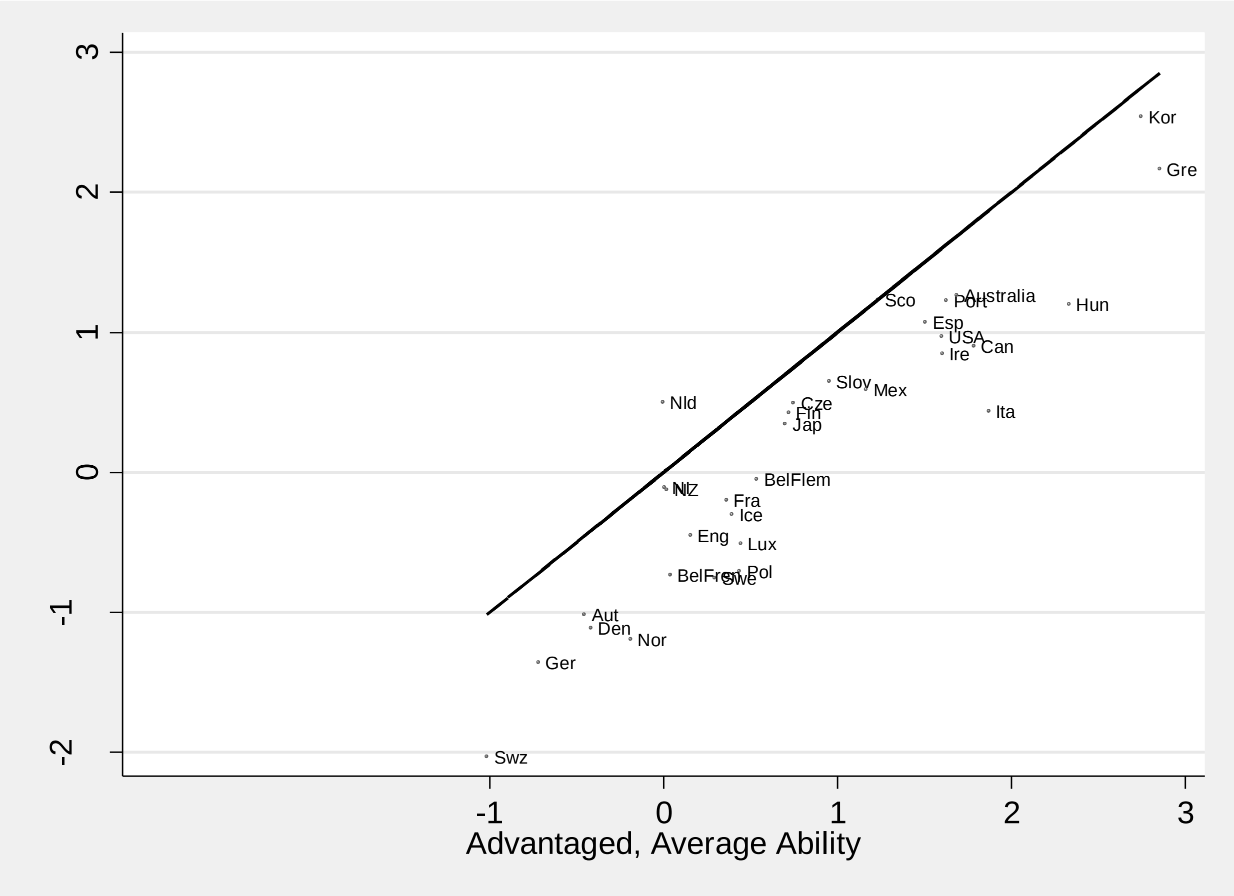
1 Advantaged children in this table are defined as those within the top quartile group of the national ESCS distribution described earlier in this section. Similarly, disadvantaged children are those in the bottom quartile group of the national ESCS distribution. The second column from the left illustrates the number of observations within these groups. The two columns on the right give the proportion of advantaged or disadvantaged children who reach the top quintile group of the maths test distribution.

2 Source: PISA 2003 data. Country by country sample sizes in Table 4.1 and Appendix 4.1.

Results are presented in Figure 4.7 (with the full set of parameter estimates provided in Appendix 4.1). The x-axis illustrates the predicted log odds of an *advantaged*, native girl scoring the national *average* mark on the PISA maths test (scoring between the 40^{th} and 60^{th} percentile) expecting to complete university. On the other hand, the y-axis demonstrates the equivalent figures for a *disadvantaged*, native girl scoring in the top 20% of the age 15 national PISA test distribution. The 45 degree line is where these children are equally likely to expect to obtain a bachelor's degree. If a country sits below this line, then children who score highly on cognitive assessment at age 15 from disadvantaged backgrounds are *less* likely to expect completion of university than their advantaged peers who score around the national average.

In almost every country, it seems that scoring highly on the PISA test (defined in this way) does not overcome the constraints of family background. 32 of the 33 countries considered sit below the 45 degree line (the Netherlands is the exception), with the difference being statistically significant at the 5% level on 28 occasions. Hence, it seems that children from disadvantaged backgrounds who score highly on age 15 cognitive assessments do not hold higher expectations than their lower scoring, but more fortunate, peers. Indeed, in a number of countries, these children are *less* likely to believe they have what it takes to successfully complete higher education.

Figure 4.7. Predicted log-odds of a *high* scoring disadvantaged native girl expecting to complete university versus an average scoring advantaged native girl



Notes:
1 Figures on the y-axis illustrate the predicted log-odds of a disadvantaged, high scoring native girl expecting to complete university. Similarly, the predicted log-odds for an advantaged native girl of average ability are presented on the x-axis. The 45 degree line represents where the predicted log-odds for the two groups described above are equal (they hold the same chance of expecting to complete university).

2 Points sit significantly below the 45 degree line in all countries at the 5% level except in Northern Ireland, Scotland, New Zealand and the Netherlands (which is above the line and significant).
I define advantage using the ESCS index described earlier in this section. Specifically, those in the top quartile group of this index were defined as advantaged, while those in the bottom quartile group were disadvantaged. High ability is defined as those in the top 20% of the national PISA math's test distribution. Average ability is defined as those in the 40^{th}-60^{th} percentile.

3 Variables in the model include the ESCS measure of family background, immigrant status, gender, children's score on the PISA maths tests and an interaction between immigrant status and ESCS. Country names corresponding to abbreviations can be found in the first column of Table 4.1

4 Source: Author's calculations using PISA 2003 data. Country by country sample sizes in Table 4.1 and Appendix 4.1. Dependent variable in regression was whether respondent expected to obtain a degree (coded 0 if they did not and 1 if they did)

England, the US and Canada are all part of this aforementioned group; each sits significantly below the 45 degree line at the 1% level. Yet Table 4.9 suggests that these countries do not particularly stand out from the rest of the OECD. The difference in both England and the US is around the OECD average, with little evidence that the expectations gap is any bigger or smaller here than other parts of the developed world. On the other hand, Scotland and Northern Ireland sit very near the 45 degree line and are two of the four countries where one can not reject the null hypothesis that high scoring disadvantaged children are equally as likely to expect to attend university as their advantaged but lower scoring peers. Indeed, Table 4.9 illustrates that the socio-economic gap is significantly weaker in Scotland than 19 other countries (including England, Canada and the US).

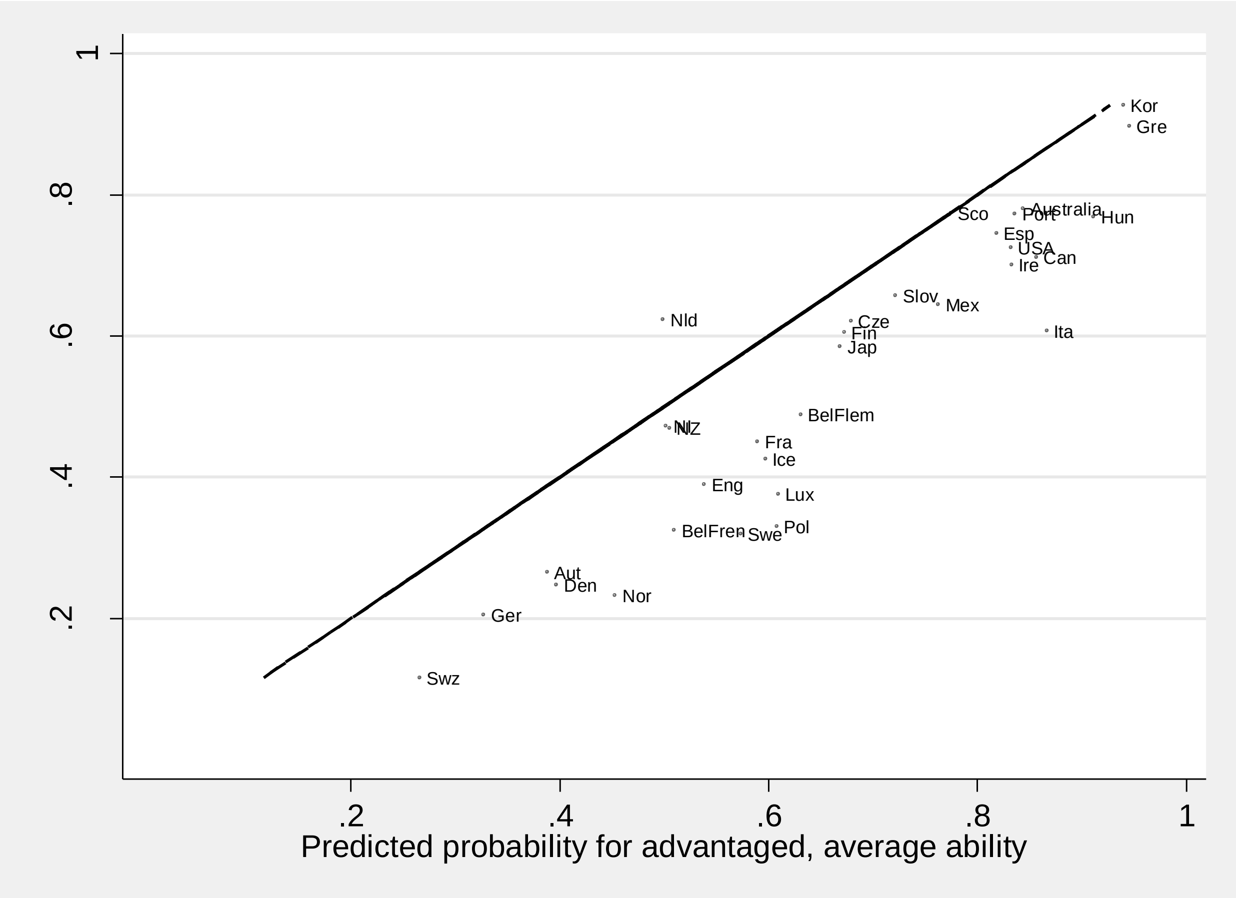
Nevertheless, the above results may still be a concern for British and American policymakers. This is perhaps more easily seen in Figure 4.8, where I convert the estimates from Figure 4.7 into predicted probabilities. In the US and Canada, an advantaged child of achieving the national average mark on the PISA maths assessment has a 85% chance of expecting to complete university. On the other hand, the probability for a high scoring child from a less fortunate background is closer to 75%. In England, the analogous figures are 55% and 40%. Thus a large socio-economic divide remains even when children from less fortunate families manage to surpass their more affluent peers academically. As stated in Section 4.2, most would argue that our best and brightest children should be aiming to obtain a degree, regardless of their family background. It should thus be a real concern in England that, of the few disadvantaged 15 year olds that manage to reach the top quintile of the maths test distribution at this age, only one in four believe that they will go on to obtain a tertiary qualification.

Table 4.9. Difference between the expectations of disadvantaged children scoring a high mark on the PISA maths assessment versus advantaged children with a mark around the national average

		'Difference' significantly greater/smaller than...				
Country	**Difference between advantaged and disadvantaged (in log odds)**	**Scotland**	**NI**	**England**	**US**	**Can**
Netherlands	-0.51	-	***	***	***	***
Scotland	0.00	-	-	**	**	***
Northern Ireland	0.12	-	-	-	-	***
New Zealand	0.14	-	-	**	*	***
Korea	0.21	-	-	-	-	**
Czech Republic	0.25	-	-	-	-	***
Finland	0.29	-	-	-	-	***
Slovakia	0.30	-	-	-	-	***
Japan	0.36	-	-	-	-	**
Portugal	0.40	-	-	-	-	**
Australia	0.42	-	-	-	-	**
Spain	0.43	-	-	-	-	**
France	0.56	-	-	-	-	-
Austria	0.56	*	-	-	-	-
Mexico	0.57	*	-	-	-	-
Belgium (Flemish)	0.58	*	-	-	-	-
England	0.60	**	-	-	-	-
USA	0.63	**	-	-	-	-
Germany	0.64	*	-	-	-	-
Greece	0.69	**	-	-	-	-
Iceland	0.69	**	*	-	-	-
Denmark	0.70	**	*	-	-	-
Ireland	0.75	**	*	-	-	-
Belgium (French)	0.77	**	*	-	-	-
Canada	0.88	***	***	-	-	-
Luxemburg	0.95	**	**	-	-	-
Norway	1.01	***	***	-	-	-
Switzerland	1.01	***	**	-	-	-
Sweden	1.05	***	***	*	-	-
Hungary	1.13	***	***	**	*	-
Poland	1.14	***	***	***	**	-
Italy	1.43	***	***	***	***	**

Notes:1 The final five columns illustrate whether the estimated log odds are significantly different to those in England, Scotland, Northern Ireland, Canada and the US. *, ** and *** indicate a statistically significant difference at the 10%, 5% and 1% level.
2 Data sorted by the difference in expectations of advantaged and disadvantaged groups
Statistical significance calculated using a two sample t-test assuming independent samples
3 Source: Author's calculations using PISA 2003 data. Country by country sample sizes in Table 4.1 and Appendix 4.1. Dependent variable in regression was whether respondent expected to obtain a degree (coded 0 if they did not and 1 if they did)

Figure 4.8. Predicted probability of a *high* scoring disadvantaged native girl expecting to complete university versus an average scoring advantaged native girl



Notes:
1 This figure presents the results given in Figure 4.7 in terms of predicted probabilities. See notes to Figure 4.7 above for further details.
2 Source: Author's calculations using PISA 2003 data. Country by country sample sizes in Table 4.1 and Appendix 4.1. Dependent variable in regression was whether respondent expected to obtain a degree (coded 0 if they did not and 1 if they did)

4.6. Summary and policy discussion for the UK

In all countries across the developed world, there are large and statistically significant differences between the educational expectations of children from different family backgrounds. This remains the case even after taking into account the fact that such groups attend different schools and achieves different marks on tests at age 15. A particularly important finding is that high achieving children from less fortunate homes hold lower expectations than their affluent but lower performing peers. A central theme amongst the research questions I set out in section 4.2 was whether such results are particularly prominent in the UK. Analysis of the PISA data has yielded little evidence that this is the case. Although the initial estimates suggested that

England sat near the top of the international ranking (third largest difference), one could not reject the hypothesis that a similarly large gap occurs in several other English speaking countries like Australia, New Zealand, Scotland and Northern Ireland. In fact, after controlling for differences in age 15 test scores, the advantaged-disadvantaged gap was only significantly bigger in England than in six of the 33 OECD nations (recalling once more my discussion of how to interpret this finding and the possibility of type two errors). Once differences in school level factors were also taken into account, England stood below with the international average. A similar finding occurs when considering the expectations of English children from disadvantaged homes who scored highly on the age 15 PISA maths test. Although this group is significantly less likely to believe that they will complete higher education than those of lower scoring children from more affluent homes, this does not stand out as an issue that is specific to the UK.

Aspects of these results conflict with evidence from the existing literature. For instance, Chevalier et al (2009) use the PISA 2003 sample for England and find that there are no social class differences in English children's educational expectations once one has taken into account differences in the PISA test scores[136]. However, the authors investigate differences based *only* upon parental occupation, and have presented conditional estimates after controlling for several related factors (e.g. parental education and number of hours studying per week). Hence one possible reason for this apparent inconsistency is my focus on a broad, multi-dimensional measure of advantage rather than their specific interest on occupation. In accordance with my findings, Chowdry et al (2009) also find large socio-economic differences in 14 and 16 year olds university plans when analysing the Longitudinal Study of Young People in England survey[137]. Although the authors do not explicitly explore the determinants of these educational intentions, there is an implicit suggestion that this is not entirely reflecting differences in academic attainment (as I also suggest in this chapter).

[136] Interestingly, they also find that although children's academic perception is related to their social class, it does explain difference in these groups attainment.
[137] Interestingly, they find that a lot higher proportion of English children expecting to attend university than in my analysis. For instance, they find half of the most disadvantaged quintile believe they are likely to apply to university and be accepted. One reason may be that the question used is very different in the PISA and LSYPE surveys, with the former perhaps better suited towards capturing children's expectations and the latter their aspirations.

A final consideration is what can be implemented, in a cost-effective manner, to address the large socio-economic gaps I have highlighted in this chapter. Emmerson et al (2005) suggest that widening access schemes like "Aim Higher" have had some success in increasing disadvantaged children's intentions of completing higher education. However, given the limited resources currently available, one must decide the age group at which it is most efficient to target such initiatives. Chowdry et al (2008) suggest moving widening participation schemes away from those enrolled in post-compulsory education (ages 16 to 18) to children in the earlier stages of secondary school. Given that I find such large socio-economic gaps in children's perceptions of higher education at age 15, I am inclined to agree with these authors. Rather than focusing on those in post 16 education, widening access initiatives to raise expectations should perhaps be targeted at younger age groups (as children are turning 14) to try and stop disadvantaged children forming negative pre-conceptions about university.

4.7. Conclusion

There is a concern in many countries that children from disadvantaged backgrounds are under-represented amongst the undergraduate population. In particular, policy-makers are worried that some young adults who could benefit from higher education decide not to seek out the returns that this investment can offer. "Widening access" schemes that try to address this issue have thus become common across the developed world. A particular feature of such programmes in the US and UK is that they explicitly aim to raise disadvantaged children's expectations of completing university. It is claimed that, in doing so, such policies will reduce the socio-economic divide in tertiary graduation rates.

By considering the size of this expectation gap in the US and UK, and how it compares to other developed countries, I have made a number of contributions to the existing literature and current policy debate. Firstly, this is the only study (that I am aware of) to place the socio-economic divide in British and American children's educational expectations into an internationally comparative perspective. Secondly, this is the first study, to my knowledge, to consider whether the gap between advantaged and disadvantaged children's educational expectations can be fully explained by differences in their test scores at age 15. Finally, other studies within this literature have typically focused upon the educational intentions of the “average” child. In contrast, I explicitly consider whether young adults from disadvantaged backgrounds who score highly on the age 15 PISA maths assessment hold higher or lower expectations than their advantaged, but lower scoring, peers.

My results show that there are large differences between advantaged and disadvantaged children's educational plans, and that this holds true across all countries in the developed world. There is little evidence that the UK stands out when compared to other members of the OECD, particularly once differences in age 15 test performance and school level factors have been taken into account. On the other hand, there is some evidence that the socio-economic gap is atypically small in the US. However, perhaps my most concerning finding is that a significant proportion of 15 year olds who score high marks on the PISA assessment do not expect to complete university. In both England and the US, an advantaged child who scores the national average mark is *more* likely to expect to complete higher education than a disadvantaged child who reaches the top quintile of the test distribution. Indeed, I predict that from my regression model that little over a third of high-performing disadvantaged children in England believe they will obtain a bachelor’s degree if they come from a disadvantaged family background. Yet one should bear in mind this is not a unique British trait, and that similar patterns can be found across the developed world.

It is of course important to note the limitations of this study and, in doing so, aspects of the wider literature. To begin, I remind the reader that the estimates presented should not be treated as causal. In particular, it is again worth pointing out that I have not explored differences in expectations between children of the same “natural” or inherent talent. Rather, I have only been able to investigate differences in expectations conditional on test performance at age 15 (which is, in itself, likely to reflect differences in socio-economic background). On a related issue, I have undertaken this research based on the assumption that adolescents' educational expectations have an important influence on their later behaviour and schooling attainment. Although there is evidence supporting this from a broad range of disciplines, including sociology, social psychology and economics, further work in this area still needs to be done. In particular, future research should focus on untangling the relationship between these variables and whether such associations vary across different national settings. Finally, all my analyses and subsequent inferences are based on the assumption that children are reporting their educational *expectations,* rather than their *aspirations,* in the PISA survey. Although general patterns within the data are consistent with this view, formal validation of this type of question would represent a significant forward step for the wider literature.

Despite these caveats, this research should make an important contribution to debates on widening university participation in the UK. Given the large socio-economic gap in English children's expectations, I suggest that policymakers should continue to promote the benefits of higher education to disadvantaged groups (particularly those with the greatest academic potential), though perhaps targeting slightly younger age groups. Of course, within the current financial climate, it is also important that such schemes represent an efficient allocation of finite government resources. Recent work by the Sutton Trust (2010) vehemently argues that this is the case. Yet it also seems prudent to stress the need for future research which thoroughly investigates the assumptions on which such policies are based. In particular, social scientists need to confirm that raising adolescents' educational expectations has a *causal* influence on later attainment (and possibly other behaviour as per Cowan 2009). The main conclusion that this chapter draws is that, although the socio-economic gap in children's educational expectations is large in all countries within the UK, it is not atypically big with respect to other developed nations. Nevertheless, the low expectations of children from disadvantaged backgrounds remain one plausible reason why our undergraduate population has such an unbalanced social mix.

Chapter 5

Conclusions

This thesis began by explaining the role of young adults' expectations in social scientists' models of schooling choice. Firstly, I considered the approach favoured by economists, where adolescents' educational decisions are driven by what they *expect* the costs and benefits of this investment to be. In doing so, I made clear my concern that information on children's expectations is not routinely collected within this discipline. Likewise, I noted the worries of others, such as Manski (1993, 2004), that traditional economic models of schooling behaviour rely heavily on the assumption that there are no systematic differences between young adults' (ex-ante) expected returns to education and later (ex-post) realisations. Using these concerns as motivation, this thesis has explored the variation in young adults' labour market expectations, and whether such beliefs about the future are (on average) realistic. In chapter 2 I considered this in reference to undergraduates' expected starting salaries within the UK. This was followed in chapter 3 by an analysis of whether 20 year old American men were able to make realistic predictions of their labour market outcomes at age 30. On the basis of this research, I have reached the following seven conclusions:

- The wage expectations of UK students vary significantly with how far they have progressed through university, their views on the graduate labour market and quality of institution they attend. On the other hand, they are not strongly associated with ethnicity.

- Part-time students in the UK are, on average, quite realistic about their starting wage. On the other hand, those studying full-time tend to overestimate their starting salary by an average of just under 15% (roughly £2,000). This result contrasts with Webbink and Hartog (2004), the only other European study using a balanced sample of students, who found that young adults in the Netherlands could make reasonable predictions of their first wage.

- The accuracy of UK students' wage expectations depend, however, on the subject they are studying. Whereas those who are enrolled in Maths, Engineering, Computer Science and Education courses hold, on average, reasonably realistic expectations, their peers studying for an Art, Business or Humanities qualification make particularly poor predictions (overestimating their starting salary by an average of almost 20%).

- Many of these results also hold over a longer time horizon for young men in the US, who overestimate their income 10 years into the future by (on average) over 40%. Consistent with my findings for the UK, those studying Mathematics, Engineering and Physical Sciences make better predictions than their peers completing an Art, Humanities or Social Science degree.

- There is, however, little evidence that young American men in employment are more realistic about the labour market than their peers who are enrolled in higher education. Both groups vastly overestimate their future salary and chances of entering a professional career.

- On the basis of such results, I suggest the simplistic assumptions that are often made by economists (that young adults' expectations do not systematically differ from their later realizations) are based on a rocky foundation. In particular, it seems that young adults, both those enrolled in higher education and their peers in employment, maybe lacking some important information on the labour market.

- Consequently, I advise US and UK policymakers to increase the availability of data on graduate and non-graduate wages. Specifically, information on graduate employment outcomes and wages should be set out separately by discipline and institution. Moreover, although I have focused on average outcomes in this thesis, one may argue that this should be in the form of entire wage *distributions* (given the central role of wage variability and risk aversion in economic models of choice). Likewise, such wage profiles should be made available for those undertaking alternative training routes (e.g. apprenticeships). Ideally, this should be accompanied by estimates of counterfactual outcomes under a set of alternative scenarios – for instance what a graduate would have earned if he/she chose to undertake an apprenticeship instead.

In the penultimate chapter, I turned to the role of children's *educational* expectations in determining their later schooling behaviour. Specifically, I discussed how educational expectations may differ between advantaged and disadvantaged groups, and how this may lead to different rates of educational attainment. I thus explored the gap between advantaged and disadvantaged children's educational plans, focusing on whether such differences are greater in the US and UK than other parts of the developed world. This has brought me to the following three additional conclusions:

- There are large and statistically significant differences between the educational expectations of advantaged and disadvantaged children. In fact, high achieving children from less fortunate homes hold lower expectations than their affluent but lower achieving peers. This holds true across almost every country in the developed world.

- Yet there is little evidence that the difference between advantaged and disadvantaged children's educational expectations stands out as particularly big in the UK (compared to other members of the OECD). On the other hand, there is some evidence that this socio-economic gap is atypically small in the US.

- I thus suggest that, although widening access schemes should still continue to be a prominent part of UK higher education policy, such initiatives should be of no more concern here than in other parts of the developed world.

Although these conclusions have partly addressed my concerns with the current literature, set out in chapter one, there are limitations to this work and (consequently) the results that I have found. Three particular limitations that I wish to highlight are:

- The lack of information on counter-factual outcomes. In section 1.2 I set out an economic framework of the HE decision making process. This model stipulates that a young person will be more likely to go to university the higher their expected economic <u>return</u> (the wages they expect to get if they obtain a degree versus the wages they expect if they don't). However, in chapters 2 and 3 I only analysed expected and actual wages conditional upon obtaining a degree – I do not have any information on students' views of outcomes under the counterfactual (not obtaining a degree). It is important to consider the limitations that this puts on the approach taken in this thesis. It is possible that, although young people overestimate their wage post-university by a certain percentage, they may also overestimate outcomes under the counterfactual (e.g. not going to university) by the same amount. In this situation their expected return to higher education (and hence their decision of whether to go to university) would remain completely unchanged. More generally, over-estimation of post-university wages is only important in the HE decision making process if it takes the expected <u>return</u> to that choice (e.g. going to university) above that of the counterfactual (e.g. not going to university). Hence young people can still rationally decide to enter higher education even when overestimating their future wage. Without information on young peoples' views of outcomes under the counterfactual (e.g. what they would earn had they not gone to university) I am unable to make a firm statement as to whether this group are just generally over-optimistic about the future (i.e. over-estimate all outcomes) or if their mistaken views on post-university wages is actually altering the educational choices that they make.

- One of the main findings in this thesis is that students in certain subjects (mainly those in Maths and Science based courses) make better predictions of future employment outcomes than others (mainly Social Science, Language and Humanities degrees). However, as noted in chapter 3, some care needs to be taken when drawing this interpretation from my results. An alternative explanation for this finding is that most students do not have any idea what they will be earning in the future (particularly when looking over a relatively long time horizon). Consequently, all higher education students may be reporting only some rough idea of an average graduate salary, which may not vary much by the subject they study. When expectations are then compared to realisations, it seems that students in some subjects (e.g. Maths, Engineering etc) make more accurate predictions than others. But the only reason for this is that they happen to be in a discipline that leads to high earning jobs (i.e. it is not so much to do with them having "more realistic" expectations, but rather that they just happen to be enrolled in a subject that leads to high later earnings). Hence the greater "accuracy" in their expectations is reflecting the fact that all students anchor their expectations around the same point – but that some groups tend to earn more than others. As shown in Table 2.12, Table A3.14 and discussed in chapter 3, there is reasonable variation in average wage expectations by subject studied – suggesting that the above interpretation is not the sole factor driving my results. Nevertheless, I can not rule the above out as at least a partial reason for why I find students in certain subjects making better predictions than others.

- A final limitation that I wish to highlight is that, for the majority of this thesis, I have worked within the constraints of human capital theory and a mainly economic paradigm (though with some attempt to integrate sociological and social psychological views in the penultimate chapter). As set out in section 1.3, there are a number of valuable models and theories from other disciplines that attempt to explain the educational decision making process, in which wage expectations play only a minor (if any) role. A fine example is the Breen and Goldthorpe (1997) model described in the introduction to this thesis. In this set-up, it is not wages and economic returns that drove young peoples' decision to enter higher education, but their relative risk aversion and the desire to not suffer downward mobility (i.e. to minimise their risk of entering a lower social position than that of their parents). As noted by Davies (2002), this is also a "rational action theory", but where young people are assumed to optimise something other than their future wages and economic returns. Future work should take such alternative perspectives into account. There seems, in particular, much to be gained by greater integration and interaction amongst the economic and sociological literatures. The work of Morgan (2005) is one explicit example that attempts such a synthesis of human capital theory, relative risk aversion and status attainment research. This needs to be built upon for our understanding of educational behaviour to progress.

On the basis of my finding and the limitations cited above, I have identified four areas that I believe are a priority for future research:

1. Developing empirical models of schooling choice that do not rely on data collected "ex-post"

I opened this work by stating my scepticism over economists' use of data collected after schooling decisions have been made ("ex-post") to make inferences about the decision making process itself; and thus the assumption that young adults' ex-ante expectations do not systematically differ to their later ex-post realisations. Yet there are, to my knowledge, very few examples of schooling choice models that relax these constraints and use data on young adults' subjective expectations as an alternative[138]. This must change in order for the literature on young adults' schooling choice to progress.

2. Better measurement, and further validation, of expectation data

On a number of occasions, I have raised concerns about the quality of data available. This includes whether such information actually captures individuals' expectations rather than their aspirations, whether respondents account for inflation in their estimates and the bunching of observations at particular points on the distribution. Manski (2004) has recently devised a range of methods, whereby respondents make probabilistic assessment of possible future events, to try and resolve such issues[139]. Yet these techniques are still in their infancy, and are rarely found within British or European surveys. Likewise, these methods have been the subject of little independent validation, with most evidence on this issue having been presented by Manski himself. Further testing and development of these methods is fundamental for

[138] Although, with the emergence of recent working papers in the US (Kaufmann and Attanasio 2009, Arcidiacono et al 2010), there is evidence that this may be changing, such advances, to my knowledge, have not been made in Europe.
[139] Though, to my knowledge, there are no examples that use these methods within educational research in the UK or Europe.

this literature to develop.

3. Further evidence of a causal relationship between adolescents' educational expectations and later outcomes

As suggested within chapter four, the evidence that educational expectations have a causal impact on later attainment is not yet conclusive. This issue of causality needs to be addressed before further policy initiatives to 'raise children's expectations' are introduced. Further developing the theory and empirical methods set out by Morgan (2004) seems like an ideal place to start. Likewise, establishing how the strength of association between expectations and outcomes varies across countries, and whether there is a causal association in some nations but not others, would be of great interest to all those actively involved in research on social mobility and university access.

4. Collecting longitudinal data on educational and financial expectations in order to investigate how children's views develop over time.

As economists and sociologists believe that expectations influence behaviour, both groups should have shown great interest in how such beliefs develop. Yet empirical evidence on this issue, particularly within economics, is scant. Collection of large-scale, longitudinal data is thus required to provide a detailed insight into how children and young adults make plans for the future. This would enable researchers to identify the stability of expectations over time, how they are related to progress at school and whether adolescents' views are actually altered by policy initiatives such as AIMHIGHER (as per Emmerson et al (2005)). One could also thoroughly investigate the impact of careers advice on young adults' expectations and how this, in turn, influences the educational decisions they make. Likewise, further information on young adults' wage expectations should be collected, with a view over longer time horizons and under various counterfactual options. Both would help economists to understand whether the assumptions often used to identify their models are correct. Such insights also highlight why future research into children's expectations should not only be a concern of academics, but also stimulate thought and debate within policy environments.

There is, evidently, still much ground to be covered in developing this literature. Indeed, despite almost half a century of the social science theory stressing the importance of children's expectations on their later behaviour and attainment, empirical evidence on this matter remains limited. This thesis has taken a step towards resolving some of the issues that this literature has faced. In doing so, I have attempted to open up several avenues for future work that may bring about a better understanding of adolescents' educational behaviour. Indeed, the next decade is likely to see further advancement in this area, to which this thesis will hopefully provide some insight.

References

Ammermueller A. & Pischke J. (2009) “Peer Effects in European Primary Schools: Evidence from the Progress in International Reading Literacy Study”, Journal of Labor Economics, volume 27, p 315-48

Angrist, J. & Krueger, A. (1991) “Does Compulsory Schooling Attendance Affect Schooling and Earnings?”, Quarterly Journal of Economics, volume 106, p 979-1014

Arcidiacono, P. Hotz, J. & Kang, S. (2010) "Modeling College Major Choices Using Elicited Measures of Expectations and Counterfactuals", IZA Discussion Papers 4738

Baird, C. Burge, S. & Reynolds, J. (2008) “Absurdly Ambitious? Teenagers' Expectations for the Future and Realities of Social Structure”, Sociology Compass, volume 2, p 944-962

Becker, G. (1993) “Human Capital: A Theoretical and Empirical Analysis with Special Reference to Education”, Third Edition, University of Chicago Press

Berger, M. (1988) “Predicted Future Earnings and Choice of College Major”, Industrial and Labor Relations Review, volume 41, p 418-429

Betts, J. (1996) “What Do Students Know About Wages? Evidence from a Survey of Undergraduates”, Journal of Human Resources, volume 31, p 27-56

Black, D. Sanders, S. & Taylor, L. (2003) “The Economic Reward for Studying Economics”, Economic Inquiry, volume 41, p 365-377

Blanden, J. & Machin, S. (2004) “Educational Inequality and the Expansion of UK Higher Education”, Scottish Journal of Political Economy, Special Issue on the Economics of Education, volume 51, p 230–49

Blau, F. & Ferber, M. (1991) “Career Plans and Expectations of Young Women and Men: The Earnings Gap and Labor Force Participation” Journal of Human Resources, volume 26, p 581-607

Borjas, G. (1999) “The Economic Analysis of Immigration”, Handbook of Labor Economics, p 1697-1760

Bosworth, L. & Ford, J. (1985) “Income Expectations and the Decision to Enter Higher Education”, Studies in Higher Education, volume 10, p 21–32

Botelho, A. & Pinto, L. (2004) “Students’ Expectations of the Economic Returns to College Education: Results of a Controlled Experiment”, Economics of Education Review, volume 23, p 645–653

Bound, J. Brown, C. & Mathiowetz, N. (2001) “Measurement Error in Survey Data”, in Heckman, J. & Leamer , E. (ed.) Handbook of Econometrics, edition 1, volume 5, chapter 59, p 3705-3843

Boudarbat, B. (2004) “Earnings and Community College Field of Study Choice in Canada”, IZA Discussion Papers 1156

Breen, R. (2001) “A Rational Choice Model of Educational Inequality”, Institutojuan March, Madrid, Working Paper number 2001/166

Breen, R. & Goldthorpe, J. (1997) “Explaining Educational Differentials: Towards a Formal Rational Action Theory”, Rationality and Society, volume 9, issue 3, pp 275-305

Breen, R. (2004) “Social Mobility in Europe”, Oxford University Press

Brown, S. Sessions, J. & Taylor, K. (2004) “What Will I Be When I Grow Pp? An Analysis of Childhood Expectations and Career Outcomes”, University of Leicester Department of Economics Working Paper Number 05/2

Brunello, G. Lucifora, C. & Winter-Ebmer, R. (2001) “The Wage Expectations of European College Students”, CEPR Discussion Papers 2817

Brunello, G. Lucifora C. & Winter-Ebmer, R. (2004) “The Wage Expectations of European Business and Economics Students”, Journal of Human Resources, volume 39, p 1116-1142

Bukodi, E. Dex, S. & Goldthorpe, J. (forthcoming) “The Conceptualisation and Measurement of Occupational Hierarchies: A Review, a Proposal and Some Illustrative Analysis”

Cameron, S. & Heckman. J. (1999) "Can Tuition Combat Rising Wage Inequality?", In Kosters, M. (ed), Financing College Tuition: Government Policies & Educational Priorities

Card, D. (1995) “Using Geographic Variation in College Proximity to Estimate the Return to Schooling”, in Christofides, L., Grant,E. & Swidinsky, R. (eds) Aspects of Labour Market Behaviour: Essays in Honour of John Vanderkamp, p 201-222

Carstensen, L. & Woltman, H. (1979) “Comparing Earning Data from the CPS and Employer’s Records”, Proceedings of the Social Statistics Section, American Statistical Association, p 168-173

Carneiro, P. & Heckman J. (2003) “Estimating Distributions of Treatment Effects with an Application to the Returns to Schooling and Measurement of the Effects of Uncertainty on College”, NBER Working Paper 9546

Carvajal, M. Bendana, D. Bozorgmanesh, A. Castillo, M. Pourmasiha K. Rao, P. & Torres, J. (2000) “Inter-Gender Differentials Between College Students’ Earnings Expectations and the Experience of Recent Graduates”, Economics of Education Review, volume 19, p 229-243

Chevalier, A. & Conlon, G. (2003) “Does it pay to attend a prestigious university?”, CEE Discussion Paper 33

Chevalier, A. Gibbons, S. Thorpe, A. Snell, M. & Hoskins, S. (2009) “Students’ Academic Self-Perception”, Economics of Education Review, volume 28, p 716-727

Chowdry, H. Crawford, C. Dearden, L. Goodman, A. & Vignoles, A. (2008) “Widening Participation in Higher Education: Analysis using Linked Administrative Data”, Institute for Fiscal Studies report

Chowdry, H. Crawford, C. & Goodman, A. (2009) “Drivers and Barriers to Educational Success Evidence from the Longitudinal Study of Young People in England”, DCSF research report RR102

Cowan, B. (2009) “Forward-Thinking Teens: The Effects of College Costs on Adolescent Risky Behavior”, Universiy of Madison, Wisconsin job market paper

Cunha, F. Heckman, J. & Lochner, L. (2006) "Interpreting the Evidence on Life Cycle Skill Formation," Handbook of the Economics of Education, Elsevier.

Curtin, T. Ingels, S. Wu,S. & Heuer, R. (2002) “National Education Longitudinal Study of 1988, Base-Year to Fourth Follow-up Data File User's Manual”, National Center for Education Statistics

Davies, R., Heinesen, E. & Holm, A. (2002) “The Relative Risk Aversion Hypothesis of Educational Choice”, Journal of Population Economics, volume 15, issue 4, pp 683–713

Department for Business, Innovation and Skills (2008) “Higher Education at Work, High Skills: High Value”

Department for Business, Innovation and Skills (2009) “Widening Participation Synthesis of International Evidence”, Department for Business, Innovation and Skills Research Paper Number 2

Dickson, M. (2009) “The Causal Effect of Education on Wages Revisited”, IZA Discussion Paper 4419

Dominitz, J. & Manski, C. (1994) “Eliciting Student Expectations of the Returns to Schooling”, NBER Working Paper 4936

Dominitz, J. & Manski, C. (1996) “Eliciting Student Expectations of the Returns to Schooling”, Journal of Human Resources, volume 31, p 1-26

Dominitz, J. (1998) “Earnings Expectations, Revisions, And Realizations”, The Review of Economics and Statistics, volume 80, p 374-388

Dustmann, C (2004) "Parental Background, Secondary School Track Choice, and Wages," Oxford Economic Papers, Oxford University Press, volume 56, p 209-230

Duncan, G. & Hill, D. (1985) “An Investigation of the Extent and Consequences of Measurement Error in Labor Economic Survey Data”, Journal of Labor Economics, volume 3, p 508-522

Emmerson, C. Frayne, C. McNally, S. & Silva, O. (2005) “Evaluation of Aimhigher: Excellence Challenge: The Early Impact of Aimhigher: Excellence Challenge on Pre-16 Outcomes: An Economic Evaluation”, Department for Education and Skills Research Report 652

Erikson, R. & Goldthorpe, J. (1992) “The Constant Flux: A Study of Class Mobility in Industrial Societies”, Clarendon Press

Finch, S. Jones, A. Parfrement, J. Cebulla, A. Connor, H. Hillage, J. Pollard, E. Tyers, C. Hunt, W. & Loukas, G. (2006a) “Student Income and Expenditure Survey 2004-2005 Report”, Department for Education and Skills Research Report 725

Finch, S. Jones, A. Parfrement, J. Cebulla, A. Connor, H. Hillage, J. Pollard, E. Tyers, C. Hunt, W. & Loukas, G. (2006b) “Student Income and Expenditure Survey 2004-2005 Technical Report”, Department for Education and Skills Research Report 725

Freeman, R. (1971) “The Market for College-Trained Manpower”, Cambridge: Harvard University Press

Friedman, M. & Friedman, R. (1980) “Free to Choose”

Ganzeboom, H., Luijkx, R. & Treiman, D. (1989) “Intergenerational Class Mobility in Comparative Perspective”, Research in Social Stratification and Mobility, volume 8, pp 3-84

Ganzeboom, H. De Graaf, P. & Treiman, D. (1992) "A Standard International Socio-Economic Index of Occupational Status", Social Science Research , volume 21, p 1-56

Giannelli, G. & Micklewright, J. (1993) “Estimating Fixed Effect Binary Choice Models with Long Panels: a Practical Approach to the Conditional Logit Model”, Statistica, volume 3, p 1-14

Gottfredson, S. (2002) “Gottfredson’s Theory of Circumscription, Compromise, and Self Creation”, In D. Brown (Ed.) Career Choice and Development , 4th edition, p 85-148

Gustman, A. & Stafford, F. (1972) “Income Expectations and the Consumption of Graduate Students”, Journal of Political Economy, volume 80, p 1246-1258

Gutman, L. & Akerman, R. (2008) “Determinants of Aspirations”, Department for Children, Schools and Families Research Report 27

Hsiao, C. (1986) “Analysis of Panel Data”, Cambridge University Press

Hanushek, E. & Woessmann, L. (2010) “The Economics of International Differences in Educational Achievement”, IZA Discussion Paper 4925

Jerrim, J & Micklewright, J (forthcoming) “Children’s Education and Parents’ Socio-economic Status: Distinguishing the Impact of Mothers and Fathers”

Kahneman, D. & Tversky, A. (1979) “Prospect Theory: An Analysis of Decision Under Risk”, Econometrica, volume 47, issue 2, pp 263-91

Kaufmann, K. & Attanasio, O. (2009) “Educational Choices, Subjective Expectations, and Credit Constraints”, NBER Working Paper 15087

Khoo, S. & Ainley, J. (2005) “Attitudes, Intentions and Participation”, LSAY Research Report 41

Machin, S. & Vignoles, A. (2004) “Educational Inequality: The Widening Socioeconomic Gap”, Fiscal Studies, volume 25, p 107–28

Manski, C. (1993) “Adolescent Econometricians: How Do Youth Infer the Returns to Schooling?”, in Clotfelder, C. & Rothschild, M. (eds.) Studies of Supply and Demand in Higher Education, p 43-60

Manski, C. (2004) “Measuring Expectations”, Econometrica, volume 72, p 1329-1376

Manski, C. & Molinari, F. (2010) “Rounding Probabilistic Expectations in Surveys”, Journal of Business and Economic Statistics, volume 28, p 219-231

Micklewright, J. Schnepf, S. & Skinner, C. (2010) “Non-response Biases in Surveys of School Children: The Case of the English PISA samples”, S3RI Applications & Policy Series Working Paper A10/01

Minnicozi, A. (2003) “Estimation of Sons’ Intergenerational Earnings Mobility in the Presence of Censoring”, Journal of Applied Econometrics, volume 18, p 291-314

Montmarquette, C. Cannings, K. & Mahseredjian, S. (2002) “How Do Young People Choose College Majors?”, Economics of Education Review, volume 21, p 543-556

Morgan, S. (1998) “Adolescent Educational Expectations. Rationalized, Fantasized, or Both?” Rationality and Society, volume 10, p 131–162

Morgan, S. (2004) “Methodologist as Arbitrator: Five Methods for Black-White Differences in the Causal Effect of Expectations on Attainment”, Sociological Methods and Research, volume 33, p 3–53

Morgan, S. (2005) “On the Edge of Commitment: Educational Attainment and Race in the United States”, Stanford University Press

Murphy, K. & Welch, M. (1990) “Empirical Age-Earnings Profiles”, Journal of Labor Economics, volume 8, p 202-229

Office for Fair Access (2010) “What More Can be Done to Widen Access to Highly Selective Universities?”, Office for Fair Access Report

Oosterbeek, H. & Ophem, H. (2000) “Schooling Choices: Preferences, Discount Rates and Rates of Return”, Empirical Economics, volume 25, p 15-34

Organisation for Economic Co-operation and Development (1998) “Education At a Glance: OECD Indicators 1998”, OECD Paris

Organisation for Economic Co-operation and Development (2004a) “Learning for Tomorrow’s World – First Results from PISA 2003”, Paris: Organisation for Economic Co-operation and Development

Organisation for Economic Co-operation and Development (2004b) “PISA 2003 Technical Report”, Paris: Organisation for Economic Co-operation and Development

Organisation for Economic Co-operation and Development (2004c) “PISA 2003 data analysis manuals for SAS and SPSS users”, Paris: Organisation for Economic Co-operation and Development

Organisation for Economic Co-operation and Development (2009) “Education At a Glance 2009: OECD Indicators”, Paris: Organisation for Economic Co-operation and Development

Patton, W. & Creed, P. (2007) “The Relationship between Career Variables and Occupational Aspirations and Expectations for Australian High School Adolescents”, Journal of Career Development, volume 34, p 127-148

Purcell, K. Elias, P. Davies, R. & Wilton, N. (2005) “The Class of ’99:A Study of the Early Labour Market Experience of Recent Graduates”, Department for Education and Skills Research Report 691

Reynolds, J. & Pemberton, J. (2001) "Rising College Expectations among Youth in the United States: A Comparison of 15 and 16 Year Olds in the 1979 and 1997 NLSY", Journal of Human Resources, volume 36, p 703-726

Rouse, C. (2004) “Low-Income Students and College Attendance: An Exploration of Income Expectations”, Social Science Quarterly, volume 85, p 1299-1317

Rubinstein, Y. & Weiss, Y. (2006) “Post Schooling Wage Growth: Investment, Search and Learning”, In Hanushek, E. & Welch,F. (eds.), Handbook of the Economics of Education, volume 1, p 1-67

Schulz, W. (2005) “Measuring the Socio-economic Background of Students’ Achievement in PISA 2000 and PISA 2003”, Paper presented at the Annual American Educational Research Association (AERA), San Francisco

Schütz G., Ursprung, H. & Wössmann, L. (2008) “Education policy and equality of opportunity”, Kyklos, volume 61, p 279–308

Sewell, W. Haller, . & Ohlendorf, G. (1970) “The Educational and Early Occupational Status Attainment Process: Replication and Revision”, American Sociological Review, volume 35, p 1014-1027

Shavit, Y. & Blossfeld, H. (1993) “Persistent Inequality: Changing Educational Attainment in Thirteen Countries”, Boulder, Colorado: Westview Press

Shields, C. & Mohan, J. (2008) “High-Quality Education for All Students: Putting Social Justice at its Heart”, Teacher Development, volume 12, p 289-300

Smith, A. (1776) “An Inquiry into the Nature and Causes of the Wealth of Nations”

Smith, H. & Powell, B. (1990) “Great Expectations: Variations in Income Expectations Among College Seniors Sociology of Education”, volume 63, p 194-207

Steedman, H. & McIntosh, S. (2001) “Measuring Low Skills in Europe: How Useful is the ISCED Framework?”, Oxford Economic Papers, volume 53, p 564-581

Sutton Trust (2008), “Increasing Higher Education Participation Amongst Disadvantaged Young People and Schools in Poor Communities”, Sutton Trust report to the National Council for Educational Excellence

Sutton Trust (2010), “The Mobility Manifesto”

Vignoles, A. (2007) “The Value of Higher Education in the Post-Expansion Era”, Presentation to the British Educational Research Association at the Institute of Education

Webbink, D. & Hartog, J. (2004) “Can Students Predict their Starting Salary? Yes!”, Economics of Education Review, volume 23, p103-113

Williams, G. & Gordon, A. (1981) **“**Perceived Earnings Functions and Ex Ante Rates of Return to Post Compulsory Education in England”**,** Higher Education, volume 10, p 199-227

Willis**,** R. & Rosen, S. **(**1979) "Education and Self-Selection", Journal of Political Economy, volume 87, p S7-36

Wolter , S. (2000) “Wage Expectations: A Comparison of Swiss and US Students”, Kyklos, volume 52, p 51-69

Wolter, S. & Zbinden, A. (2002) “Labour Market Expectations of Swiss University Students”, International Journal of Manpower, volume 23, p 458-470

Appendices

Appendix 2.1 Definition of part-time and dependant students[140]

The SIES (Finch et al 2006a) used the following definition for whether a student is classed as a dependant or independent student:

Dependant Student:

These are full-time students who:

(a) Had applied for student support and their parents, step parent or legal guardian's income taken into account
(b) Were aged under 25 years, were unmarried and had not applied for student support

Independent Students:

All those not in the above two groups.

I also contacted The National Centre for Social Research to clarify the definition used for a dependant student. The questions asked and answers given appear below.

Q: Is a person, who has applied for student support, unmarried, under 25, but *has not had* their parent's income taken into account classed as a dependant or independent student?

A: Independent student

[140] Note a version of this Appendix was also included in my 2007 University of Southampton MSc thesis.

Q: Is a person, who has applied for student support, unmarried, under 25, *had* their parents' income taken into account**,** *but is above the threshold* for any further support other than the basic level, a dependant or independent student?

A: Dependant student

This highlights that whether parents' income has been taken into consideration is important to deciding how the student is classed. Anyone who is 25 or over, married or has not had their parents' income taken into account is classed as an independent student. Those whose parent's income has been taken into account are dependant students. One problem identified from the responses may be that parents know they are above the income threshold for any further support and hence do not disclose this information on the UCAS application form. Therefore it is possible that some respondents, who are actually dependant students, are mistakenly classified as independent.

Appendix 2.2 Construction and difficulties in measuring parental income[141]

Annual parental income in the SIES has been recorded in band widths of £5,000 or £10,000, though the respondent could decide whether to report the information in a gross or net amount. I decided to put all net data onto the gross scale (as this was the most commonly recorded). To do this, I adjust the “net” band limits upwards to their gross equivalent, taking into account income tax and national insurance rates for the 2004/05 financial year (under the assumption that all individuals are contracted into the state earnings-related pension system). The modifications to the data are shown below.

Table A2.1 Conversion between net and gross parental income

Net value £	Gross value £	Category generated £
0-5,000	0-5,000	20,000 and below
5001-10,000	5,000-12,000	20,000 and below
10,001-15,000	11,900-19,400	20,000 and below
15,001-20,000	19,400-27,000	20,001-40,000
20,001-25,000	27,000-34,000	20,001-40,000
25,001-30,000	34,000-41,000	20,001-40,000
30,001-40,000	41,000-58,000	40,001+
40,001-50,000	58,000-75,000	40,001+
50,001+	75,001+	40,001+

Notes: 1 Author’s own calculations

There are some limitations to this technique. The choice of groups is largely dictated by the data. For instance a net salary of £15,000 is roughly equal to a gross salary of £20,000. As £15,000 and £20,000 are both cut off points, it is sensible to create a category of income £20,000 or below to minimize overlap between groups. With the categorizations used, there should be little overlap, though this can not be totally avoided. For instance a student may know that their parents’ net earnings are £15,100, equivalent to £19,500 gross, per year. However this student would be put into the group £20,000-£40,001 gross per annum because of the overlap problem. It is reasonable to suggest that the analysis will not be severely affected by this, as the overlap is small.

Some other assumptions must be made about this variable. The question asked is

[141] Note a version of this Appendix was also included in my 2007 University of Southampton MSc thesis.

about the total income of parents. This may complicate the conversion between net and gross. In particular the tax, if only one parent is earning the income, is greater than if two parents are working. For example, consider two households with £30,000 net income. Household A has one parent working who earns the whole £30,000. The gross equivalent is £41,000 per year. Household B however has two parents earning £15,000, with gross equivalent being £38,800. Hence there is a difference between the gross equivalence due to the tax system, which would put household A into the £40,000+ bracket and household B into the £20,001-£40,000 group. Furthermore there is an issue that some forms of income are not taxable, such as child benefit. For simplicity, it has been assumed that all parental income is taxable and has been generated by one adult in the household. With the boundaries chosen the effect is probably quite small, but it is still important to note this difficulty.

It is also important to recognize that this variable may suffer a reasonably large degree of measurement error, as it relies on students reporting their parents' income. A further point to note is that some (36) students failed to state whether they were reporting figures in gross or net terms. In this instance it has been assumed that students are reporting gross figures.

It should also be noted that there are other ways in which one may re-code this variable. One drawback in what I have done is that it results in quite a coarse measurement of parental income (i.e. there are only three groups). An alternative is to take the midpoint of the groups, and create a quasi-continuous variable. This would have the benefit of providing a broader sense of parents' income, though the difficulty of conversion between net and gross figures would still exist. Likewise, there would be the issue of an undefined upper threshold for the top earning group. In both ways of handling the data there is a significant chance of measurement error. Thus the variable should be viewed as an approximate measure of parents' income.

Appendix 2.3 Interval regression results

Students' wage expectations in the SIES tend to bunch around round numbers, despite an open text field allowing precise estimates to be recorded. One can take this into account when estimating regression coefficients by assuming that students do not expect to get exactly the salary they report, but give a "ball-park" figure. As an example, a student predicting a salary of £13,000 may be presenting their midpoint estimate or reporting to the nearest thousand. In reality they expect a salary between £12,500 and £13,499. Hence their actual estimate is unknown, but lies within a interval.

If it is hence assumed that:

- Students who do <u>not</u> report their expected salary on a multiple of £5,000 (i.e. they report a figure of say £22,000 rather than a value such as £15,000, £20,000, £25,000 etc) round their estimate to the nearest thousand. So I assume someone who reports £22,000 really means between £21,500 and £22,499. For the few observations (roughly 5% of the sample) that give a very precise figure (e.g. £22,300) I place them within the most appropriate interval (e.g. I would place someone reporting £22,300 into the £21,500 to £22,499 interval).
- Students who report an expected wage which <u>is</u> a multiple of £5,000 (e.g. £15,000, £20,000, £25,000, £30,000 and £35,000) report their expected wage to the nearest £5,000. Hence I assume someone who reports £25,000 is actually expecting somewhere between £22,500 and £27,499

I make these different assumptions to try and account for the large number of observations at certain points in the expected salary distribution. In particular, examination of the data suggests that extra clustering occurs at numbers that are rounded to the nearest £5,000 (see Table 2.2). Consequently it seems appropriate to assume that students who are reporting these figures are exercising a greater degree of rounding and have a wider anticipated salary range. It is also necessary to assume that the unobserved response (expected starting wage) is normally distributed.

Results for the censored regression model appear in the Table A2.2 below. Compared to the original OLS regression (Table 2.4), little changes with the introduction of the censoring assumption and use of interval regression. Most coefficients alter by around 0.2 to 0.3% suggesting that, even when assuming quite extreme rounding by students, there are limited differences compared to using ordinary least squares.

Table A2.2 Interval regression results

	Specification 4	
Variable	**Coefficient**	**Standard error**
Future plans (Ref: Career Job Only)		
Temporary job only	-0.228	0.025
Either a career or temporary job	-0.096	0.023
Further study or travel	-0.005	0.013
Hard to get graduate job (Ref: Agree)		
Neutral	0.021	0.018
Disagree	0.059	0.012
Missing	0.081	0.033
Proximity to graduation (Ref: Final year)		
1 year	0.014	0.014
2 or more years	0.035	0.017
University type (Ref: Post-1992)		
Other Pre-1992	-0.031	0.018
Russell group	-0.024	0.026
Parents earnings (Ref: Below £20,000)		
£20,001-£40,000	0.031	0.019
£40,001+	0.053	0.020
Independent student or missing data	0.087	0.030
How parents earns (Ref: Work)		
Benefits	-0.067	0.043
Investments	0.116	0.029
Ethnic group (Ref: White)		
Black/Asian	0.055	0.025
Mixed/Other	0.019	0.023

Notes:

1* Significant at 5% level

2 Source: Authors calculations using the SIES data. Sample size in all regressions is 2,659. Dependent variable is the natural logarithm of students' expected wages

Table A2.2 Interval regression results continued

	Specification 4	
	Coefficient	Standard error
Total income		
Mean centred (per £0000)	0.036	0.017
Study mode (Ref: Full-time)		
Part-time student	-0.033	0.025
Earnings from work		
Mean centred (per £0000)	0.030	0.021
Part-time student *Earnings from work	0.141	0.032
Subject area (Ref: Medicine)		
Allied To Medicine	-0.162	0.032
Sciences	-0.183	0.032
Maths, Computer Science	-0.141	0.035
Engineering, Technology	-0.111	0.034
Architecture, Building	-0.182	0.036
Social Studies	-0.155	0.029
Law	-0.114	0.043
Business	-0.149	0.036
English, Languages, Classics	-0.225	0.036
History, Philosophy	-0.235	0.046
Arts	-0.285	0.035
Education	-0.181	0.028
Combined	-0.199	0.038
Other	-0.215	0.048
Entry qualification (Ref: A-levels)		
GNVQ/AVCE	-0.072	0.023
Other	-0.013	0.015
Age		
Mean centred	0.013	0.005
University location (Ref: Other England)		
London	0.105	0.021
Wales	-0.036	0.022
Gender (Ref: Male)		
Female	-0.052	0.013
University dummies		
Constant	10.009	0.040

Appendix 2.4 Probit results for salary non-response in the DLHE

Table A2.3 Results for probit model of non-response

	Group	Coefficient	SE
Domicile (Ref: London)	North East	0.221*	0.027
	North West	-0.092*	0.019
	Yorkshire	0.005	0.021
	East Midlands	0.027	0.021
	West Midlands	-0.009	0.020
	East	-0.054*	0.020
	South East	0.011	0.018
	South West	0.009	0.021
	Isle of Man/ Channel Islands	-0.391*	0.110
	Unknown	-0.072*	0.024
University location (Ref: England)	Wales	-0.240*	0.021
Term-time accommodation (Ref: University maintained property)	Parental home	-0.136*	0.016
	Own home	-0.043*	0.014
	Other	-0.090*	0.019
	Unknown	-0.025	0.021
Degree class (Ref: 1st)	2.1	-0.082*	0.017
	2.2	-0.180*	0.018
	3rd	-0.276*	0.027
	Unclassified	-0.166*	0.046
	Not applicable	-0.107*	0.024
UCAS score mean centred (100 point increase)		0.039	0.006
University type (Ref: Post-1992)	Russell Group	-0.097*	0.014
	Pre-1992	-0.133*	0.013
Subject (Ref: Medicine)	Allied to Medicine	0.126*	0.040
	Biology	0.174*	0.043
	Physical Sciences	0.227*	0.042
	Maths	0.183*	0.046
	Computer Science	0.165*	0.040
	Engineering	0.249*	0.042
	Social Sciences	0.210*	0.039
	Law	0.139*	0.046
	Business	0.202*	0.038
	Mass Communication	0.067	0.044
	Languages	0.102*	0.040
	History	0.069	0.042
	Art	-0.117*	0.039
	Education	0.012	0.042
	Combined	0.050	0.071
	Psychology	0.223*	0.043
	Sports Science	0.144*	0.044
	Other	0.089	0.049

Table A2.3 Results for probit model of non-response (continued)

	Group	Coefficient	SE
Graduate level job (Ref: Yes)	Non-graduate job	-0.115*	0.030
Degree required for job (Ref: Formal requirement)	Expected	-0.066*	0.018
	Advantageous	-0.075*	0.015
	No	-0.271*	0.015
	Don't know	-0.702*	0.034
Job type (Ref: Managerial)	Professional	0.042*	0.020
	Associate professional	0.059*	0.018
	Administrative	0.100*	0.033
	Skilled labour	-0.090	0.059
	Personal services	-0.064	0.039
	Sales or customer service	-0.053	0.036
	Construction	0.042	0.074
	Elementary	-0.366*	0.043
Ethnicity (Ref: Asian)	Black	0.015	0.039
	Other/ Mixed	0.135*	0.037
	Unknown	0.035	0.036
	White	0.148*	0.018
Disabled (Ref: No)	Yes	-0.072*	0.019
Gender (Ref: Male)	Female	-0.098*	0.010
	Constant	0.308	0.052

Notes:
1 * indicates significance at the 5% level
2 Figures refer to marginal effects. A negative coefficient suggests a lower probability of reporting salary
3 UCAS score refers to student performance on tests typically taken at age 18 in England and Wales, that largely determine university entry.
4 Source: Authors calculations using the DLHE dataset. Total sample size in the regression was 87,327. Dependent variable in the regression model these figures were based upon was a binary indicator of whether the graduate reported their salary (coded 0 if they did not and 1 if they did).

Appendix 2.5 The difference between HESA's official average graduate wage and the figures used in chapter 2

HESA report the average graduate wage for the 2004/2005 year group as £19,000, calculated from the DLHE survey used in chapter 2[142]. This is in fact the starting wage calculated from the data (£18,531), rounded to the nearest thousand. The average graduate wage presented in chapter 2 is lower than this official figure. The table below shows how the figures presented in chapter 2 relate to the official HESA average. As can be seen, the difference in graduate salaries is largely due to the different samples being considered.

Among the most important points is that I have calculated wages *separately* for those who studied part-time and those who studied full-time (selection rule 1). In comparison, the official HESA figure relates to when these groups are analysed *together***.** An important result, shown in section 2.7, is that part-time students, on average, are actually quite realistic in their wage expectations. It is full-time students who, on average, overestimate their starting wage and this should be made explicitly clear when reporting results.

A further adjustment is that I have scaled the HESA data back to 2005 prices using RPI inflation, recorded as 2.8% in 2005 (selection rule 4). As explained in section 2.6, the SIES was conducted in January to March 2005, whereas the DLHE for this cohort was conducted in early 2006. The existing academic literature suggests that individuals report their wage expectations in current prices. Since data on actual wages has been collected a year later than expectations, it is necessary to account for the inflation over this period. It should be noted, however, that this has only a moderate influence on the results, and the general findings of chapter 2 would still hold if it is not made.

[142] See http://www.hesa.ac.uk/index.php/content/view/126/161/

Other points to note are that I have excluded medical students (selection rule 5) and those over 25 from the study of full-time students (selection rule 3). Medics have been excluded due to the different proportion of these students in the SIES compared to the DLHE. Although by excluding medics the DLHE starting salary is reduced, the same selection rule has been applied to the SIES, with an even sharper fall in expectations. Furthermore, only students under 25 have been considered in my analysis. My intention in studying full-time and part-time students separately was to investigate how realistic "traditional" university students are, who have little pre-existing experience of the labour market. Again this restriction has a largely negligible influence on results.

A final point is that the official HESA data does not take into account the large number of graduates not reporting their salary, who are generally working in non-graduate jobs and have 2.2 or 3rd class degrees. When this is taken into account (via the weights I produce as part of chapter 2), I estimate that HESA's official figure is upwardly biased by around 3%, or £500 (see rule 6).

Table A2.4 Comparison between the average graduate salary reported by HESA and the average graduate salary used in this chapter

	Average salary £	**Selection rule**
Raw HESA data	20,314	
1st degree students Only	18,381	
Only those in full-time jobs	18,540	
Only those employed in UK *	18,531	
Only those who were studying full-time when at university	17,720	1
English and Welsh universities only	17,743	2
Under 25 years old only	17,336	3
Scaled to 2005 prices	16,788	4
Medics excluded **	16,455	5
Weighting for non-response	15,996	6

Notes:
1 The figures in the average salary column relate to the cumulative effect of all the sample selection rules.
2 * Official HESA figure
3 ** Figure used in this chapter
4 Authors own calculations using DLHE data

Appendix 3.1 NELS sample design[143] (Source: Curtin et al 2002)

This bulk of this Appendix has been taken (on occasion word for word) from Curtin et al (2002). These authors explain the NELS sample design at great length. I reproduce their work to help the reader understand some of the technicalities of the NELS sampling process. Although I have rephrased and edited part of their text, it should be noted that I claim none of Appendix 3.1 to be my own independent work.

The sample for NELS: 88/94 (i.e. age 20 sampling frame) was created by dividing the NELS:88/92 (i.e. age 18) sample into 18 groups based on their response history, dropout status, eligibility status, school sector type, race, test scores, socioeconomic status, and freshened status. Each sampling group was assigned an overall selection probability. Cases within a group were selected such that the overall group probability was met, but the probability of selection within the group was proportional to each sample member's second follow-up (age 18) design weight. Assigning selection probabilities proportional to the second follow-up (age 18) design weight, reduced the variability of the NELS:88/94 (age 20) raw weights and consequently increased the efficiency of the resulting sample from 40.1 percent to 44.0 percent. The groups were:

0. Excluded from age 20 follow-up

The age 20 follow-up sample is a spring defined sample. Therefore students who had been brought in through the freshening process, but who had dropped out by the time of data collection, as well as the age 14 dropouts were assigned to this group. As these groups have been excluded from the age 20 follow-up, they have a sampling probability of zero. In addition, sample members who were ineligible or out of scope (dead or out of country) for the age 18 follow-up were also assigned to this group.

1. Nonresponders

These sample members had never completed a questionnaire in any round

2. Poor responders

These are sample members who did not complete a age 18 questionnaire (but had

[143] See http://nces.ed.gov/pubs2002/2002323.pdf for more details

responded at either age 14 or 16)

3. Ever dropped out

Sample members for whom Curtin et al (2002) have evidence that they ever dropped out of school (including those who were in school during periods of data collection) were included in this group.

4. Ineligible to participate (due to language barriers or mental or physical impairment) prior to age 18

5. Attended a private school at age 14

6. Attended a private school in either age 16 or 18

7. Hispanic

8. Asian or Pacific Islander (API)

9. Native American

10. Black, top quartile in cognitive tests

11. Black, other test scores

12. White, lowest socioeconomic quartile

13. White, highest socioeconomic quartile

14. White, middle socioeconomic quartiles

15. Freshened in at age 16

16. Freshened in at age 18

17. Other

The Table below lists the groups, their selection probabilities and their age 16 and 18 follow-up distributions. While some sample members qualified for more than one of the sample groups, each member was assigned to only one group. The groups were created in order of priority, so that each sample member was assigned to the first group for which they qualified. For example, if someone was both a dropout (group 3) and was in a private school at age 14 (group 5), he or she was assigned to group 3.

The data used to assign the students to groups was drawn from a variety of possible sources, including questionnaire data for variables such as race and school sector type. If status at time of data collection was relevant and was not determined at the time of data collection, the imputed status developed during the age 18 weighting process was used.

Table A3.1 Sampling frame and selection probabilities NELS age 18 and 20 follow-up

	Selection Probability of being included in age 20 sample	N (Age 18 Sample)	N (Age 20 Sample)
TOTAL		21635	15964
Excluded	0	731	0
Non-responders	0.15	288	43
Poor responders	0.25	2383	596
Ever dropped out	1	2351	2351
Ineligible to participate	0.9	212	191
Attended private school at age 14	0.8	2984	2387
Attended private school at either age 16 or 18	0.8	122	98
Hispanic	0.9	1629	1466
Asian or Pacific Islander	1	874	874
Native American	1	132	132
Black, top quartile of cognitive tests	1	79	79
Black, other	0.9	1238	1114
White, lowest socio-economic group	1	1295	1295
White, highest socio-economic group	0.6	2536	1522
White, middle socio-economic group	0.8	4763	3810
Brought into sample at age 16	0.3	4	1
Brought into sample at age 18	0.3	6	2
Other	0.4	8	3

Notes: 1 Source Curtin et al 2002

Table A3.2 Sampling frame and response rates – NELS age 20 follow-up

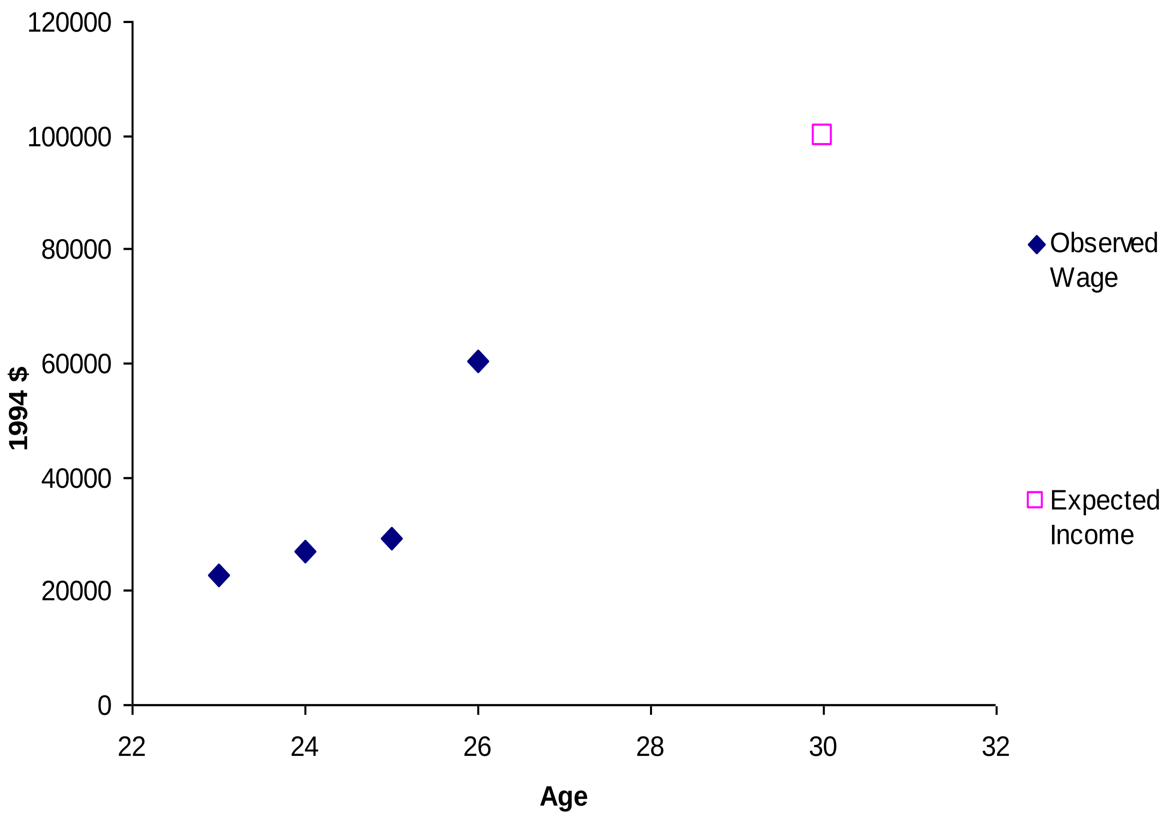
	Sampling frame age 20	Respondents to survey age 20
Gender		
Male	7,895	7,354
Female	7,980	7,561
Race/Ethnicity		
Asian	1,151	1,088
Hispanic	2,288	2,107
Black	1,840	1,681
White	10,303	9,787
Native American	230	211
Missing	63	41
Age 18 test quartile		
Lowest	2,669	2,497
2^{nd}	2,850	2,710
3^{rd}	2,836	2,746
4^{th}	2,982	2,923
Missing	55	53
Did not complete test	4,483	3,986
Socio-economic status		
Lowest	4,062	3,788
2^{nd}	3,784	3,587
3^{rd}	3,742	3,570
4^{th}	3,635	3,507
Missing	652	463
Drop out Status		
Never dropped Out	13,337	12,654
Ever dropped out	2,538	2,261
Age 14 school type		
Public	13,383	12,540
Catholic	1,355	1,292
NAIS private	595	568
Other private	542	515
Total	15,875	14,915

Notes: 1 Source Curtin et al 2002

Appendix 3.2 Methods to predict age 30 income

I now return to the two problems with the NELS data that I highlighted at the end of section 3.3, and briefly overviewed in section 3.4. To begin, consider Figure A3.1[144]. This illustrates the data observed for one particular individual in the NELS.

Figure A3.1 Observable wage and income expectation data for ID 7286532 in the NELS



Note:
1 All data in 1994 wages
2 This individual reported zero unearned income at age 26; therefore his wages are equivalent to his income. Note that in these diagrams, I am simply trying to explain my extrapolation method for *wages*.
3 Discussion of unearned income can be found later in this Appendix.
4 This individual is not an example of a "typical" NELS respondent. Rather, I have chosen this observation as it provides a good example of the points I am tying to make. Most respondents see a gradual increase in their wage between 23 and 26, and not such a large increase at age 26.
5 Source: Authors calculations from the NELS dataset.

[144] For this particular individual, the income they expect is significantly higher than their predicted income at age 30. This is not necessarily typical of all other respondents in the dataset. Rather I have chosen this individual as he is a good example of the substantial points I make throughout this section.

Respondents are asked what they expect their annual income to be when they turn 30. However, information on realisations is only available for wages between the ages of 23 and 26. Using the available data, I must make a prediction of each individual’s age 30 income. I separate this into two parts: (a) the estimation of wages, and (b) the estimation of unearned income.

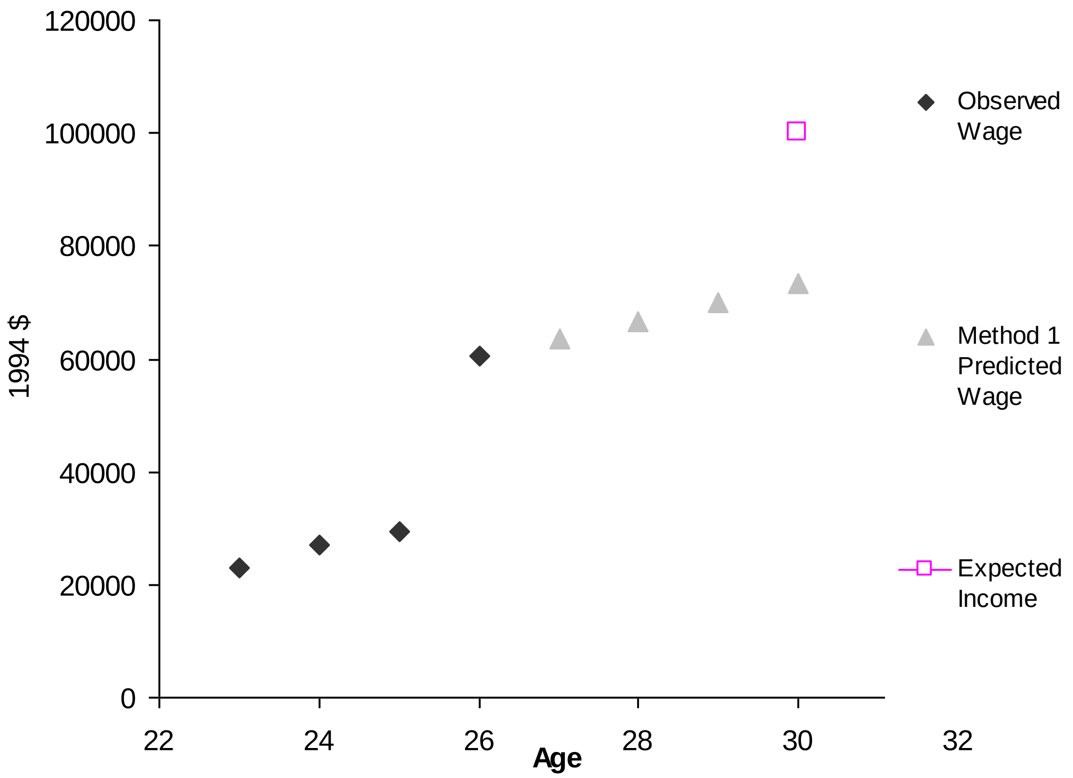
Wages

On average, wages grow quite substantially with the first ten years of labour market experience. Yet the path of wages for a given individual between 26 and 30 may be quite unstable. Both upwards and downwards shocks are possible due to job or career changes (promotion or redundancy), different family and location choices and preferences (prefer leisure to work due to the birth of a child) or simply macroeconomic conditions and “luck” (a particularly large bonus or commission in a given year). Given the factors described above, one may or may not wish to consider an individual’s labour market history in predicting their age 30 wage. For instance, a sudden job change may make the information conveyed in past wages irrelevant. Recall Figure A3.1. At age 26, this individual has a particularly large wage by his "historical" (age 23- 25) standards. This may be the result of him changing job, perhaps into a much more lucrative career, where his labour market history is irrelevant for his future wages. Indeed, I do not know the reason for this potential change of job, but it could be a change in his preferences. For instance, this could be a graduate who took a low intensity job to enjoy life when young, though at age 26 made the decision to start a career. One may view this as a *permanent* shift in his wage profile. What he was previously earning, before this permanent shift, is irrelevant in predicting his future wage.

In contrast, it is equally possible for this to be a *temporary* increase in his wage, for instance from sheer good (or bad) luck. Take a 26 year old salesman who has had a particularly good year. For the salesman, the jump in wages could reflect a large bonus. However, in the future things may not be so good, with his wage reverting to his historical average (e.g. the average of the previous three years). Hence this observation at 26 may be treated more as an outlier, a sudden (but *temporary)* change in income.

Given the range of possibilities, I use two methods to predict age 30 wages. Method 1 views large changes in wages as a *permanent* shift in an individual's circumstance. Therefore, his previous wage profile is treated as irrelevant; it is only the most recently observed (e.g. age 26) wage that contains any useful information about his wage at age 30. Under this method, I simply take the most recently observed wage for each individual and extrapolate it forward, using external estimates of wage growth for young workers. Figure A3.2 presents a hypothetical example for the illustrative individual in Figure A3.1, assuming a real growth rate of 5% per annum. Observe that only the wage at age 26 influences my prediction, and that the large shock at age 26 gets carried forward. The previous income profile of the individual (between 23 and 25) has no influence at all.

Figure A3.2 Illustration of wage prediction method 1 for ID 7286532 in the NELS



Note: 1See notes to Figure A3.1
2 The above is a hypothetical example of extrapolation Method 1. I assume that his wage will grow at 5% per annum between the ages 26 and 30. His previous wage history (the wages he received between 23 and 25) play no part in the age 30 income prediction
3 Source: Authors calculations from the NELS dataset.

To implement this method, I require an external estimate of the annual real wage growth for young workers. Rubinstein and Weiss (2007) provide a table of average annual real wage growth rates, as implied by a Mincer wage equation, broken down by labour market experience and educational attainment for three surveys; the Current Population Survey (CPS), Panel Survey of Income Dynamics (PSID) and National Longitudinal Survey of Youth 1979 (NLSY 79) [145]. Furthermore, Rubinstein and Weiss restrict each of the above datasets to full-time, male, American workers (as I have done with the NELS). One should note, however, that these surveys all relate to different years[146]. The growth rates they calculate from the CPS, PSID and NLSY are provided in Table A3.3, with further details available on page 14 and Appendix 5 of Rubinstein and Weiss (2007).

Table A3.3 Average, annual (real) wage growth rates for young workers: Rubenstein and Weiss estimates

		% Average (real) wage growth rate per annum by education level				
Number of years experience in the labour force	**Data source**	**Below high school**	**High school**	**Some college**	**College graduates**	**MA/PhD**
0-10	CPS	2.4	3.2	3.3	3.6	2.9
	PSID	2.8	3.0	3.8	3.9	3.2
	NLSY	2.4	3.4	4.6	5.2	5.5
11-15	CPS	1.6	2.2	-	-	-
	PSID	1.9	2.0	-	-	-
	NLSY	1.3	2.3	-	-	-

Notes:
1 Source: Table 1, page 14 of Rubinstein and Weiss (2007) Post Schooling Wage Growth: Investment, Search and Learning. Handbook of the Economics of Education, Volume 1

2 CPS: Current Population Survey Annual March Supplement
3 PSID: Panel Survey of Income Dynamics
4 NLSY: National Longitudinal Study of Youth 1979

[145] It should be noted that Rubinstein and Weiss provide two sets of growth rates, one based on a Mincer quadratic specification (experience and experience squared), and the other based on cell means. Their justification for the latter method is based on work by Murphy and Welch (1990), who claim the quadratic specification fits the age-earnings profile poorly, especially at the early stages of workers careers. One may worry that using the growth rates implied by a Mincer equation here could lead to underestimation of future wages. However, the paper by Murphy and Welch shows that the error in the quadratic wage specification is small after 3 years labour market experience and reaches zero at around 5 years. This means that for the period I am trying to extrapolate to, the quadratic Mincer specification fits the actual data quite well.

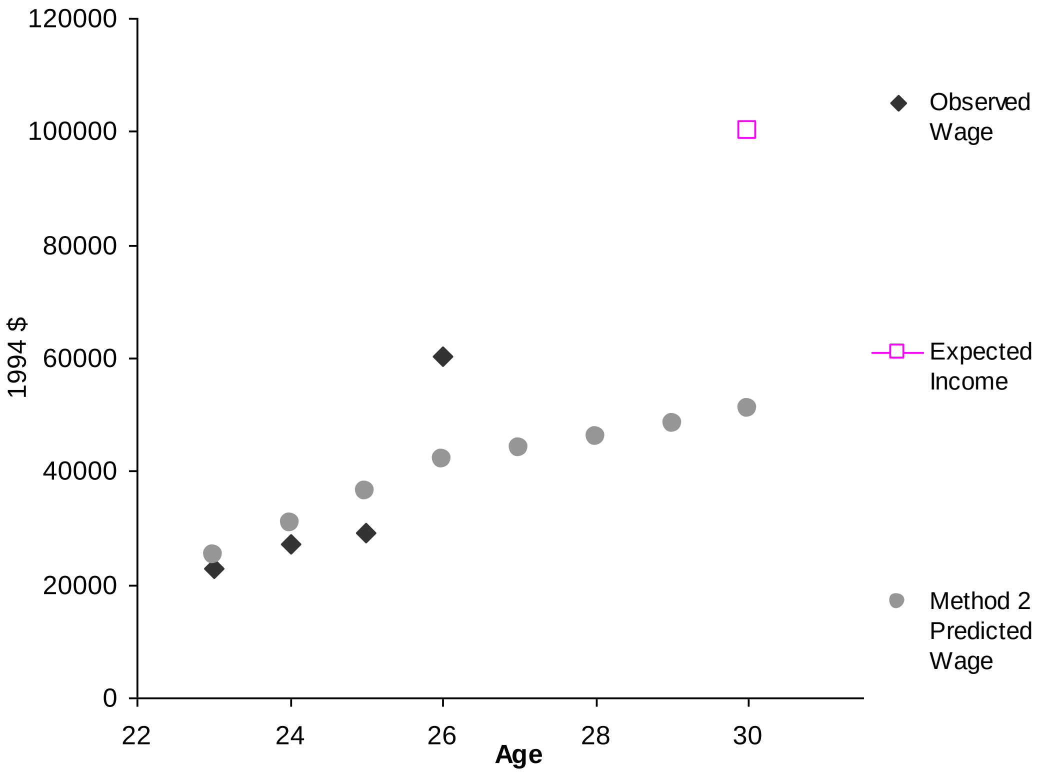
[146] The CPS data relates to wages between 1998 and 2002, the PSID is for all years after 1968, while the NLSY draws its information between 1979 and 2000.

These growth rates are applied to each individual in the NELS, depending on their highest qualification achieved by age 26. For example, an individual with college education, and who was earning $50,000 dollars at age 26, would be estimated to be earning $61,240 at age 30 (all in 1994 prices) [147]. In the event that wages go unobserved at age 26 (e.g. the individual was unemployed) I extrapolate from their last observed full-time wage[148]. From this point on, I shall call this prediction "Method 1"[149].

As suggested, this may not be an appropriate method if wage shocks (as for the respondent in Figure A3.1) are only temporary (e.g. a salesman with a large bonus). If this is the case, an individual's future wage will be randomly scattered around his time mean. The wage *history*, rather than just the most recent observation, is now informative. In prediction "Method 2" I take this into account. A hypothetical example is shown in Figure A3.3. To make the difference absolutely clear, contrast this with Figure A3.2 (that uses prediction Method 1). Even though both assume the same real growth rate (5% per annum) they generate quite different predictions of wages at age 30.

[147] \$50,000 * (1.052^4), using the NLSY data and "College graduates" column in Table A3.3.
[148] For example, if someone was earning \$50,000 at age 25, and their wage was not recorded at age 26, I would predict their age 30 income to be \$50,000 * (1.052^5) = \$64,577
[149] Rubenstein and Weiss provide growth estimates for young workers using three separate surveys (PSID, CPS and NLSY). I shall therefore use the notation "Method 1a" for my predictions when using their CPS growth estimates, "Method 1b" for their PSID estimates and "Method 1c" for their NLSY estimates.

Figure A3.3 Illustration of wage prediction method 2 for ID 7286532 in the NELS



Note:
1See notes to Figure A3.1
2The above is a hypothetical example of extrapolation Method 2.
3 Source: Authors calculations from the NELS dataset.

Hence "Method 2" uses individual's wage history, rather than just the most recent observation, to predict future income. To implement Method 2 I use a fixed effects regression model, following the methodology of Carneiro and Heckman (2003)[150]. The natural logarithm of wages is the dependent variable, with age and age-squared as (time varying) explanatory terms that capture wage growth[151]. Other specifications, such as including a cubic age term or using a set of age dummies as an alternative, were also estimated. However results did not differ significantly from the parsimonious quadratic specification. I run separate regressions for five different

[150] Carneiro and Heckman (2003) faced a similar problem in having to estimate wages up to age 65 for a sample whose last observed wage was at age 35. In particular, they pool their data with an additional source, and use a similarly specified fixed effects model to estimate wages into the future. However, unlike Carneiro and Heckman, I run separate regressions for each educational group. Moreover, they specify an autoregressive error term, whereas I assume it to be random. In other words, they allow the particularly large wage shown in Figure A3.3 to revert to the mean (estimated fixed effect) over a series of years, whereas I assume it returns there instantly in the next period. Hence the method of Carneiro and Heckman is a sort of middle ground between the two extrapolation methods that I am proposing here. I also experimented with an auto-regressive error term, but found no change to my substantial results.
[151] Age enters as a quadratic term to allow for flattening of the age-earnings wage profile.

educational groups, based on highest qualification achieved by age 26[152]. This allows the age coefficients, and therefore wage growth, to vary between groups with different levels of human capital. I have also experimented with alternative specifications that allowed wage growth to vary within these educational groups, and found similar results[153]. In all models, I assume the error term is independent and identically distributed, scattered randomly around each individual's fixed effect. Formally, this model can be expressed as:

$$\mathrm{Y}_{\mathrm{ia}} = \alpha + \beta_0 \mathrm{A}_{\mathrm{ia}} + \beta_1 \mathrm{A}_{\mathrm{ia}}^2 + \eta_{\mathrm{i}} + \varepsilon_{\mathrm{ia}} \ \nabla \in \mathrm{E}$$

Where:

Y_{ia} = log earnings of individual i at time a

A_{ia} = Age of individual i at time a

η_i = Individual (or fixed) effect

ε_{ia} = Error term, assumed to be normally distributed

E = Five education groups (Less than high school, high school, associates degree, bachelors degree, MSc/PhD)

In this model, it is the estimated fixed effect, η_i, that captures the influence of all wages for individual i between 23 and 26. Note that this specification is quite different to a "standard" wage equation, where the aim is to estimate the impact of various regression coefficients on the outcome (wages). My concern, on the other hand, is not in estimating the importance or effect sizes of various explanatory variables, but in predicting future wages. Therefore I allow the individual fixed effect to capture all the factors that are usually included on the right hand side of "standard" wage equation. This includes geographic location, individual ability and socio-economic background[154].

[152] Minicozzi (2003) suggests using separate regressions for different education-occupation combinations. Here, separate regressions are estimated only for different educational groups. Considering occupation would have lead to vastly reduced sample sizes and imprecisely estimated coefficients.

[153] In particular, I estimate a model where I allow wage growth to differ between college students who study different subjects. All the results presented in section 3.5 are robust to these additional specifications.

[154] However, I do account for human capital separately by estimating five regressions based on each individual's highest educational attainment at age 26.

The estimated coefficients from the five regressions enter a prediction equation for age 30 wages, formally specified as:

$$\hat{Y}_{i30} = \alpha + \hat{\beta}_0 A_{i30} + \hat{\beta}_1 A^2_{i30} + \hat{\eta}_i + \hat{\varepsilon}_{i30} \qquad \nabla \in E$$

With

$\hat{Y}$ = Predicted log wage at age 30

A= Age

$\hat{\eta}$ = Individual fixed effect

$\hat{\varepsilon}_{30}$= Random draw from the distribution of errors at age 26 (assumed to be normally distributed)

E = Achieved education at age 26

This prediction includes an error term. I assume the errors at age 30 are normally distributed with mean zero and variance equal to that in the estimated error distribution at age 26. I then take a random draw from this normal distribution for each individual.

One could estimate the preceding model, and form predictions, based solely on the NELS data. However this would have some fairly significant disadvantages. The age coefficients, which reflect wage growth, would be estimated solely from data in the observed period (wages recorded between the ages of 23 and 26). One would be assuming that the annual wage growth rate between 26 and 30 is the same as the wage growth rate between 23 and 26. This seems unlikely. Murphy and Welch (1990) show that earnings between 23 and 26 grow substantially faster than between 26 and 30. Moreover, with wages recorded at only 4 time points in the NELS, the quadratic age function would be poorly defined. On the other hand, using just a linear age function would miss an important empirical feature (flattening) of the age-earnings profile.

Thus the NELS data must be complemented with additional information on how wages grow in the unobserved period (27 to 30 years old). One method, used by Carneiro and Heckman (2003), is to pool the truncated survey (the NELS, which only contains wages until 26), with a second comparable data source that follows individuals to the point of interest (up to age 30). This pooled dataset will therefore contain information on wages between 23 and 30. However, certain criteria must be checked and some assumptions must be made. In particular, one implicit assumption is that the (unobserved) wage growth rate experienced by NELS sample members between the ages of 26 and 30 will be the same as the (observed) growth rate experienced by sample members from the second pooling survey. A further assumption is that structural differences in the economy, and between the two samples, do not lead to differences in wage *growth*[155]. It is also vital the two surveys are collecting comparable data, with similar wording of key questions.

The survey chosen is the National Longitudinal Study of Youth (NLSY) 1979[156]. The survey began in 1979, with 12,686 men and women surveyed who were between the ages of 14 and 22. These individuals were then followed each year, and have information on their income at age 30 collected between 1987 and 1995. I make similar restrictions in the NLSY as I have done in the NELS (I have excluded women and only consider the wages of individuals when they are working full-time).

The NLSY has numerous attractions as a source to pool with the NELS. Critically, wages are collected for individuals between the ages of 23 and 30, providing information on wage growth during the period not observed (between 26 and 30) in the NELS. Secondly, the wording of the questions regarding wages is broadly similar. Whereas the NELS asks:

[155] Of course, overall wage levels are likely to be higher for cohorts from later periods. However this general rise in the wage level should be captured by the person specific fixed effect.
[156] The United States Department of Labor describes "The NLSY79 is a nationally representative sample of 12,686 young men and women who were 14-22 years old when they were first surveyed in 1979". The National Bureau of Economic Research describe the quality of this survey, and the high response rate. In particular, it notes that 87% of those selected for interview responded in the base year (1979), while 86.7% of eligible respondents took part in 1996. Further details can be found at www.nber.org/~kling/surveys/NLSY79.html and http://www.nlsinfo.org/nlsy79/docs/79html/79text/front.htm

First, including all of the wages, salaries, and commissions you earned in (1997/1998/1999), about how much did you earn from employment before taxes and all other deductions?

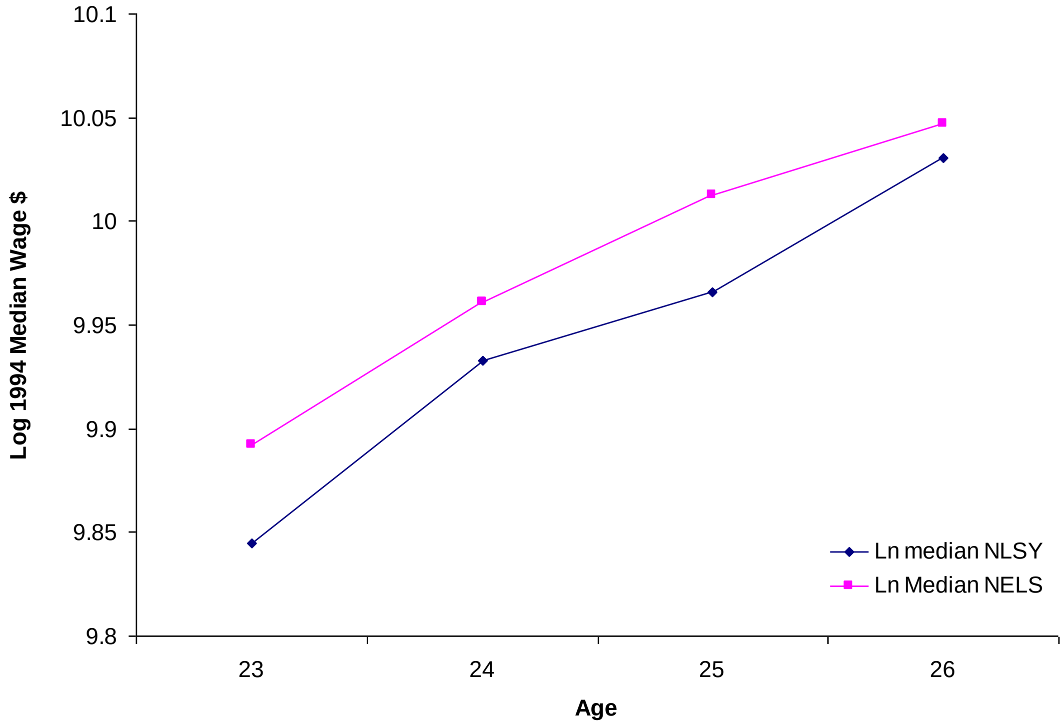
The NLSY uses the similar phrase:

Not counting any money you received from your military service during (YEAR), how much did you receive from wages, salaries, commissions or tips from all jobs, before deductions for taxes or anything else?

Although the NLSY excludes military wages in this question, this has been recorded separately in another item. Thus this is easily added in to also include those who were working in the military. Otherwise, the comparability seems good. Both ask for gross wages, before tax or other deductions. Moreover, it is made clear that the respondent should be taking into account all aspects of employment related income, including all wages and commissions, from all employment held.

It is also important to check that during the observable period (ages 23 to 26), wages in the two surveys grow at a similar rate. If wage growth is vastly different during the period observed in both surveys, it would be difficult to justify the assumption that the NELS will have similar wage growth to the NLSY in the unobserved (ages 27 to 30) period. Appendix Figure A3.4 presents the median wage recorded at each age in the NELS and NLSY. Although the median wage in the NELS is above that in the NLSY, the growth in wages between 23 and 26 is similar. For instance, mean (median) wages grew by 23% (17%) in the NELS between the ages 23 and 26, compared to 23% (22%) in the NLSY. In terms of the mean, this difference is very small. Though the gap is larger for the medians (around 1% per year), it still seems that the growth rate in the two surveys is reasonably similar. This, hopefully, means the wages that go unobserved in the NELS, between 27 and 30, will also follow a similar growth pattern to the NLSY.

Figure A3.4 Log median wages in the NELS and NLSY between the ages 23 and 26



Notes 1 Source: Authors calculations from the NELS and NLSY datasets.

To further investigate this point, I conduct a Chow test to investigate whether the age coefficient (which reflects wage growth) differs between the NELS and NLSY for wages observed between 23 and 26. Since wages are only available for four time points, age is kept as a linear function. Separate regressions are ran for each education group, and takes the form:

$$\mathrm{Y}_{\mathrm{ia}} = \boldsymbol{\alpha} + \boldsymbol{\beta}_0 \mathrm{A}_{\mathrm{ia}} + \boldsymbol{\beta}_1 \mathrm{A}_{\mathrm{ia}} * \mathrm{S}_{\mathrm{i}} + \boldsymbol{\eta}_{\mathrm{i}} + \boldsymbol{\varepsilon}_{\mathrm{ia}} \ \nabla \in \mathrm{E}$$

Where:

Y_{ia} = log earnings of individual i at time a

A_{ia} = Age of individual i at time a

S_{i} = Dummy variable indicating respondent was part of the NELS survey (NLSY reference group)

η_{i} = Individual (fixed) effect

$\varepsilon_{\mathrm{ia}}$ = Random error term

E = One of five education groups (Less than high school, high school, associates degree, bachelors degree, MA/PhD)

The test is formally specified as:

H_O: $\beta_1 = 0$

H_A: $\beta_1 \neq 0$

The results are given in the Table A3.4. As expected, all age terms are significant. However the real interest rests on the Age-Survey interaction terms (the column labelled "Difference"). These show whether growth in average wages, for each schooling group, differs between the NELS and NLSY between the ages of 23 and 26. Out of the five regressions, two are statistically significant. However, these are for the two smallest education groups, which in total make up only around 10% of the NELS sample. Indeed the coefficient for MSc/PhD graduates is likely to be poorly defined due to the limited number of wage observations in the NELS for these individuals. In the other three regressions, which account for 90% of observations, the age-survey coefficient is very small. Indeed, if the Chow test is performed on the sample as a whole, with no distinction between education groups, the results suggest a difference in wage growth rates of under 1% per year. This is consistent with my claim that the NELS and NLSY cohorts experience similar wage growth patterns between 23 and 26 years of age. The assumption, based on this result, is that the NELS and NLSY samples also experience similar growth rates between 26 and 30.

Table A3.4 Chow test to investigate whether wage growth is similar for young men between the ages 23 and 26 for the NELS and NLSY surveys

Education group	Average per annum real wage growth NLSY	Average per annum real wage growth NELS	Difference
Below high school	0.049 (0.0055)	0.047	-0.002 (0.0128)
High school graduate	0.057 (0.003)	0.051	-0.006 (0.006)
Associates degree	0.093 (0.0061)	0.072	-0.021 (0.010)*
Bachelors degree	0.145 (0.0067)	0.141	-0.004 (0.008)
MSc/PhD	0.155 (0.0155)	0.250	0.095 (0.025)*

Notes:
1 The difference column relates to the Chow test of whether wage growth differs between the ages of 23 and 26 for each education group in the two surveys. In other words, this is the test of Ho: $\beta_0 = 0$ in the hypothesis test specified above.
2 Standard errors are presented in parenthesis
3* Indicates statistical significant at the 5% level
4 Source: Authors calculations from the NELS and NLSY datasets. NELS sample size 4,434. Dependent variable is the natural logarithm of each individual at each time point (between ages 23 and 26) that they have data available.

A final issue may be that structural changes in the economy led to differential growth rates in real wages between the two surveys. However between 2000 and 2004 (the period unobserved in the NELS) average annual real wage growth in the US was around 3.2%. Between 1987 and 1995 (when members of the NLSY were turning 30) wage growth averaged around 4%. Although there is a difference, it appears to not be substantial, though it could lead to a slight overestimation of the predicted age 30 NELS wage.

Having established the comparability of the two surveys, I proceed to pool the information from these two datasets together (which, from this point on, I will call the NELS-NLSY pooled data). Using this data, I predict age 30 wages by estimating the fixed effect regression specified on pages 235-236. Table A3.5 provides the regression coefficients for the age and age squared terms from the five estimated models (recall I estimate five separate regressions based on educational attainment by age 26).

Table A3.5 Estimated age coefficients from prediction method 2 (fixed effects regression model)

Education level at age 26	Variable	Coefficient	SE
Below high school	Age	0.051	0.0071
	Age^2	-0.003	0.0010
High school	Age	0.053	0.0037
	Age^2	-0.003	0.0005
Associates degree	Age	0.091	0.0072
	Age^2	-0.004	0.0010
Bachelors	Age	0.171	0.0051
	Age^2	-0.011	0.0008
MA/PhD	Age	0.200	0.0152
	Age^2	-0.011	0.0020

Notes:
1These are the estimated age coefficients from the fixed effects regression model described above.
2These coefficients reflect the estimated wage growth between 26 and 30. Table A3.6 converts these coefficients into estimated annual wage growth for ease of interpretation and comparison to the wage growth rates suggested by Rubenstein and Weiss.
3 All coefficients are statistically significant at the 5% level
4 Source: Authors calculations from the NELS and NLSY datasets. NELS sample size 4,434. Dependent variable is the natural logarithm of each individual at each time point (between ages 23 and 26) that they have data available.

All coefficients are statistically significant at the 5% level, capturing the quadratic effect of age and the flattening of the age-earnings profile. For interpretation purposes, however, it is easier to convert these coefficients into estimated annual growth rates (as per Rubinstein and Weiss). I present these in the final column of Table A3.6, which also contains the average annual wage growth estimates from prediction Method 1 for comparison.

Table A3.6 Predicted average, annual real wage growth rates for young American men between the ages 26 and 30

	Estimated % real growth rate per annum			
	Method 1			Method 2
	(a) CPS	(b) PSID	(c) NLSY	NELS-NLSY Pooled
Below high school	1.6	1.9	1.3	1.6
High school	2.7	3.6	2.8	2.5
Associates degree	3.3	3.8	4.6	5.2
Bachelors	3.6	3.9	5.2	5.0
MA/PhD	2.9	3.2	5.5	6.6

Notes:
1"Below high school" annual growth rate is taken from the 11-15 years experience row in Rubenstein and Weiss, under the assumption that they would have (potentially) entered the labour market at 16/17. For annual growth rates of high school graduates, I use a simple average of the Rubenstein and Weiss 0-10 years experience and 11-15 years experience columns. The assumption is that high school graduates enter the labour market at 18, and thus have around 8 years labour market experience at age 26 increasing to 12 years experience at age 30.
2 Individuals in all other educational groups (associates degree, bachelors degree and MA/PhD) are assumed to have 0-10 years labour market experience between 26 and 30.
3 Source: Authors calculations from the NELS dataset. NELS sample size 4,434.

It appears predicted wage growth is similar across all methods for each of the educational groups. As expected, the estimated growth rates from Method 2 are the best aligned with Method 1c (Rubenstein and Weiss's estimated wage growth when they use the NLSY data), though are slightly higher (lower) for the top (bottom) educational group[157].

[157] The Rubenstein and Weiss growth rate "Method 1" is based upon an average for the first ten years labour market experience. On the other hand, in the NELS, those with MSc/PhD level education are only likely to have up to 5 years experience between 26 and 30. Likewise, those with high school education in the NELS will have between 8-12 years experience. This is the most likely reason for the slightly differences, and that Method 1 growth rates are slight under (over) estimates for the most (least) educated.

I also present the average predicted age 30 wage for the two methods in Table A3.7. Predicted age 30 wages are similar across estimation methods, with differences typically less than 5% for both the mean and the median. However, the spread of Method 2 (the fixed effect model) is smaller. Indeed this is as expected; outlying observations get moderated in Method 2 by the influence of previous wages (it is a time mean). This does not occur in Method 1, where it is only the most recent observation that is used for prediction. Hence if there is a large shock to the most recent observation, this gets carried forward to the future prediction in Method 1, as opposed to being averaged out in Method 2.

Table A3.7 Predicted annual wage at age 30 for NELS sample members

Prediction method	Dataset	Median predicted wage $000	Mean predicted wage $000 (standard deviation)
1	CPS	25.7 (3.4)	29.4 (19.2)
1	PSID	25.5 (3.4)	29.5 (19.3)
1	NLSY	26.1 (3.4)	30.4 (20.1)
2	NELS-NLSY Pooled	26.7 (3.3)	29.6 (14.8)

Notes:
1 All figures presented in 1994 wages
2 Figures in parenthesis represent the spread of the data (p90/p10 for median, standard deviation for the mean)
3 Source: Authors calculations from the NELS dataset. Sample size = 4,434..

I also compare my predictions of average age 30 wages for different groups to similar information recorded for 30 year olds in an external data source (the 2003-2005 CPS March Annual Supplement)[158]. The results appear in Table A3.8.

[158] The exact wording in the CPS is as follows: “How much did (name/you) earn from this employer before taxes and other deductions during (Year)?”. This is supplemented with other questions to check the robustness of answers and to calculate other wage sources. In particular, respondents are asked “How much did (name/you) earn in tips, bonuses, overtime pay or commissions from this employer in (Year)” and “What is your best estimate of (name's/your) correct total amount of earnings from all other employers during (Year)?”. All of these responses are used to calculate respondents earned income, making the definition comparable to the other surveys in question (NELS, NLSY, PSID).

Table A3.8 Predicted *mean* age 30 NELS wage compared to the *mean* age 30 Current Population Survey (CPS) wage

	% of observations in NELS	Predicted wage method 1c ($000)	Predicted wage method 2 ($000)	% of observations in CPS	CPS wage ($000)
Highest qualification at age 26					
Below high School	5.4	20.9	20.8	12.4	15.8
High school	56.4	27.0	25.5	47.8	24.7
Associates degree	7.0	29.9	30.5	8.2	28.1
Bachelors	28.0	37.7	37.8	24.6	38.1
Masters degree / PhD	2.8	42.4	43.6	6.8	44.5
Race					
White	69.8	31.4	31.3	60.6	31.5
American Indian	0.1	23.6	24.6	0.2	NA
Asian or Pacific Islander	5.1	38.5	36.4	6.2	35.4
Black (not Hispanic)	8.3	24.6	25.1	9.6	26.9
Hispanic	13.3	27.4	27.1	21.7	20.7
Other	4.6	27.1	26.6	1.7	NA
All respondents	100.0	30.4	29.6	100.0	28.9

Notes:
1 All observations in 1994 $
2 Total sample size in the NELS is 4,434. In CPS, the total sample size is 1,412
3 CPS Wages for American Indian and Other ethnic groups are not reported due to the small sample size.
4 CPS data from 2002-2004 March Annual Supplement, restricted to men working full-time, all year round, at age 30 available from : http://www.census.gov/hhes/www/cpstc/cps_table_creator.html
5 The proportion of respondents with below high school education is lower in the NELS than CPS. This is because CPS is a general population survey. The NELS data I am using represents the population who were in high school as seniors in 1992. Hence the NELS and CPS cover slightly different populations, particularly regarding those with less than high school education. On the other hand, the CPS contains a lot more individuals with MA or PhD level qualifications. This is due to the last NELS wave being conducted at age 26. Some of the NELS cohort would still be in higher education, and are still studying for these qualifications.
6 The three left-hand most columns refers to authors calculations from the NELS dataset, the two columns on the right are based on data from the CPS taken from from 2002-2004 March Annual Supplement, restricted to men working full-time, all year round, at age 30 available from : http://www.census.gov/hhes/www/cpstc/cps_table_creator.html. Total sample size in the NELS is 4,434. In CPS, the total sample size is 1,412

In general, predicted age 30 wages are similar to those in the external CPS data. I predict average wages to be $29,600 in the NELS, while in the CPS the equivalent figure is $28,900. Even when looking at subgroups of the population, differences tend to be quite small. For example, I predict the mean wage of those with a Bachelors degree to be $37,800, while in the CPS the average wage for those with a degree is $38,100. Likewise, I predict White respondents to earn a mean wage of $31,300, while in the CPS the figure is $31,500. Nevertheless, there are some groups where predicted wages are quite different from average wages in the CPS. For instance, I estimate the average wage of those with less than high school education to be around $21,000 while the CPS figure stands at just over $15,000. This may be because the NELS data represents the population of high school seniors in the spring of 1992. Hence my definition of “less than high school education” is those who made it into the final year of high school but did not graduate. On the other hand the CPS represents the whole US population, and defines less than high school education as everyone who did not graduate from high school, *including* those who dropped out *before* their senior year. This is probably the reason why, in the NELS compared to the CPS:

(a) my predicted wage is higher

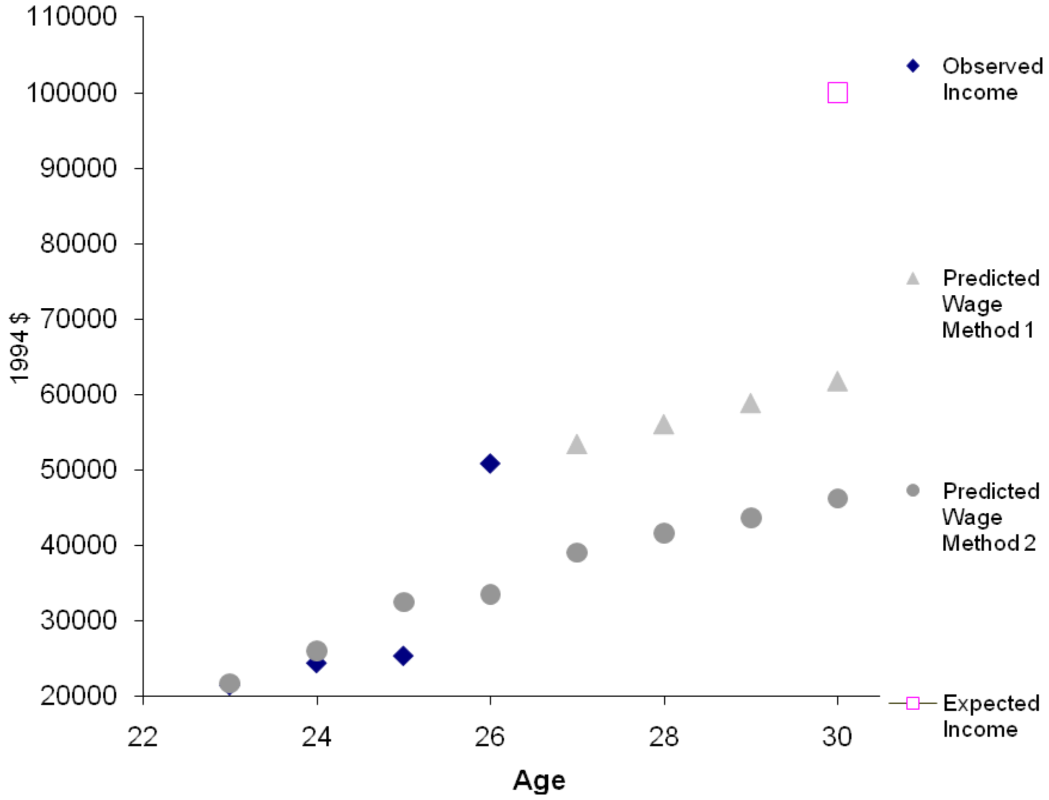
and

(b) there are a smaller proportion of respondents with below high school education

In a similar manner, I predict average wages for Hispanics to be around $27,000, while the CPS figure is closer to $20,000. This again could be due to slight differences in wording or response to the question regarding race and ethnicity in the CPS and NELS surveys. However, it is worth noting that both these groups only form a small part of the overall sample. The general message is that my prediction methods seem to generate a reasonable estimate of average age 30 wages.

In Figure A3.5, however, I show my two predictions for the illustrative NELS respondent. This highlights the difficulty of analysis at the individual level.

Figure A3.5 Comparison of wage prediction methods for ID 7286532 in the NELS



Notes:
1 See notes to Figure A3.3, A3.4 and A3.5
2 Source: Authors calculations from the NELS dataset.

The two predictions of his age 30 wage are $15,000 (30%) apart. Given my discussion at the start of the section about whether the large jump in wages is temporary or permanent, anywhere between the two predictions, or even a figure outside of this range, could be possible. When dealing with group averages, over-estimates of wage growth for some individuals will be compensated by underestimates for others. However a comparison of expected and actual wages at the individual level is troublesome, as there is a large range of possible values for each individual's predicted age 30 wage.

Unearned Income

Thus far I have only considered wages; I now turn my attention to how other sources of finance contribute to total age 30 income. Some details were collected in the NELS about respondents' non-wage income at age 25 (the full previous year, 1999, prior to the survey). They were asked:

Without considering the earnings from employment that you just reported, approximately how much did you and your (spouse/partner) receive from other sources of income in 1999? $..

These sources might include stocks and bonds, savings interest, insurance, alimony or child support, amily members, and disability payments

As this information is collected in just one question, measurement error may be a concern. Another difficulty is that the question asks for joint unearned income for the respondent and their partner (if married or cohabiting). Fortunately, the majority of those who did report a figure were not in a marriage or marriage like partnership. Appendix Table A3.9 presents the distribution of unearned income by marital status.

Table A3.9 Distribution of unearned income by marital status, for those reporting a value above zero

Percentiles of unearned income distribution	Single	Married/Cohabiting
1	30	100
5	200	200
10	500	500
25	1,200	1,100
50	4,000	3,250
Mean	8,528	7,335
75	10,000	7,000
90	20,000	15,000
95	30,000	30,000
99	60,000	65,000
Standard Deviation	16,045	13,298
% of observations where reported unearned income >0	25.7	19.0
% of observations missing	2.6	8.9
Observations (including 0's)	3,924	510

Notes:
1 Around 75% (80%) of single (married or cohabiting) individuals report 0 unearned income. The distributions above relate only to those who reported some form of unearned income (a value greater than 0)
2 Source: Authors calculations from the NELS dataset. Sample size 4,434

The reported distributions of unearned income by married and single individuals are quite similar, though the former are more likely to not respond and the latter more likely to report 0.

The first question to ask is how much unearned income do young adults receive, and what proportion of total income does it make up at (a) the group and (b) the individual level? The second column of Table A3.10 provides details of the unearned income distribution at age 26 for NELS respondents. One striking feature is that the majority of individuals (74%) report no unearned income. It would appear that, even though unearned income is important conceptually, empirically it has relatively little influence on average total income (at least at the group level).

To investigate this proposition further, I again turn to the NLSY where, as opposed to the NELS, respondents are asked several questions about each aspect of their non-earned income. For instance, they were asked about their income from businesses, public support, educational grants and any other sources in a series of separate questions[159]. Table A3.10 compares the distributions of unearned income for NELS and NLSY sample members at age 26, for those reporting a value above 0.

Table A3.10 Distribution of unearned income at age 26 in the NELS and NLSY, for those reporting a value greater than 0

Percentiles of unearned income distribution	Reported unearned income in NELS	Reported unearned income in NLSY	Reported unearned income in NLSY for those reporting a value greater than $500
1	50	4	510
5	200	14	578
10	500	29	662
25	1,200	116	1,156
50	3,876	578	2,553
Mean	8,079	3,347	6,202
75	9,000	2,753	6,314
90	20,000	7,967	14,287
95	30,000	14,571	26,893
99	70,000	42,563	57,150
% reporting>0	26%	38%	20%

Notes:
1 All data are for individuals at age 26 in 1994 prices
2 Distribution is for respondents reporting a value greater than 0
3 Source: Authors calculations from the NELS and NLSY datasets. NELS sample size 4,434 (1,153 observations with unearned income greater than 0)

Notice firstly the NLSY has a greater proportion (38% compared to 26%) of people reporting positive unearned income. However, the distribution shows almost a quarter of these observations are less than $100. It seems that the NELS, by recording this data in a single question, misses many individuals who have a small quantity of unearned income. In any case, both the NELS and NLSY suggest that unearned income, *on average*, has only a small influence on total income at age 26. The median respondent indicates they have no unearned income. Even of the minority that do report a figure above 0, unearned income (on average) is relatively small compared to wages in most cases.

---

[159] Note in the NLSY, respondents were asked separate questions about their spouses unearned income aswell.

One may suggest that unearned income may make up a more significant proportion of total income at age 30 than at age 26. To investigate this, I compare mean wages to the mean total income for men in the 2003-2005 CPS March Annual Supplement[160]. On average (mean), unearned sources of finance contribute only $500 (2%) to total income. I performed a similar analysis on the NLSY 79 sample when they turned age 30, and found a similar result (unearned income makes a very small contribution to total income at the group or population level).

Overall it seems that, on average, unearned income makes up only a very small part of total age 30 incomes. Hence it should be of limited importance when one compares expectations to realisations at the group level. Therefore, to incorporate unearned income into my predictions, I simply use the value recorded at age 25 in the NELS. Implicitly this means that anyone with zero unearned income at 25 will also have zero predicted unearned income at age 30. Given its minor role, this should not introduce substantial bias at the group or population level.

On the other hand, unearned income is a non-trivial matter at the *individual* level. Table A3.10 shows some individuals to report a figure over $10,000 at age 25 in the NELS data. But this could be a one-off inheritance from a relative dying, or sudden good luck with a stock option (especially given the technology boom at the time of the survey in 2000). There is no indication about how this unearned income may change in the future. Hence predicting unearned income at age 30 for a given individual is an even harder task than for wages. Thus the NELS simply does not contain the data to make estimation of unearned income at the individual level a realistic possibility.

[160] Several questions about other (unearned) sources of income were asked in the CPS. This includes how much they received from benefits, welfare, assistance, dividends and interest. The data I use is drawn from the CPS "Table Creator", available from http://www.census.gov/hhes/www/cpstc/cps_table_creator.html I produce two values, one looking at men's average wages, the other their total income. I assume that the difference between these two figures (average wages and average income) equals total income from unearned sources.

Summary

Drawing together the results from this appendix, it seems that inferences at the group and population level should be reasonably robust to the problems identified with the NELS data. I have presented two methods to predict age 30 wages, which provide similar estimates of average wages at age 30, and that are comparable with external estimates from population level data. Moreover, even though age 30 unearned income is difficult to predict, I have shown that this makes up only a small proportion of total average income. I am therefore confident that the substantive inferences in section 3.5 regarding population and group level averages are robust to the data issues discussed throughout this section.

However, my concerns for analysis at the individual level remain. Figure A3.5 illustrates how two very different predictions, over $15,000 (30%) apart, can be made for any one individual. I have also assumed this person has no unearned income at age 30, as he did not report any at age 25. This would be quite a bold assumption to make. The implication is that inferences made at the individual level are likely to suffer from what may be quite severe biases. Consequently, I focus on group level analysis (mean and median outcomes), that I believe are robust to the assumptions I have made about the data. Though analysis at the individual level would be of great interest, I do not believe this to be sensible with the NELS data.

Appendix 3.3 Comparison of OLS results using Method 1 to Method 2

Tables A3.11 to A3.13 provide regression results, analogous to these in Table 3.11, except that I now predict age 30 income using "Method 1". This method is described in more detail in section 3.4 and Appendix 3.2. Note that I implement Method 1 three ways, using different estimates of young adults wage growth from different surveys.

Comparing the results to those in Table 3.11, it seems that most of the patterns I describe in section 3.5 still hold. For example, specification 1 consistently shows that workers hold more realistic expectations than Art, Biology and Communication students. And, as described in section 3.5, the inclusion of the college drop-out dummy in specification 4 causes the subject coefficients to drop dramatically. However, it is worth noting that statistical significance has been lost for some groups of students in comparison to Table 3.11. For example, using the PSID to extrapolate wage growth ("Method 1b"), the coefficient estimates for Accounting, Finance and Biological Science groups are now only statistically significant at the 10% level (compared to the 5% level in Table 3.11). This seems to be a result of both a decrease in the estimated coefficient, and more variability in the data (recall my discussion of Table A3.7, where I show the standard deviation of predicted wages to be lower in Method 2 than Method 1). Nevertheless, I can confidently say that, on average, there is still little evidence that workers hold more realistic expectations of their future income than students. Moreover, although some of the coefficients have been reduced to lower levels of statistical significance, the general patterns found regarding specific groups of students still seem to hold. In particular, engineering, maths and computer science students hold more realistic expectations than workers (and those in creative disciplines) across all results. Likewise, I always find those who drop out of university have the least realistic expectations.

Turning to the other coefficients, there is again strong agreement across the prediction methods. Family income, and whether the student also holds a job at age 20, is rarely of statistical significance at any of the conventional levels. On the other hand, cognitive maths ability and the Black race dummy are always significant at the 5% level. Hence the general message from these tables is that the results presented in section 3.5 seem relatively robust to the prediction method that I use.

Table A3.11 OLS regression results comparing how realistic students are to workers (Prediction "Method 1a" using CPS wage growth estimates)

	Specification 1		Specification 2		Specification 3		Specification 4	
	Co	**SE**	**Co**	**SE**	**Co**	**SE**	**Co**	**SE**
Work-student status at age 20 (Ref: Working)								
Agriculture student	-0.10	0.08	-0.05	0.08	-0.07	0.07	-0.12	0.08
Economics, Finance student	-0.09	0.10	-0.02	0.11	-0.05	0.10	-0.12	0.10
Business, Management student	0.01	0.05	0.05	0.05	0.03	0.05	-0.05	0.05
Journalism, Communication student	0.20*	0.08	0.24*	0.08	0.21*	0.08	0.13	0.08
Computer Science, Maths student	-0.11	0.09	-0.06	0.09	-0.10	0.10	-0.20*	0.09
Education student	-0.02	0.06	0.02	0.06	0.00	0.07	-0.06	0.07
Engineering, Physical sciences student	-0.16*	0.05	-0.11*	0.05	-0.14*	0.05	-0.21*	0.06
Language student	-0.08	0.13	-0.02	0.13	-0.02	0.13	-0.10	0.14
Health student	0.19*	0.08	0.24*	0.08	0.22*	0.08	0.15*	0.08
Law student	0.44*	0.20	0.44*	0.19	0.44*	0.20	0.32	0.19
Biological science student	0.24*	0.08	0.30*	0.08	0.29*	0.08	0.21*	0.08
Social sciences, Humanities student	0.17*	0.07	0.22*	0.07	0.18*	0.07	0.10	0.08
Art student	0.31*	0.13	0.36*	0.13	0.35*	0.14	0.25*	0.13
Other student	0.05	0.06	0.08	0.06	0.06	0.06	-0.03	0.06
Not student or working	0.15*	0.06	0.16*	0.06	0.12*	0.05	0.12*	0.05
Missing	0.30*	0.07	0.33*	0.07	0.28*	0.07	0.30*	0.07
Maths ability at age 18	-	-	-0.07*	0.02	-0.06*	0.02	-0.05*	0.02
Race (Ref: White)								
American Indian or Alaska Native	-	-	-	-	0.07	0.08	0.06	0.08
Asian or Pacific Islander	-	-	-	-	0.06	0.07	0.06	0.07
Black, not Hispanic	-	-	-	-	0.22*	0.05	0.20*	0.05
Hispanic or Latino	-	-	-	-	0.11*	0.05	0.10*	0.05
More than one race	-	-	-	-	0.18*	0.07	0.17*	0.07
Missing	-	-	-	-	0.02	0.08	0.00	0.08
Family income parents reported when respondent was age 18 (Ref: Bottom quintile)								
2nd quintile	-	-	-	-	-0.03	0.05	-0.03	0.05
3rd quintile	-	-	-	-	-0.02	0.05	-0.01	0.05
4th quintile	-	-	-	-	0.08	0.05	0.08	0.05
Top quintile	-	-	-	-	0.03	0.06	0.05	0.06
Student at 20, who also held a part-time job (Ref: No)								
Yes	-	-	-	-	0.01	0.04	-0.02	0.04
College Dropout (ref: No)								
Yes	-	-	-	-	-	-	0.21*	0.05
Constant	0.44*	0.03	0.41*	0.03	0.38*	0.05	0.39*	0.05

Notes:
1 These results refer to when I use Rubenstein and Weiss (2007) CPS estimates of wage growth (see Table 3.6) to predict NELS sample members age 30 income.
2 Source: Authors calculations from the NELS dataset. Sample size = 4,434. Dependent variable is the natural logarithm of respondents expected age 30 income divided by their predicted actual age 30 income.

Table A3.12 OLS regression results comparing how realistic students are to workers (Prediction "Method 1b" using PSID wage growth estimates)

	Specification 1		Specification 2		Specification 3		Specification 4	
	Co	**SE**	**Co**	**SE**	**Co**	**SE**	**Co**	**SE**
Work-student status at age 20 (Ref: Working)								
Agriculture student	-0.12	0.08	-0.07	0.08	-0.09	0.07	-0.15*	0.08
Economics, Finance student	-0.09	0.10	-0.03	0.11	-0.05	0.10	-0.14	0.10
Business, Management student	0.01	0.05	0.05	0.05	0.02	0.05	-0.06	0.06
Journalism, Communication student	0.20*	0.08	0.24*	0.08	0.21*	0.08	0.12	0.08
Computer Science, Maths student	-0.12	0.09	-0.06	0.09	-0.11	0.10	-0.22*	0.09
Education student	-0.03	0.06	0.01	0.07	-0.01	0.07	-0.08	0.07
Engineering, Physical sciences student	-0.17*	0.05	-0.11*	0.05	-0.15*	0.05	-0.22*	0.06
Language student	-0.08	0.13	-0.02	0.13	-0.02	0.13	-0.11	0.14
Health student	0.18*	0.08	0.23*	0.08	0.21*	0.08	0.13	0.08
Law student	0.44*	0.20	0.44*	0.19	0.44*	0.20	0.30	0.19
Biological science student	0.22*	0.08	0.28*	0.08	0.27*	0.09	0.19*	0.08
Social sciences, Humanities student	0.15*	0.07	0.21*	0.07	0.16*	0.07	0.07	0.08
Art student	0.30*	0.13	0.35*	0.13	0.34*	0.14	0.23	0.13
Other student	0.05	0.06	0.08	0.06	0.05	0.06	-0.05	0.06
Not student or working	0.14*	0.06	0.15*	0.06	0.12*	0.05	0.12*	0.05
Missing	0.31*	0.07	0.33*	0.07	0.28*	0.07	0.29*	0.07
Maths ability at age 18	-	-	-0.08*	0.02	-0.06*	0.02	-0.05*	0.02
Race (Ref: White)								
American Indian or Alaska Native	-	-	-	-	0.07	0.08	0.06	0.08
Asian or Pacific Islander	-	-	-	-	0.06	0.08	0.06	0.07
Black, not Hispanic	-	-	-	-	0.22*	0.05	0.21*	0.05
Hispanic or Latino	-	-	-	-	0.12*	0.05	0.10*	0.05
More than one race	-	-	-	-	0.16*	0.08	0.14	0.08
Missing	-	-	-	-	0.03	0.08	0.00	0.08
Family income parents reported when respondent was age 18 (Ref: Bottom quintile)								
2nd quintile	-	-	-	-	-0.03	0.05	-0.03	0.05
3rd quintile	-	-	-	-	-0.02	0.06	-0.01	0.06
4th quintile	-	-	-	-	0.08	0.05	0.08	0.05
Top quintile	-	-	-	-	0.02	0.06	0.04	0.06
Student at 20, who also held a part-time job (Ref: No)								
Yes	-	-	-	-	0.02	0.04	-0.02	0.04
College Dropout (ref: No)								
Yes	-	-	-	-	-	-	0.23*	0.05
Constant	0.43*	0.03	0.40*	0.03	0.37*	0.05	0.38*	0.05

Notes:
1 These results refer to when I use Rubenstein and Weiss (2007) PSID estimates of wage growth (see Table 3.6) to predict NELS sample members age 30 income.
2 Source: Authors calculations from the NELS dataset. Sample size = 4,434. Dependent variable is the natural logarithm of respondents expected age 30 income divided by their predicted actual age 30 income

Table A3.13 OLS regression results comparing how realistic students are to workers (Prediction "Method 1c" using NLSY wage growth estimates)

	Specification 1		Specification 2		Specification 3		Specification 4	
	Co	**SE**	**Co**	**SE**	**Co**	**SE**	**Co**	**SE**
Work-student status at age 20 (Ref: Working)								
Agriculture student	-0.15*	0.08	-0.09	0.08	-0.11	0.07	-0.18*	0.08
Economics, Finance student	-0.12	0.11	-0.05	0.11	-0.08	0.11	-0.17	0.10
Business, Management student	-0.02	0.05	0.03	0.05	-0.01	0.06	-0.10	0.06
Journalism, Communication student	0.18*	0.08	0.22*	0.08	0.19*	0.08	0.08	0.08
Computer Science, Maths student	-0.14	0.09	-0.08	0.09	-0.13	0.10	-0.25*	0.09
Education student	-0.06	0.06	-0.01	0.07	-0.03	0.07	-0.11	0.07
Engineering, Physical sciences student	-0.19*	0.05	-0.13*	0.05	-0.17*	0.06	-0.26*	0.06
Language student	-0.11	0.12	-0.04	0.12	-0.04	0.13	-0.15	0.14
Health student	0.16*	0.08	0.21*	0.08	0.19*	0.08	0.10	0.08
Law student	0.43*	0.20	0.43*	0.20	0.42*	0.20	0.26	0.19
Biological science student	0.20*	0.08	0.26*	0.09	0.25*	0.09	0.16*	0.08
Social sciences, Humanities student	0.13*	0.07	0.19*	0.07	0.14*	0.07	0.04	0.08
Art student	0.28*	0.13	0.34*	0.13	0.32*	0.14	0.20	0.13
Other student	0.03	0.06	0.06	0.06	0.03	0.06	-0.08	0.06
Not student or working	0.14*	0.06	0.15*	0.06	0.12*	0.05	0.12*	0.05
Missing	0.30*	0.07	0.32*	0.07	0.27*	0.07	0.28*	0.07
Maths ability at age 18	-	-	-0.08*	0.02	-0.07*	0.02	-0.05*	0.02
Race (Ref: White)								
American Indian or Alaska Native	-	-	-	-	0.08	0.08	0.06	0.08
Asian or Pacific Islander	-	-	-	-	0.05	0.08	0.06	0.07
Black, not Hispanic	-	-	-	-	0.22*	0.05	0.21*	0.05
Hispanic or Latino	-	-	-	-	0.12*	0.05	0.10*	0.05
More than one race	-	-	-	-	0.16*	0.08	0.15*	0.08
Missing	-	-	-	-	0.04	0.08	0.00	0.08
Family income parents reported when respondent was age 18 (Ref: Bottom quintile)								
2nd quintile	-	-	-	-	-0.03	0.05	-0.03	0.05
3rd quintile	-	-	-	-	-0.02	0.06	-0.01	0.06
4th quintile	-	-	-	-	0.07	0.05	0.08	0.05
Top quintile	-	-	-	-	0.01	0.06	0.03	0.06
Student at 20, who also held a part-time job (Ref: No)								
Yes	-	-	-	-	0.02	0.04	-0.02	0.04
College Dropout (ref: No)								
Yes	-	-	-	-	-	-	0.26*	0.05
Constant	0.42*	0.03	0.38*	0.03	0.36*	0.05	0.37*	0.05

Notes:
1 These results refer to when I use Rubenstein and Weiss (2007) NLSY estimates of wage growth (see Table 3.6) to predict NELS sample members age 30 income.
2 Source: Authors calculations from the NELS dataset. Sample size = 4,434. Dependent variable is the natural logarithm of respondents expected age 30 income divided by their predicted actual age 30 income

Table A3.14 Expected wages at age 30 in the NELS (raw figures as reported by students)

	N	Expected wage (mean)	Expected wage (median)
Education	122	39,749	35,500
Agriculture	51	44,000	35,000
Computer science/ Maths	123	46,852	40,000
Engineering/Physical sciences	397	48,101	45,000
Arts	105	49,600	40,000
Other	330	50,198	45,000
Languages	38	50,342	40,000
Journalism	81	51,604	45,000
Law	90	56,000	50,000
Business/Management	295	59,368	50,000
Social science/ Humanities	205	60,202	50,000
Health	90	61,428	50,000
Finance/Accounting	140	62,446	52,500
Biological sciences	147	64,242	50,000

Appendix 4.1. Full set of model parameter estimates

This appendix provides the full set of parameter estimates for the models presented in chapter four.

Model 1 -3

For models 1-3, I present the difference between the educational expectations of a hypothetical "advantaged" and "disadvantaged" child (in terms of log-odds). Recall that these hypothetical children differ in terms of highest parental education, highest parental occupation and number of books in the home as described in section 4.3.

Parental education and books in the home are dummy variables, where the reference group refers to characteristics of the disadvantaged child (less than 100 books, neither parent completed any more then compulsory schooling). One must sum the relevant coefficients (those on the "high" books and "high" education dummies) to form this part of the prediction of the difference between the hypothetical advantaged and disadvantaged children.

Adding in the contribution of parental occupation to these predictions is a little trickier. Recall that parental occupation is based upon the continuous ISEI index. Also recall that I define my hypothetical "advantaged" child as having a parent at the (national) 75th percentile of this continuous index, while a "disadvantaged" child is defined as having highest parental occupation at the (national) 25th percentile. Hence to add this into the predicted difference between "advantaged" and "disadvantaged" groups, one needs to know:

(1) How many ISEI points there are between the 25th and 75th national percentile
(2) How a one point increase in ISEI changes a child's expectations

Note that BOTH of these factors may differ across countries.

If the ISEI index had been entered as a single, simple linear term, one would simply multiply (1)*(2) to calculate the contribution parental occupation makes to the difference between advantaged and disadvantaged groups. However, I have entered the ISEI index as a series of piece-wise linear components, with a knot at the 50th percentile. Hence one must break point (1) and point (2) above into two further components:

(1a) How many ISEI points there are between the 25th and 50th percentile of the national ISEI distribution
(1b) How many ISEI points there are between the 50th and 75th percentile of the national ISEI distribution
(2a) How a one point change in ISEI between the 25th and 50th national percentile alters a child's expectations
(2b) How a one point change in ISEI between the 50th and 75th national percentile alters a child's expectations

Now, to calculate the contribution of occupation to the advantaged-disadvantaged expectation gap, once must sum {(1a*2a) + (1b*2b)}. Information on 1a and 1b (i.e. percentiles of the ISEI distribution) can be easily calculated from Table 4.3 (1a is, for example, simple the 50th percentile minus the 25th percentile- for England this is 51 - 35). Details on 2a and 2b are provided in the following Tables (e.g. under the label "occupation spline 26-50th percentile coefficient" for 2b).

To summerise, to get the difference between the expectations of advantaged and disadvantaged children for models 1-3, one must sum[161]:

Parental education "high" coefficient
+
Books in the home "high" coefficient
+
Occupation spline 26-50th percentile coefficient * Number of ISEI points between 25th and 50th percentile
+
Occupation spline 51-75th percentile coefficient * Number of ISEI points between 50th and 75th percentile

[161] The coefficients in the tables below refer to a one point increase in the ISEI index over the relevant range (e.g. so the coefficient for "Occupation spline 26-50th percentile" refers to how the log-odds of expecting to go to university change with a one point increase in the ISEI scale that occurs between the 26th and 50th national percentile).

A worked example for England is given below:

	Model 1	Model 2	Model 3
Coefficient on High (over 100) Books (Ref: Low)	0.803	0.437	0.404
+			
Coefficient on High (tertiary) Parental Education (Ref: Low)	1.300	1.295	0.678
+			
Occupation spline 26-50th percentile coefficient * Number of ISEI points between 25th and 50th percentile	0.025*16	0.012*16	0.010*16
+			
Occupation spline 51-75th percentile coefficient * Number of ISEI points between 51th and 75th percentile	0.034*15	0.026*15	0.006*15
=			
Difference (in log odds) between advantaged and disadvantaged children's expectations	3.014	2.314	1.332

1 Source: Author's calculations using PISA 2003 data. Sample size = 3,817. Dependent variable in regression was whether respondent expected to obtain a degree (coded 0 if they did not and 1 if they did)

Model 4

In model 4, I change specification (with details provided in section 4.5) and compare the predicted log-odds of expecting to complete university for an advantaged children of average ability to a disadvantaged child of high ability. Recall that, in this model, "advantaged" is based upon quartiles of the ESCS distribution described in section 4.5. Likewise, ability enters the model as a set of dummy variables reflecting quintiles of the (national) PISA math's ability distribution.

To calculate the predicted log-odds of an advantaged child of average ability expecting to go to university, one must sum the coefficients of the relevant dummy variables:

ESCS Highest Quartile "Advantaged" coefficient
+
Maths Ability third quintile coefficient
+
Constant

To calculate the predicted log-odds of a disadvantaged child of high ability expecting to go to university, one must sum the coefficients:

Maths Ability Highest quintile coefficient
+
Constant

An example for England is given below (figures correspond to those presented in Figure 4.7 and Table 4.10):

Advantaged child of average ability

	Model 4 Advantaged
ESCS (Ref: Lowest Quartile "Disadvantaged")	
Highest Quartile "Advantaged"	1.838
Maths Ability (Ref: Lowest Ability Quintile)	
Third Quartile	1.214
Constant	-2.898
Predicted log odds for average ability, advantaged child	0.154

Disadvantaged child of high ability

	Model 4 Disadvantaged
Maths Ability (Ref: Lowest Ability Quintile)	
Highest Quartile	2.449
Constant	-2.898
Predicted log odds for high ability, disadvantaged child	-0.449

Notes:
1 Source: Author’s calculations using PISA 2003 data. Sample size = 3,817. Dependent variable in regression was whether respondent expected to obtain a degree (coded 0 if they did not and 1 if they did)

Australia

	Model 1		Model 2		Model 3	
	Beta	**SE**	**Beta**	**SE**	**Beta**	**SE**
Gender (Ref: Girl)						
Boy	-0.665	0.061	-0.809	0.063	-0.866	0.058
Books (Ref: Low)						
High	0.745	0.062	0.458	0.065	0.424	0.066
Missing	-0.050	0.232	0.101	0.256	0.021	0.250
Parental Education (Ref: Low)						
Medium	0.334	0.072	0.330	0.076	0.275	0.092
High	1.086	0.085	0.988	0.087	0.720	0.078
Missing	-0.777	0.310	-0.592	0.325	-0.469	0.320
Occupation spline 0-10th percentile	0.021	0.016	0.017	0.016	0.012	0.016
Occupation spline 11-25th percentile	0.031	0.010	0.020	0.010	0.015	0.011
Occupation spline 26-50th percentile	0.011	0.012	-0.001	0.012	-0.014	0.013
Occupation spline 50-75th percentile	0.036	0.011	0.029	0.012	0.033	0.011
Occupation spline 76-90th percentile	-0.185	0.163	-0.183	0.178	-0.201	0.173
Occupation spline 91-100th percentile	0.012	0.013	0.013	0.014	0.004	0.014
Immigrant Status (Ref: Native)						
Immigrant	1.040	0.584	1.160	0.595	0.709	0.591
Books*Immigrant						
High Books, Immigrant	-0.493	0.093	-0.417	0.099	-0.358	0.104
Missing Books, Immigrant	-0.866	0.411	-0.918	0.442	-0.800	0.443
Parental Education*Immigrant						
Medium Education, Immigrant	-0.330	0.112	-0.372	0.120	-0.391	0.150
High Education, Immigrant	0.146	0.134	0.175	0.139	0.074	0.130
Missing Education, Immigrant	0.629	0.416	0.677	0.435	0.747	0.447
Occupation 0-10 * Immigrant	0.007	0.021	0.003	0.021	0.002	0.021
Occupation 11-25 * Immigrant	-0.044	0.015	-0.042	0.016	-0.036	0.017
Occupation 26-50 * Immigrant	0.029	0.020	0.022	0.020	0.039	0.021
Occupation 51-75 * Immigrant	-0.034	0.016	-0.031	0.018	-0.026	0.018
Occupation 76-90 * Immigrant	0.345	0.253	0.267	0.284	0.304	0.280
Occupation 91-100 * Immigrant	0.020	0.021	0.009	0.023	0.012	0.023
Ability spline 0-10th percentile	-	-	0.007	0.002	0.007	0.002
Ability spline 11-25th percentile	-	-	0.005	0.002	0.007	0.002
Ability spline 26-50th percentile	-	-	0.009	0.002	0.010	0.002
Ability spline 50-75th percentile	-	-	0.010	0.002	0.009	0.002
Ability spline 76-90th percentile	-	-	0.012	0.003	0.013	0.003
Ability spline 91-100th percentile	-	-	0.015	0.003	0.013	0.003
School FE	-	-			Yes	Yes

1 Source: Author's calculations using PISA 2003 data. Sample size = 12,492 in models 1,2, 4 and 5. Sample size = 11,794 in model 3 where some observations are dropped due to the fixed effect perfectly predicting the response. Dependent variable in all models was whether respondent expected to obtain a degree (coded 0 if they did not and 1 if they did)

	Model 4	
	Beta	SE
Gender (Ref: Girl)		
Boy	-0.770	0.058
ESCS (Ref: Lowest Quartile "Disadvantaged")		
Second Quartile	0.571	0.109
Third Quartile	1.081	0.084
Highest Quartile "Advantaged"	1.768	0.110
Maths Ability (Ref: Lowest Ability Quintile)		
Second Quintile	0.463	0.069
Third Quintile	1.001	0.080
Fourth Quintile	1.483	0.097
Highest Quintile	2.348	0.100
Immigrant Status (Ref: Native)		
Immigrant	0.873	0.127
ESCS*Immigrant		
Second Quartile, Immigrant	-0.300	0.148
Third Quartile, Immigrant	-0.388	0.165
Highest Quartile, Immigrant	-0.356	0.190
Constant	-1.081	0.083

	Model 5	
	Beta	SE
Gender (Ref: Girl)		
Boy	-0.411	0.063
ESCS (Ref: Lowest Quartile "Disadvantaged")		
Second Quartile	0.564	0.113
Third Quartile	1.035	0.087
Highest Quartile	1.682	0.108
Maths Ability	0.934	0.079
ESCS*Maths Ability		
Second ESCS Quartile*Maths Ability	0.028	0.103
Third ESCS Quartile*Maths Ability	0.085	0.118
Highest ESCS Quartile*Maths Ability	0.104	0.128
Constant	-0.261	0.083

Austria

	Model 1		Model 2		Model 3	
	Beta	**SE**	**Beta**	**SE**	**Beta**	**SE**
Gender (Ref: Girl)						
Boy	-0.095	0.125	-0.337	0.118	0.073	0.105
Books (Ref: Low)						
High	0.939	0.107	0.524	0.108	0.247	0.114
Missing	-0.123	0.610	-0.278	0.674	-0.384	0.716
Parental Education (Ref: Low)						
Medium	0.340	0.088	0.479	0.090	0.402	0.205
High	1.145	0.140	1.295	0.149	0.500	0.094
Missing	0.027	0.617	0.510	0.761	0.183	0.668
Occupation spline 0-10th percentile	-0.025	0.036	-0.047	0.042	-0.052	0.038
Occupation spline 11-25th percentile	-0.024	0.044	-0.006	0.047	-0.064	0.051
Occupation spline 26-50th percentile	0.070	0.025	0.037	0.025	0.034	0.028
Occupation spline 50-75th percentile	0.018	0.017	0.013	0.017	-0.005	0.020
Occupation spline 76-90th percentile	0.034	0.013	0.019	0.014	0.020	0.015
Occupation spline 91-100th percentile	0.019	0.012	0.030	0.014	0.032	0.014
Immigrant Status (Ref: Native)						
Immigrant	1.992	1.368	2.255	1.424	0.892	1.530
Books*Immigrant						
High Books, Immigrant	-0.019	0.245	-0.088	0.271	-0.065	0.283
Missing Books, Immigrant	-0.182	0.858	-0.371	0.811	-0.113	0.903
Parental Education*Immigrant						
Medium Education, Immigrant	-0.084	0.252	-0.296	0.254	-0.591	0.502
High Education, Immigrant	0.159	0.289	-0.114	0.309	-0.093	0.273
Missing Education, Immigrant	-0.432	0.969	-0.726	1.169	-0.693	1.181
Occupation 0-10 * Immigrant	-0.052	0.059	-0.044	0.064	-0.017	0.069
Occupation 11-25 * Immigrant	0.004	0.081	-0.010	0.087	0.039	0.105
Occupation 26-50 * Immigrant	-0.053	0.042	-0.040	0.043	-0.056	0.053
Occupation 51-75 * Immigrant	0.046	0.044	0.053	0.050	0.089	0.046
Occupation 76-90 * Immigrant	-0.013	0.035	-0.008	0.040	-0.018	0.036
Occupation 91-100 * Immigrant	-0.027	0.033	-0.037	0.041	-0.022	0.042
Ability spline 0-10th percentile	-	-	-0.009	0.006	-0.011	0.006
Ability spline 11-25th percentile	-	-	0.007	0.006	0.000	0.007
Ability spline 26-50th percentile	-	-	0.014	0.003	0.005	0.004
Ability spline 50-75th percentile	-	-	0.016	0.002	0.011	0.003
Ability spline 76-90th percentile	-	-	0.009	0.003	0.005	0.003
Ability spline 91-100th percentile	-	-	0.010	0.002	0.009	0.002
School FE	-	-			Yes	Yes

1 Source: Author's calculations using PISA 2003 data. Sample size = 4,558 in models 1,2, 4 and 5. Sample size = 3,530 in model 3 where some observations are dropped due to the fixed effect perfectly predicting the response. Dependent variable in all models was whether respondent expected to obtain a degree (coded 0 if they did not and 1 if they did)

	Model 4	
	Beta	**SE**
Gender (Ref: Girl)		
Boy	-0.312	0.121
ESCS (Ref: Lowest Quartile "Disadvantaged")		
Second Quartile	0.518	0.165
Third Quartile	1.082	0.177
Highest Quartile "Advantaged"	2.108	0.170
Maths Ability (Ref: Lowest Ability Quintile)		
Second Quintile	0.481	0.221
Third Quintile	1.207	0.201
Fourth Quintile	1.891	0.194
Highest Quintile	2.754	0.191
Immigrant Status (Ref: Native)		
Immigrant	0.864	0.263
ESCS*Immigrant		
Second Quartile, Immigrant	-0.085	0.347
Third Quartile, Immigrant	-0.137	0.306
Highest Quartile, Immigrant	-0.140	0.325
Constant	-3.772	0.238

	Model 5	
	Beta	**SE**
Gender (Ref: Girl)		
Boy	0.314	0.129
ESCS (Ref: Lowest Quartile "Disadvantaged")		
Second Quartile	0.359	0.168
Third Quartile	0.950	0.165
Highest Quartile	1.898	0.163
Maths Ability	1.003	0.248
ESCS*Maths Ability		
Second ESCS Quartile*Maths Ability	0.183	0.276
Third ESCS Quartile*Maths Ability	0.002	0.269
Highest ESCS Quartile*Maths Ability	-0.013	0.266
Constant	-2.526	0.158

Belgium (French)

	Model 1		Model 2		Model 3	
	Beta	**SE**	**Beta**	**SE**	**Beta**	**SE**
Gender (Ref: Girl)						
Boy	-0.293	0.118	-0.421	0.114	-0.212	0.116
Books (Ref: Low)						
High	0.983	0.136	0.685	0.142	0.689	0.150
Missing	0.004	0.570	0.642	0.568	0.639	0.613
Parental Education (Ref: Low)						
Medium	0.545	0.175	0.426	0.190	0.201	0.276
High	0.878	0.192	0.852	0.196	0.567	0.187
Missing	-1.000	0.382	-0.571	0.412	-0.415	0.455
Occupation spline 0-10th percentile	0.027	0.046	0.026	0.053	0.030	0.059
Occupation spline 11-25th percentile	0.039	0.041	0.029	0.044	0.025	0.045
Occupation spline 26-50th percentile	0.024	0.020	0.007	0.022	-0.011	0.021
Occupation spline 50-75th percentile	0.011	0.021	0.017	0.020	0.030	0.021
Occupation spline 76-90th percentile	0.079	0.089	0.006	0.090	-0.050	0.090
Occupation spline 91-100th percentile	0.005	0.018	-0.003	0.019	-0.007	0.020
Immigrant Status (Ref: Native)						
Immigrant	0.365	1.535	1.159	1.687	1.250	1.830
Books*Immigrant						
High Books, Immigrant	-0.443	0.218	-0.485	0.242	-0.454	0.253
Missing Books, Immigrant	-0.693	1.008	-0.436	0.974	0.201	0.951
Parental Education*Immigrant						
Medium Education, Immigrant	0.259	0.257	0.256	0.268	0.663	0.420
High Education, Immigrant	0.224	0.276	0.162	0.289	0.314	0.270
Missing Education, Immigrant	0.101	0.653	-0.235	0.671	-0.351	0.697
Occupation 0-10 * Immigrant	0.002	0.057	-0.012	0.063	-0.030	0.068
Occupation 11-25 * Immigrant	-0.027	0.059	-0.023	0.063	0.000	0.064
Occupation 26-50 * Immigrant	-0.024	0.030	-0.034	0.032	-0.028	0.033
Occupation 51-75 * Immigrant	0.028	0.036	0.016	0.039	0.011	0.041
Occupation 76-90 * Immigrant	-0.105	0.169	-0.053	0.182	-0.056	0.189
Occupation 91-100 * Immigrant	0.013	0.032	0.024	0.034	0.031	0.033
Ability spline 0-10th percentile	-	-	0.006	0.009	-0.002	0.011
Ability spline 11-25th percentile	-	-	0.007	0.005	0.006	0.005
Ability spline 26-50th percentile	-	-	0.015	0.003	0.013	0.003
Ability spline 50-75th percentile	-	-	0.006	0.003	0.003	0.002
Ability spline 76-90th percentile	-	-	0.006	0.003	0.006	0.003
Ability spline 91-100th percentile	-	-	0.014	0.003	0.011	0.003
School FE	-	-			Yes	Yes

1 Source: Author’s calculations using PISA 2003 data. Sample size = 2,931 in models 1,2, 4 and 5. Sample size = 2,598 in model 3 where some observations are dropped due to the fixed effect perfectly predicting the response. Dependent variable in all models was whether respondent expected to obtain a degree (coded 0 if they did not and 1 if they did)

	Model 4	
	Beta	**SE**
Gender (Ref: Girl)		
Boy	-0.339	0.113
ESCS (Ref: Lowest Quartile "Disadvantaged")		
Second Quartile	0.660	0.215
Third Quartile	1.117	0.206
Highest Quartile "Advantaged"	1.875	0.206
Maths Ability (Ref: Lowest Ability Quintile)		
Second Quintile	0.936	0.246
Third Quintile	1.607	0.236
Fourth Quintile	2.051	0.252
Highest Quintile	2.711	0.241
Immigrant Status (Ref: Native)		
Immigrant	0.606	0.263
ESCS*Immigrant		
Second Quartile, Immigrant	-0.370	0.354
Third Quartile, Immigrant	-0.108	0.312
Highest Quartile, Immigrant	-0.793	0.326
Constant	-3.444	0.292

	Model 5	
	Beta	**SE**
Gender (Ref: Girl)		
Boy	-0.005	0.122
ESCS (Ref: Lowest Quartile "Disadvantaged")		
Second Quartile	0.588	0.194
Third Quartile	1.075	0.191
Highest Quartile	1.773	0.189
Maths Ability	0.803	0.224
ESCS*Maths Ability		
Second ESCS Quartile*Maths Ability	0.235	0.249
Third ESCS Quartile*Maths Ability	0.167	0.281
Highest ESCS Quartile*Maths Ability	0.118	0.264
Constant	-1.856	0.173

Belgium (Flemish)

	Model 1		Model 2		Model 3	
	Beta	SE	Beta	SE	Beta	SE
Gender (Ref: Girl)						
Boy	-0.419	0.107	-0.742	0.097	-0.430	0.093
Books (Ref: Low)						
High	0.744	0.082	0.567	0.092	0.507	0.100
Missing	-0.753	0.395	-0.215	0.501	0.020	0.462
Parental Education (Ref: Low)						
Medium	0.162	0.105	0.262	0.107	0.180	0.133
High	0.963	0.110	1.039	0.120	0.650	0.112
Missing	-1.519	0.369	-0.825	0.394	-0.713	0.414
Occupation spline 0-10th percentile	0.031	0.035	0.006	0.039	0.011	0.039
Occupation spline 11-25th percentile	0.011	0.035	-0.013	0.037	-0.037	0.039
Occupation spline 26-50th percentile	0.050	0.012	0.030	0.012	0.015	0.012
Occupation spline 50-75th percentile	0.026	0.011	0.011	0.013	0.020	0.013
Occupation spline 76-90th percentile	0.122	0.064	0.131	0.072	0.080	0.076
Occupation spline 91-100th percentile	-0.011	0.013	-0.010	0.015	0.003	0.015
Immigrant Status (Ref: Native)						
Immigrant	-0.346	1.658	0.229	2.243	-0.314	2.426
Books*Immigrant						
High Books, Immigrant	-0.039	0.194	-0.377	0.236	-0.228	0.246
Missing Books, Immigrant	1.732	0.911	1.919	1.232	1.437	1.684
Parental Education*Immigrant						
Medium Education, Immigrant	0.123	0.283	0.154	0.331	0.104	0.462
High Education, Immigrant	0.097	0.281	0.330	0.326	0.528	0.337
Missing Education, Immigrant	-0.320	0.877	-0.221	0.834	0.288	0.838
Occupation 0-10 * Immigrant	0.029	0.064	0.034	0.084	0.038	0.088
Occupation 11-25 * Immigrant	-0.056	0.073	-0.041	0.088	-0.048	0.095
Occupation 26-50 * Immigrant	-0.007	0.028	-0.010	0.030	0.012	0.033
Occupation 51-75 * Immigrant	-0.046	0.026	-0.046	0.032	-0.055	0.034
Occupation 76-90 * Immigrant	0.222	0.186	0.240	0.221	0.289	0.234
Occupation 91-100 * Immigrant	0.019	0.036	-0.014	0.040	-0.042	0.037
Ability spline 0-10th percentile	-	-	0.003	0.006	0.003	0.007
Ability spline 11-25th percentile	-	-	0.011	0.004	0.006	0.005
Ability spline 26-50th percentile	-	-	0.016	0.002	0.011	0.003
Ability spline 50-75th percentile	-	-	0.017	0.003	0.013	0.003
Ability spline 76-90th percentile	-	-	0.005	0.003	0.001	0.003
Ability spline 91-100th percentile	-	-	0.025	0.005	0.023	0.005
School FE	-	-			Yes	Yes

1 Source: Author's calculations using PISA 2003 data. Sample size = 5,696 in models 1,2, 4 and 5. Sample size = 5,279 in model 3 where some observations are dropped due to the fixed effect perfectly predicting the response. Dependent variable in all models was whether respondent expected to obtain a degree (coded 0 if they did not and 1 if they did)

	Model 4	
	Beta	**SE**
Gender (Ref: Girl)		
Boy	-0.677	0.099
ESCS (Ref: Lowest Quartile "Disadvantaged")		
Second Quartile	0.669	0.137
Third Quartile	1.358	0.134
Highest Quartile "Advantaged"	2.149	0.150
Maths Ability (Ref: Lowest Ability Quintile)		
Second Quintile	0.986	0.186
Third Quintile	1.914	0.172
Fourth Quintile	2.768	0.165
Highest Quintile	3.480	0.177
Immigrant Status (Ref: Native)		
Immigrant	1.154	0.220
ESCS*Immigrant		
Second Quartile, Immigrant	-0.404	0.298
Third Quartile, Immigrant	-0.269	0.333
Highest Quartile, Immigrant	-0.593	0.307
Constant	-3.528	0.185

	Model 5	
	Beta	**SE**
Gender (Ref: Girl)		
Boy	-0.161	0.105
ESCS (Ref: Lowest Quartile "Disadvantaged")		
Second Quartile	0.924	0.139
Third Quartile	1.617	0.141
Highest Quartile	2.322	0.149
Maths Ability	1.619	0.142
ESCS*Maths Ability		
Second ESCS Quartile*Maths Ability	-0.288	0.175
Third ESCS Quartile*Maths Ability	-0.449	0.179
Highest ESCS Quartile*Maths Ability	-0.336	0.184
Constant	-2.070	0.138

Canada

	Model 1		Model 2		Model 3	
	Beta	**SE**	**Beta**	**SE**	**Beta**	**SE**
Gender (Ref: Girl)						
Boy	-0.680	0.047	-0.847	0.048	-0.867	0.054
Books (Ref: Low)						
High	0.566	0.055	0.380	0.060	0.328	0.061
Missing	-0.043	0.130	-0.165	0.142	-0.048	0.145
Parental Education (Ref: Low)						
Medium	0.290	0.061	0.213	0.062	0.123	0.083
High	1.163	0.080	1.136	0.082	0.683	0.070
Missing	-0.047	0.307	0.372	0.314	0.261	0.313
Occupation spline 0-10th percentile	-0.012	0.018	-0.017	0.019	-0.005	0.019
Occupation spline 11-25th percentile	0.032	0.009	0.021	0.009	0.015	0.010
Occupation spline 26-50th percentile	0.016	0.009	0.011	0.009	0.007	0.010
Occupation spline 50-75th percentile	0.025	0.009	0.020	0.009	0.028	0.010
Occupation spline 76-90th percentile	0.033	0.044	-0.007	0.046	0.006	0.049
Occupation spline 91-100th percentile	-0.007	0.012	0.000	0.013	0.000	0.015
Immigrant Status (Ref: Native)						
Immigrant	0.640	0.922	0.909	0.910	1.178	0.954
Books*Immigrant						
High Books, Immigrant	-0.269	0.121	-0.345	0.124	-0.293	0.129
Missing Books, Immigrant	-0.110	0.290	0.060	0.275	-0.006	0.279
Parental Education*Immigrant						
Medium Education, Immigrant	-0.176	0.144	-0.043	0.140	-0.020	0.199
High Education, Immigrant	-0.143	0.179	-0.098	0.183	-0.044	0.157
Missing Education, Immigrant	0.605	0.429	0.331	0.453	0.267	0.448
Occupation 0-10 * Immigrant	0.023	0.033	0.014	0.032	-0.001	0.035
Occupation 11-25 * Immigrant	-0.024	0.019	-0.015	0.019	-0.002	0.020
Occupation 26-50 * Immigrant	-0.029	0.022	-0.045	0.022	-0.049	0.023
Occupation 51-75 * Immigrant	0.009	0.019	0.021	0.020	0.016	0.021
Occupation 76-90 * Immigrant	0.065	0.095	0.040	0.099	0.041	0.096
Occupation 91-100 * Immigrant	-0.003	0.025	-0.007	0.026	-0.019	0.025
Ability spline 0-10th percentile	-	-	0.003	0.002	0.004	0.002
Ability spline 11-25th percentile	-	-	0.007	0.002	0.007	0.002
Ability spline 26-50th percentile	-	-	0.010	0.002	0.011	0.002
Ability spline 50-75th percentile	-	-	0.009	0.002	0.010	0.002
Ability spline 76-90th percentile	-	-	0.009	0.003	0.010	0.003
Ability spline 91-100th percentile	-	-	0.007	0.003	0.008	0.003
School FE	-	-			Yes	Yes

1 Source: Author's calculations using PISA 2003 data. Sample size = 26,707 in models 1,2, 4 and 5. Sample size = 25,619 in model 3 where some observations are dropped due to the fixed effect perfectly predicting the response. Dependent variable in all models was whether respondent expected to obtain a degree (coded 0 if they did not and 1 if they did)

	Model 4	
	Beta	**SE**
Gender (Ref: Girl)		
Boy	-0.755	0.047
ESCS (Ref: Lowest Quartile "Disadvantaged")		
Second Quartile	0.624	0.067
Third Quartile	1.304	0.068
Highest Quartile "Advantaged"	1.880	0.114
Maths Ability (Ref: Lowest Ability Quintile)		
Second Quintile	0.401	0.065
Third Quintile	0.902	0.077
Fourth Quintile	1.260	0.075
Highest Quintile	1.900	0.088
Immigrant Status (Ref: Native)		
Immigrant	1.083	0.114
ESCS*Immigrant		
Second Quartile, Immigrant	-0.269	0.130
Third Quartile, Immigrant	-0.490	0.153
Highest Quartile, Immigrant	-0.342	0.222
Constant	-0.997	0.076

	Model 5	
	Beta	**SE**
Gender (Ref: Girl)		
Boy	-0.452	0.055
ESCS (Ref: Lowest Quartile "Disadvantaged")		
Second Quartile	0.619	0.065
Third Quartile	1.306	0.068
Highest Quartile	1.919	0.118
Maths Ability	0.564	0.055
ESCS*Maths Ability		
Second ESCS Quartile*Maths Ability	-0.016	0.073
Third ESCS Quartile*Maths Ability	0.119	0.081
Highest ESCS Quartile*Maths Ability	0.044	0.120
Constant	-0.367	0.060

Czech

	Model 1		Model 2		Model 3	
	Beta	**SE**	**Beta**	**SE**	**Beta**	**SE**
Gender (Ref: Girl)						
Boy	-0.511	0.088	-0.880	0.095	-0.688	0.095
Books (Ref: Low)						
High	0.810	0.085	0.367	0.087	0.347	0.091
Missing	-0.253	0.332	-0.365	0.339	-0.286	0.360
Parental Education (Ref: Low)						
Medium	0.152	0.113	0.163	0.136	0.147	0.171
High	1.044	0.098	0.957	0.119	0.704	0.129
Missing	-0.868	0.412	-0.469	0.497	-0.699	0.516
Occupation spline 0-10th percentile	0.039	0.038	0.006	0.042	0.015	0.046
Occupation spline 11-25th percentile	0.055	0.022	0.048	0.022	0.037	0.024
Occupation spline 26-50th percentile	0.047	0.013	0.031	0.015	0.025	0.015
Occupation spline 50-75th percentile	0.018	0.030	0.015	0.032	0.024	0.035
Occupation spline 76-90th percentile	0.026	0.011	0.015	0.013	0.021	0.015
Occupation spline 91-100th percentile	0.025	0.013	0.023	0.015	0.010	0.017
Immigrant Status (Ref: Native)						
Immigrant	-2.463	2.665	-3.414	2.434	-1.305	2.570
Books*Immigrant						
High Books, Immigrant	-0.746	0.331	-0.565	0.335	-0.542	0.394
Missing Books, Immigrant	0.497	1.341	1.181	0.852	0.401	0.871
Parental Education*Immigrant						
Medium Education, Immigrant	0.273	0.538	0.174	0.536	0.092	0.651
High Education, Immigrant	0.284	0.313	0.273	0.338	0.409	0.406
Missing Education, Immigrant	1.244	1.122	1.748	1.192	1.445	1.164
Occupation 0-10 * Immigrant	0.105	0.092	0.141	0.083	0.066	0.090
Occupation 11-25 * Immigrant	-0.049	0.063	-0.063	0.070	-0.065	0.081
Occupation 26-50 * Immigrant	-0.052	0.044	-0.037	0.050	-0.016	0.053
Occupation 51-75 * Immigrant	0.066	0.103	-0.040	0.109	-0.056	0.111
Occupation 76-90 * Immigrant	0.022	0.045	0.056	0.045	0.066	0.050
Occupation 91-100 * Immigrant	-0.059	0.041	-0.089	0.062	-0.134	0.062
Ability spline 0-10th percentile	-	-	0.013	0.008	0.012	0.009
Ability spline 11-25th percentile	-	-	0.015	0.006	0.017	0.006
Ability spline 26-50th percentile	-	-	0.012	0.003	0.012	0.003
Ability spline 50-75th percentile	-	-	0.012	0.002	0.010	0.002
Ability spline 76-90th percentile	-	-	0.020	0.003	0.017	0.004
Ability spline 91-100th percentile	-	-	0.010	0.004	0.008	0.005
School FE	-	-			Yes	Yes

1 Source: Author's calculations using PISA 2003 data. Sample size = 6,076 in models 1,2, 4 and 5. Sample size = 5,398 in model 3 where some observations are dropped due to the fixed effect perfectly predicting the response. Dependent variable in all models was whether respondent expected to obtain a degree (coded 0 if they did not and 1 if they did)

	Model 4	
	Beta	SE
Gender (Ref: Girl)		
Boy	-0.898	0.093
ESCS (Ref: Lowest Quartile "Disadvantaged")		
Second Quartile	0.507	0.116
Third Quartile	1.124	0.114
Highest Quartile "Advantaged"	2.063	0.127
Maths Ability (Ref: Lowest Ability Quintile)		
Second Quintile	0.876	0.185
Third Quintile	1.534	0.176
Fourth Quintile	2.221	0.188
Highest Quintile	3.345	0.192
Immigrant Status (Ref: Native)		
Immigrant	0.239	0.323
ESCS*Immigrant		
Second Quartile, Immigrant	-0.475	0.467
Third Quartile, Immigrant	-0.128	0.443
Highest Quartile, Immigrant	-0.477	0.465
Constant	-2.849	0.188

	Model 5	
	Beta	SE
Gender (Ref: Girl)		
Boy	-0.282	0.091
ESCS (Ref: Lowest Quartile "Disadvantaged")		
Second Quartile	0.491	0.115
Third Quartile	1.096	0.114
Highest Quartile	2.113	0.127
Maths Ability	1.489	0.161
ESCS*Maths Ability		
Second ESCS Quartile*Maths Ability	-0.349	0.180
Third ESCS Quartile*Maths Ability	-0.335	0.183
Highest ESCS Quartile*Maths Ability	-0.419	0.193
Constant	-1.336	0.118

Denmark

	Model 1		Model 2		Model 3	
	Beta	**SE**	**Beta**	**SE**	**Beta**	**SE**
Gender (Ref: Girl)						
Boy	-0.191	0.077	-0.324	0.075	-0.259	0.076
Books (Ref: Low)						
High	0.615	0.098	0.392	0.100	0.426	0.102
Missing	0.157	0.347	0.094	0.352	0.050	0.324
Parental Education (Ref: Low)						
Medium	0.553	0.123	0.413	0.126	0.286	0.170
High	1.411	0.147	1.375	0.148	0.665	0.135
Missing	-0.593	0.436	-0.295	0.426	-0.386	0.380
Occupation spline 0-10th percentile	-0.039	0.037	-0.049	0.036	-0.048	0.035
Occupation spline 11-25th percentile	0.002	0.028	0.000	0.028	0.000	0.028
Occupation spline 26-50th percentile	0.038	0.017	0.025	0.017	0.027	0.018
Occupation spline 50-75th percentile	0.014	0.027	0.013	0.028	0.005	0.031
Occupation spline 76-90th percentile	0.007	0.023	-0.010	0.024	0.008	0.025
Occupation spline 91-100th percentile	0.017	0.014	0.013	0.015	0.029	0.017
Immigrant Status (Ref: Native)						
Immigrant	-2.503	1.855	-1.334	1.938	-2.454	2.140
Books*Immigrant						
High Books, Immigrant	-0.495	0.232	-0.568	0.244	-0.438	0.246
Missing Books, Immigrant	-0.270	0.780	-0.090	0.805	0.156	0.742
Parental Education*Immigrant						
Medium Education, Immigrant	-0.184	0.296	-0.323	0.303	-0.577	0.552
High Education, Immigrant	-0.397	0.339	-0.426	0.370	0.018	0.339
Missing Education, Immigrant	-0.196	0.886	-0.348	0.975	0.366	0.891
Occupation 0-10 * Immigrant	0.136	0.071	0.107	0.074	0.136	0.079
Occupation 11-25 * Immigrant	0.037	0.060	0.040	0.064	0.011	0.065
Occupation 26-50 * Immigrant	-0.123	0.041	-0.127	0.044	-0.133	0.045
Occupation 51-75 * Immigrant	0.142	0.064	0.149	0.073	0.227	0.072
Occupation 76-90 * Immigrant	-0.033	0.052	-0.024	0.058	-0.066	0.058
Occupation 91-100 * Immigrant	-0.046	0.032	-0.039	0.035	-0.052	0.039
Ability spline 0-10th percentile	-	-	-0.001	0.004	-0.001	0.004
Ability spline 11-25th percentile	-	-	0.017	0.005	0.016	0.005
Ability spline 26-50th percentile	-	-	0.005	0.003	0.006	0.003
Ability spline 50-75th percentile	-	-	0.011	0.003	0.010	0.003
Ability spline 76-90th percentile	-	-	0.009	0.003	0.010	0.003
Ability spline 91-100th percentile	-	-	0.000	0.003	-0.001	0.003
School FE	-	-			Yes	Yes

1 Source: Author's calculations using PISA 2003 data. Sample size = 4,191 in models 1,2, 4 and 5. Sample size = 4,033 in model 3 where some observations are dropped due to the fixed effect perfectly predicting the response. Dependent variable in all models was whether respondent expected to obtain a degree (coded 0 if they did not and 1 if they did)

	Model 4	
	Beta	**SE**
Gender (Ref: Girl)		
Boy	-0.274	0.075
ESCS (Ref: Lowest Quartile "Disadvantaged")		
Second Quartile	0.263	0.171
Third Quartile	0.831	0.153
Highest Quartile "Advantaged"	1.669	0.153
Maths Ability (Ref: Lowest Ability Quintile)		
Second Quintile	0.790	0.162
Third Quintile	1.018	0.169
Fourth Quintile	1.595	0.162
Highest Quintile	1.991	0.157
Immigrant Status (Ref: Native)		
Immigrant	1.624	0.249
ESCS*Immigrant		
Second Quartile, Immigrant	-0.505	0.321
Third Quartile, Immigrant	-1.282	0.339
Highest Quartile, Immigrant	-1.111	0.343
Constant	-3.105	0.187

	Model 5	
	Beta	**SE**
Gender (Ref: Girl)		
Boy	0.061	0.084
ESCS (Ref: Lowest Quartile "Disadvantaged")		
Second Quartile	0.315	0.172
Third Quartile	0.753	0.162
Highest Quartile	1.747	0.159
Maths Ability	1.010	0.173
ESCS*Maths Ability		
Second ESCS Quartile*Maths Ability	-0.039	0.192
Third ESCS Quartile*Maths Ability	-0.270	0.197
Highest ESCS Quartile*Maths Ability	-0.399	0.197
Constant	-2.176	0.138

England

	Model 1		Model 2		Model 3	
	Beta	**SE**	**Beta**	**SE**	**Beta**	**SE**
Gender (Ref: Girl)						
Boy	-0.470	0.098	-0.605	0.098	-0.617	0.103
Books (Ref: Low)						
High	0.803	0.107	0.437	0.103	0.404	0.108
Missing	0.363	0.512	0.408	0.489	0.235	0.464
Parental Education (Ref: Low)						
Medium	0.041	0.129	0.135	0.131	0.109	0.169
High	1.300	0.141	1.295	0.144	0.678	0.130
Missing	-0.531	0.350	-0.184	0.350	-0.125	0.344
Occupation spline 0-10th percentile	0.071	0.032	0.076	0.036	0.102	0.038
Occupation spline 11-25th percentile	-0.003	0.048	-0.034	0.048	-0.031	0.048
Occupation spline 26-50th percentile	0.025	0.012	0.012	0.013	0.010	0.013
Occupation spline 50-75th percentile	0.035	0.017	0.027	0.018	0.007	0.017
Occupation spline 76-90th percentile	-0.015	0.049	-0.049	0.053	0.052	0.053
Occupation spline 91-100th percentile	0.017	0.015	0.021	0.017	0.017	0.018
Immigrant Status (Ref: Native)						
Immigrant	4.726	2.093	7.297	2.288	7.415	2.186
Books*Immigrant						
High Books, Immigrant	-0.493	0.223	-0.719	0.248	-0.434	0.246
Missing Books, Immigrant	-0.146	0.888	0.419	1.122	0.200	1.191
Parental Education*Immigrant						
Medium Education, Immigrant	0.036	0.276	0.141	0.317	0.125	0.363
High Education, Immigrant	0.106	0.275	0.103	0.318	0.203	0.270
Missing Education, Immigrant	0.594	0.605	0.268	0.658	-0.040	0.669
Occupation 0-10 * Immigrant	-0.119	0.078	-0.206	0.085	-0.226	0.082
Occupation 11-25 * Immigrant	0.100	0.111	0.086	0.129	0.089	0.134
Occupation 26-50 * Immigrant	-0.056	0.026	-0.040	0.031	-0.045	0.033
Occupation 51-75 * Immigrant	-0.035	0.029	-0.033	0.034	0.009	0.035
Occupation 76-90 * Immigrant	0.011	0.086	-0.005	0.093	-0.108	0.101
Occupation 91-100 * Immigrant	-0.024	0.028	-0.028	0.034	-0.017	0.036
Ability spline 0-10th percentile	-	-	0.024	0.008	0.026	0.009
Ability spline 11-25th percentile	-	-	-0.005	0.005	-0.007	0.006
Ability spline 26-50th percentile	-	-	0.019	0.003	0.021	0.003
Ability spline 50-75th percentile	-	-	0.005	0.003	0.005	0.003
Ability spline 76-90th percentile	-	-	0.014	0.003	0.012	0.003
Ability spline 91-100th percentile	-	-	0.014	0.005	0.014	0.005
School FE	-	-			Yes	Yes

1 Source: Author's calculations using PISA 2003 data. Sample size = 3,817 in models 1,2, 4 and 5. Sample size = 3,655 in model 3 where some observations are dropped due to the fixed effect perfectly predicting the response. Dependent variable in all models was whether respondent expected to obtain a degree (coded 0 if they did not and 1 if they did)

	Model 4	
	Beta	**SE**
Gender (Ref: Girl)		
Boy	-0.569	0.096
ESCS (Ref: Lowest Quartile "Disadvantaged")		
Second Quartile	0.419	0.161
Third Quartile	0.762	0.161
Highest Quartile "Advantaged"	1.838	0.138
Maths Ability (Ref: Lowest Ability Quintile)		
Second Quintile	0.252	0.198
Third Quintile	1.214	0.180
Fourth Quintile	1.550	0.178
Highest Quintile	2.449	0.183
Immigrant Status (Ref: Native)		
Immigrant	1.620	0.226
ESCS*Immigrant		
Second Quartile, Immigrant	-0.510	0.340
Third Quartile, Immigrant	-0.485	0.304
Highest Quartile, Immigrant	-1.700	0.251
Constant	-2.898	0.180

	Model 5	
	Beta	**SE**
Gender (Ref: Girl)		
Boy	-0.235	0.106
ESCS (Ref: Lowest Quartile "Disadvantaged")		
Second Quartile	0.420	0.176
Third Quartile	0.747	0.177
Highest Quartile	1.681	0.169
Maths Ability	1.275	0.219
ESCS*Maths Ability		
Second ESCS Quartile*Maths Ability	-0.257	0.254
Third ESCS Quartile*Maths Ability	-0.171	0.279
Highest ESCS Quartile*Maths Ability	-0.127	0.253
Constant	-1.998	0.143

Finland

	Model 1		Model 2		Model 3	
	Beta	SE	Beta	SE	Beta	SE
Gender (Ref: Girl)						
Boy	-0.158	0.059	-0.236	0.062	-0.200	0.064
Books (Ref: Low)						
High	0.403	0.067	0.232	0.070	0.280	0.072
Missing	-0.847	0.360	-0.892	0.383	-0.997	0.426
Parental Education (Ref: Low)						
Medium	0.328	0.073	0.291	0.077	0.066	0.196
High	0.949	0.084	0.931	0.087	0.525	0.073
Missing	-0.309	0.351	-0.200	0.354	-0.127	0.357
Occupation spline 0-10th percentile	0.005	0.024	0.010	0.025	0.012	0.025
Occupation spline 11-25th percentile	0.032	0.031	0.025	0.032	0.023	0.033
Occupation spline 26-50th percentile	-0.001	0.007	-0.006	0.007	-0.005	0.008
Occupation spline 50-75th percentile	0.012	0.009	0.008	0.010	0.024	0.010
Occupation spline 76-90th percentile	0.066	0.034	0.046	0.035	0.037	0.036
Occupation spline 91-100th percentile	-0.011	0.015	-0.012	0.015	-0.002	0.016
Immigrant Status (Ref: Native)						
Immigrant	3.162	4.595	3.179	5.031	3.179	4.772
Books*Immigrant						
High Books, Immigrant	-0.246	0.274	-0.216	0.273	-0.160	0.269
Missing Books, Immigrant	1.366	1.197	0.957	1.040	1.323	0.983
Parental Education*Immigrant						
Medium Education, Immigrant	-0.011	0.377	-0.065	0.390	-0.628	1.348
High Education, Immigrant	-0.146	0.396	-0.238	0.402	-0.211	0.356
Missing Education, Immigrant	1.930	1.569	1.800	1.478	1.093	1.447
Occupation 0-10 * Immigrant	-0.104	0.168	-0.097	0.182	-0.093	0.173
Occupation 11-25 * Immigrant	0.031	0.137	0.037	0.143	0.051	0.147
Occupation 26-50 * Immigrant	-0.004	0.032	-0.012	0.033	-0.018	0.034
Occupation 51-75 * Immigrant	0.039	0.040	0.056	0.041	0.046	0.039
Occupation 76-90 * Immigrant	-0.085	0.124	-0.100	0.131	-0.128	0.121
Occupation 91-100 * Immigrant	-0.022	0.040	-0.011	0.044	-0.006	0.045
Ability spline 0-10th percentile	-	-	-0.003	0.002	-0.004	0.002
Ability spline 11-25th percentile	-	-	0.010	0.003	0.008	0.003
Ability spline 26-50th percentile	-	-	0.001	0.002	0.002	0.003
Ability spline 50-75th percentile	-	-	0.008	0.002	0.008	0.002
Ability spline 76-90th percentile	-	-	0.011	0.003	0.011	0.003
Ability spline 91-100th percentile	-	-	0.007	0.003	0.007	0.003
School FE	-	-			Yes	Yes

1 Source: Author's calculations using PISA 2003 data. Sample size = 5,793 in models 1,2, 4 and 5. Sample size = 5,706 in model 3 where some observations are dropped due to the fixed effect perfectly predicting the response. Dependent variable in all models was whether respondent expected to obtain a degree (coded 0 if they did not and 1 if they did)

	Model 4	
	Beta	**SE**
Gender (Ref: Girl)		
Boy	-0.196	0.058
ESCS (Ref: Lowest Quartile "Disadvantaged")		
Second Quartile	0.278	0.098
Third Quartile	0.677	0.087
Highest Quartile "Advantaged"	1.281	0.098
Maths Ability (Ref: Lowest Ability Quintile)		
Second Quintile	0.303	0.102
Third Quintile	0.405	0.108
Fourth Quintile	0.663	0.106
Highest Quintile	1.393	0.108
Immigrant Status (Ref: Native)		
Immigrant	0.583	0.270
ESCS*Immigrant		
Second Quartile, Immigrant	-0.550	0.363
Third Quartile, Immigrant	0.279	0.371
Highest Quartile, Immigrant	-0.342	0.373
Constant	-0.966	0.102

	Model 5	
	Beta	**SE**
Gender (Ref: Girl)		
Boy	0.049	0.060
ESCS (Ref: Lowest Quartile "Disadvantaged")		
Second Quartile	0.321	0.100
Third Quartile	0.693	0.093
Highest Quartile	1.330	0.101
Maths Ability	0.296	0.068
ESCS*Maths Ability		
Second ESCS Quartile*Maths Ability	0.046	0.094
Third ESCS Quartile*Maths Ability	0.090	0.098
Highest ESCS Quartile*Maths Ability	0.195	0.097
Constant	-0.591	0.081

France

	Model 1		Model 2		Model 3	
	Beta	**SE**	**Beta**	**SE**	**Beta**	**SE**
Gender (Ref: Girl)						
Boy	-0.536	0.087	-0.741	0.096	-0.696	0.101
Books (Ref: Low)						
High	0.757	0.093	0.431	0.099	0.358	0.108
Missing	0.296	0.395	0.191	0.494	0.049	0.542
Parental Education (Ref: Low)						
Medium	0.270	0.153	0.422	0.169	0.000	0.000
High	0.775	0.101	0.810	0.104	0.584	0.105
Missing	-0.270	0.225	0.094	0.231	0.089	0.275
Occupation spline 0-10th percentile	0.006	0.031	0.007	0.036	0.029	0.039
Occupation spline 11-25th percentile	0.022	0.058	-0.009	0.063	-0.021	0.067
Occupation spline 26-50th percentile	0.039	0.012	0.025	0.013	0.017	0.013
Occupation spline 50-75th percentile	-0.007	0.029	-0.011	0.031	-0.023	0.034
Occupation spline 76-90th percentile	0.044	0.020	0.030	0.021	0.042	0.024
Occupation spline 91-100th percentile	0.013	0.016	0.005	0.016	-0.002	0.015
Immigrant Status (Ref: Native)						
Immigrant	1.889	1.090	2.808	1.176	3.057	1.326
Books*Immigrant						
High Books, Immigrant	0.006	0.175	0.100	0.186	0.142	0.205
Missing Books, Immigrant	-0.769	0.910	-0.137	1.119	0.523	1.169
Parental Education*Immigrant						
Medium Education, Immigrant	-0.382	0.296	-0.372	0.345	0.000	0.000
High Education, Immigrant	-0.364	0.207	-0.376	0.217	-0.297	0.213
Missing Education, Immigrant	-0.314	0.419	-0.374	0.461	-0.369	0.467
Occupation 0-10 * Immigrant	-0.039	0.043	-0.057	0.046	-0.079	0.052
Occupation 11-25 * Immigrant	-0.072	0.080	-0.057	0.089	-0.013	0.099
Occupation 26-50 * Immigrant	-0.002	0.022	-0.027	0.025	-0.033	0.026
Occupation 51-75 * Immigrant	-0.012	0.054	-0.004	0.063	0.054	0.065
Occupation 76-90 * Immigrant	-0.012	0.038	-0.002	0.044	-0.041	0.045
Occupation 91-100 * Immigrant	0.031	0.032	0.033	0.035	0.055	0.039
Ability spline 0-10th percentile	-	-	0.017	0.008	0.019	0.008
Ability spline 11-25th percentile	-	-	0.015	0.005	0.007	0.006
Ability spline 26-50th percentile	-	-	0.014	0.003	0.010	0.003
Ability spline 50-75th percentile	-	-	0.009	0.003	0.004	0.003
Ability spline 76-90th percentile	-	-	0.015	0.003	0.014	0.003
Ability spline 91-100th percentile	-	-	0.007	0.004	0.005	0.004
School FE	-	-			Yes	Yes

1 Source: Author’s calculations using PISA 2003 data. Sample size = 3,997 in models 1,2, 4 and 5. Sample size = 3,489 in model 3 where some observations are dropped due to the fixed effect perfectly predicting the response. Dependent variable in all models was whether respondent expected to obtain a degree (coded 0 if they did not and 1 if they did)

	Model 4	
	Beta	**SE**
Gender (Ref: Girl)		
Boy	-0.752	0.093
ESCS (Ref: Lowest Quartile "Disadvantaged")		
Second Quartile	0.744	0.168
Third Quartile	0.844	0.163
Highest Quartile "Advantaged"	1.846	0.181
Maths Ability (Ref: Lowest Ability Quintile)		
Second Quintile	1.039	0.175
Third Quintile	1.687	0.183
Fourth Quintile	2.205	0.180
Highest Quintile	2.973	0.175
Immigrant Status (Ref: Native)		
Immigrant	1.158	0.215
ESCS*Immigrant		
Second Quartile, Immigrant	-0.690	0.254
Third Quartile, Immigrant	-0.618	0.269
Highest Quartile, Immigrant	-0.908	0.270
Constant	-3.171	0.215

	Model 5	
	Beta	**SE**
Gender (Ref: Girl)		
Boy	-0.256	0.109
ESCS (Ref: Lowest Quartile "Disadvantaged")		
Second Quartile	0.709	0.168
Third Quartile	0.897	0.161
Highest Quartile	1.868	0.177
Maths Ability	1.509	0.171
ESCS*Maths Ability		
Second ESCS Quartile*Maths Ability	-0.340	0.230
Third ESCS Quartile*Maths Ability	-0.700	0.196
Highest ESCS Quartile*Maths Ability	-0.645	0.201
Constant	-1.629	0.155

Germany

	Model 1		Model 2		Model 3	
	Beta	**SE**	**Beta**	**SE**	**Beta**	**SE**
Gender (Ref: Girl)						
Boy	-0.168	0.103	-0.351	0.105	-0.147	0.116
Books (Ref: Low)						
High	1.026	0.133	0.678	0.137	0.665	0.140
Missing	0.682	0.534	1.202	0.560	1.567	0.631
Parental Education (Ref: Low)						
Medium	0.315	0.122	0.193	0.122	0.188	0.169
High	1.158	0.125	0.966	0.132	0.463	0.118
Missing	-1.145	0.467	-1.012	0.462	-1.211	0.497
Occupation spline 0-10th percentile	0.038	0.046	0.044	0.052	0.072	0.061
Occupation spline 11-25th percentile	0.043	0.044	0.007	0.048	-0.029	0.051
Occupation spline 26-50th percentile	0.027	0.016	0.017	0.017	0.020	0.018
Occupation spline 50-75th percentile	0.047	0.030	0.035	0.033	0.031	0.032
Occupation spline 76-90th percentile	0.007	0.022	0.007	0.024	0.009	0.023
Occupation spline 91-100th percentile	0.041	0.012	0.036	0.013	0.043	0.014
Immigrant Status (Ref: Native)						
Immigrant	3.719	1.790	4.815	1.992	5.828	2.088
Books*Immigrant						
High Books, Immigrant	-0.161	0.254	-0.296	0.278	-0.550	0.300
Missing Books, Immigrant	0.906	0.873	0.780	0.976	-0.119	1.040
Parental Education*Immigrant						
Medium Education, Immigrant	0.411	0.300	0.511	0.314	-0.082	0.460
High Education, Immigrant	-0.168	0.293	0.148	0.311	0.611	0.298
Missing Education, Immigrant	-0.087	0.916	0.325	0.933	0.494	1.155
Occupation 0-10 * Immigrant	-0.114	0.066	-0.136	0.073	-0.169	0.077
Occupation 11-25 * Immigrant	0.042	0.080	0.021	0.089	0.041	0.097
Occupation 26-50 * Immigrant	-0.063	0.035	-0.061	0.037	-0.054	0.039
Occupation 51-75 * Immigrant	0.161	0.082	0.161	0.086	0.112	0.086
Occupation 76-90 * Immigrant	-0.069	0.055	-0.091	0.059	-0.059	0.063
Occupation 91-100 * Immigrant	-0.035	0.034	-0.035	0.034	-0.066	0.036
Ability spline 0-10th percentile	-	-	0.000	0.006	-0.002	0.008
Ability spline 11-25th percentile	-	-	0.008	0.006	0.008	0.007
Ability spline 26-50th percentile	-	-	0.015	0.004	0.011	0.004
Ability spline 50-75th percentile	-	-	0.010	0.003	0.006	0.003
Ability spline 76-90th percentile	-	-	0.015	0.004	0.012	0.004
Ability spline 91-100th percentile	-	-	0.003	0.003	0.001	0.003
School FE	-	-			Yes	Yes

1 Source: Author's calculations using PISA 2003 data. Sample size = 4,457 in models 1,2, 4 and 5. Sample size = 3,298 in model 3 where some observations are dropped due to the fixed effect perfectly predicting the response. Dependent variable in all models was whether respondent expected to obtain a degree (coded 0 if they did not and 1 if they did)

	Model 4	
	Beta	**SE**
Gender (Ref: Girl)		
Boy	-0.363	0.108
ESCS (Ref: Lowest Quartile "Disadvantaged")		
Second Quartile	0.393	0.248
Third Quartile	1.137	0.233
Highest Quartile "Advantaged"	2.155	0.218
Maths Ability (Ref: Lowest Ability Quintile)		
Second Quintile	0.449	0.279
Third Quintile	1.374	0.261
Fourth Quintile	1.987	0.259
Highest Quintile	2.892	0.252
Immigrant Status (Ref: Native)		
Immigrant	0.732	0.331
ESCS*Immigrant		
Second Quartile, Immigrant	0.167	0.463
Third Quartile, Immigrant	0.019	0.407
Highest Quartile, Immigrant	-0.492	0.381
Constant	-4.250	0.333

	Model 5	
	Beta	**SE**
Gender (Ref: Girl)		
Boy	-0.256	0.109
ESCS (Ref: Lowest Quartile "Disadvantaged")		
Second Quartile	0.709	0.168
Third Quartile	0.897	0.161
Highest Quartile	1.868	0.177
Maths Ability	1.509	0.171
ESCS*Maths Ability		
Second ESCS Quartile*Maths Ability	-0.340	0.230
Third ESCS Quartile*Maths Ability	-0.700	0.196
Highest ESCS Quartile*Maths Ability	-0.645	0.201
Constant	-1.629	0.155

Greece

	Model 1		Model 2		Model 3	
	Beta	**SE**	**Beta**	**SE**	**Beta**	**SE**
Gender (Ref: Girl)						
Boy	-0.693	0.093	-1.136	0.096	-0.907	0.084
Books (Ref: Low)						
High	0.768	0.089	0.591	0.098	0.488	0.124
Missing	-0.588	0.291	-0.519	0.349	-0.118	0.352
Parental Education (Ref: Low)						
Medium	0.487	0.096	0.543	0.102	0.509	0.149
High	0.809	0.143	0.938	0.145	0.651	0.130
Missing			0.000	0.000	0.000	0.000
Occupation spline 0-10th percentile	-0.079	0.039	-0.087	0.048	-0.107	0.061
Occupation spline 11-25th percentile	0.092	0.049	0.068	0.056	0.147	0.070
Occupation spline 26-50th percentile	0.038	0.014	0.024	0.016	0.000	0.019
Occupation spline 50-75th percentile	0.004	0.026	0.007	0.028	-0.003	0.028
Occupation spline 76-90th percentile	0.079	0.017	0.051	0.018	0.041	0.020
Occupation spline 91-100th percentile	0.016	0.023	0.021	0.023	0.012	0.025
Immigrant Status (Ref: Native)						
Immigrant	-1.604	1.566	-2.696	1.937	-1.788	2.385
Books*Immigrant						
High Books, Immigrant	0.227	0.192	-0.027	0.248	-0.198	0.319
Missing Books, Immigrant	-0.109	0.891	0.982	1.169	0.702	0.882
Parental Education*Immigrant						
Medium Education, Immigrant	-0.289	0.237	-0.190	0.263	-0.205	0.333
High Education, Immigrant	0.140	0.269	0.085	0.291	0.355	0.298
Missing Education, Immigrant			0.000	0.000	0.000	0.000
Occupation 0-10 * Immigrant	0.049	0.067	0.098	0.081	0.087	0.099
Occupation 11-25 * Immigrant	-0.066	0.111	-0.109	0.130	-0.139	0.156
Occupation 26-50 * Immigrant	0.011	0.037	0.015	0.044	-0.001	0.051
Occupation 51-75 * Immigrant	0.016	0.059	0.052	0.061	0.104	0.079
Occupation 76-90 * Immigrant	-0.058	0.044	-0.064	0.050	-0.065	0.057
Occupation 91-100 * Immigrant	0.048	0.041	0.034	0.044	0.007	0.043
Ability spline 0-10th percentile	-	-	0.015	0.003	0.012	0.004
Ability spline 11-25th percentile	-	-	0.012	0.003	0.008	0.004
Ability spline 26-50th percentile	-	-	0.016	0.003	0.010	0.004
Ability spline 50-75th percentile	-	-	0.013	0.003	0.008	0.004
Ability spline 76-90th percentile	-	-	0.016	0.005	0.012	0.005
Ability spline 91-100th percentile	-	-	0.019	0.010	0.019	0.010
School FE	-	-			Yes	Yes

1 Source: Author's calculations using PISA 2003 data. Sample size = 4,613 in models 1,2, 4 and 5. Sample size = 4,195 in model 3 where some observations are dropped due to the fixed effect perfectly predicting the response. Dependent variable in all models was whether respondent expected to obtain a degree (coded 0 if they did not and 1 if they did)

	Model 4	
	Beta	**SE**
Gender (Ref: Girl)		
Boy	-1.106	0.098
ESCS (Ref: Lowest Quartile "Disadvantaged")		
Second Quartile	0.800	0.122
Third Quartile	1.180	0.133
Highest Quartile "Advantaged"	2.468	0.193
Maths Ability (Ref: Lowest Ability Quintile)		
Second Quintile	1.103	0.120
Third Quintile	1.711	0.132
Fourth Quintile	2.492	0.150
Highest Quintile	3.491	0.194
Immigrant Status (Ref: Native)		
Immigrant	-0.503	0.220
ESCS*Immigrant		
Second Quartile, Immigrant	0.247	0.262
Third Quartile, Immigrant	0.411	0.316
Highest Quartile, Immigrant	-0.131	0.305
Constant	-1.324	0.172

	Model 5	
	Beta	**SE**
Gender (Ref: Girl)		
Boy	-0.490	0.104
ESCS (Ref: Lowest Quartile "Disadvantaged")		
Second Quartile	0.742	0.120
Third Quartile	1.354	0.146
Highest Quartile	2.616	0.220
Maths Ability	0.891	0.107
ESCS*Maths Ability		
Second ESCS Quartile*Maths Ability	0.097	0.143
Third ESCS Quartile*Maths Ability	0.111	0.144
Highest ESCS Quartile*Maths Ability	0.192	0.163
Constant	0.006	0.162

Hungary

	Model 1		Model 2		Model 3	
	Beta	**SE**	**Beta**	**SE**	**Beta**	**SE**
Gender (Ref: Girl)						
Boy	-0.864	0.101	-1.213	0.107	-0.731	0.122
Books (Ref: Low)						
High	0.992	0.091	0.502	0.093	0.230	0.100
Missing	-0.589	0.306	-0.429	0.470	-0.582	0.464
Parental Education (Ref: Low)						
Medium	0.219	0.080	0.322	0.087	0.148	0.116
High	1.518	0.143	1.418	0.138	0.691	0.131
Missing	-0.614	0.625	0.340	0.525	0.266	0.483
Occupation spline 0-10th percentile	0.051	0.027	0.072	0.030	0.077	0.034
Occupation spline 11-25th percentile	0.062	0.025	0.044	0.026	-0.018	0.031
Occupation spline 26-50th percentile	0.014	0.022	0.006	0.025	0.023	0.031
Occupation spline 50-75th percentile	0.068	0.014	0.042	0.014	0.034	0.016
Occupation spline 76-90th percentile	-0.012	0.019	-0.004	0.019	-0.010	0.020
Occupation spline 91-100th percentile	0.022	0.020	0.009	0.024	0.009	0.023
Immigrant Status (Ref: Native)						
Immigrant	6.744	2.459	6.518	3.272	5.560	2.622
Books*Immigrant						
High Books, Immigrant	0.447	0.418	0.346	0.471	0.806	0.579
Missing Books, Immigrant			0.000	0.000	0.000	0.000
Parental Education*Immigrant						
Medium Education, Immigrant	-0.033	0.523	-0.123	0.567	-0.555	0.740
High Education, Immigrant	-0.970	0.601	-0.533	0.696	-1.370	0.805
Missing Education, Immigrant			0.000	0.000	0.000	0.000
Occupation 0-10 * Immigrant	-0.219	0.093	-0.198	0.124	-0.168	0.104
Occupation 11-25 * Immigrant	-0.005	0.125	-0.073	0.150	-0.003	0.161
Occupation 26-50 * Immigrant	0.008	0.152	0.067	0.171	0.074	0.215
Occupation 51-75 * Immigrant	-0.077	0.067	-0.087	0.071	-0.145	0.095
Occupation 76-90 * Immigrant	0.121	0.093	0.067	0.100	0.198	0.129
Occupation 91-100 * Immigrant	-0.081	0.057	-0.095	0.060	-0.149	0.066
Ability spline 0-10th percentile	-	-	0.010	0.005	0.000	0.006
Ability spline 11-25th percentile	-	-	0.017	0.005	0.010	0.005
Ability spline 26-50th percentile	-	-	0.018	0.003	0.012	0.003
Ability spline 50-75th percentile	-	-	0.011	0.003	0.004	0.003
Ability spline 76-90th percentile	-	-	0.013	0.004	0.005	0.005
Ability spline 91-100th percentile	-	-	0.022	0.006	0.018	0.007
School FE	-	-			Yes	Yes

1 Source: Author's calculations using PISA 2003 data. Sample size = 4,756 in models 1,2, 4 and 5. Sample size = 3,728 in model 3 where some observations are dropped due to the fixed effect perfectly predicting the response. Dependent variable in all models was whether respondent expected to obtain a degree (coded 0 if they did not and 1 if they did)

	Model 4	
	Beta	**SE**
Gender (Ref: Girl)		
Boy	-1.194	0.108
ESCS (Ref: Lowest Quartile "Disadvantaged")		
Second Quartile	0.786	0.129
Third Quartile	1.554	0.127
Highest Quartile "Advantaged"	2.671	0.165
Maths Ability (Ref: Lowest Ability Quintile)		
Second Quintile	1.250	0.149
Third Quintile	1.929	0.141
Fourth Quintile	2.568	0.170
Highest Quintile	3.469	0.194
Immigrant Status (Ref: Native)		
Immigrant	1.496	0.348
ESCS*Immigrant		
Second Quartile, Immigrant	-1.409	0.509
Third Quartile, Immigrant	-1.496	0.534
Highest Quartile, Immigrant	-1.950	0.564
Constant	-2.270	0.165

	Model 5	
	Beta	**SE**
Gender (Ref: Girl)		
Boy	-0.700	0.105
ESCS (Ref: Lowest Quartile "Disadvantaged")		
Second Quartile	0.834	0.122
Third Quartile	1.608	0.131
Highest Quartile	2.835	0.159
Maths Ability	1.107	0.123
ESCS*Maths Ability		
Second ESCS Quartile*Maths Ability	-0.061	0.164
Third ESCS Quartile*Maths Ability	-0.037	0.160
Highest ESCS Quartile*Maths Ability	0.023	0.169
Constant	-0.772	0.135

Iceland

	Model 1		Model 2		Model 3	
	Beta	**SE**	**Beta**	**SE**	**Beta**	**SE**
Gender (Ref: Girl)						
Boy	-0.572	0.091	-0.542	0.101	-0.522	0.099
Books (Ref: Low)						
High	0.537	0.099	0.280	0.101	0.325	0.102
Missing	-0.029	0.450	0.410	0.465	0.326	0.434
Parental Education (Ref: Low)						
Medium	0.431	0.099	0.349	0.103	0.209	0.120
High	1.651	0.121	1.562	0.121	1.056	0.113
Missing	-0.941	0.703	-1.024	0.692	-1.021	0.661
Occupation spline 0-10th percentile	0.022	0.045	0.029	0.047	-0.026	0.047
Occupation spline 11-25th percentile	0.015	0.013	0.013	0.014	0.006	0.015
Occupation spline 26-50th percentile	0.007	0.016	0.001	0.016	-0.006	0.016
Occupation spline 50-75th percentile	0.002	0.011	0.002	0.011	0.010	0.011
Occupation spline 76-90th percentile	0.032	0.049	-0.005	0.051	-0.027	0.050
Occupation spline 91-100th percentile	0.010	0.017	0.019	0.018	0.028	0.018
Immigrant Status (Ref: Native)						
Immigrant	5.646	4.115	3.125	4.210	2.871	4.099
Books*Immigrant						
High Books, Immigrant	-0.472	0.303	-0.424	0.314	-0.299	0.318
Missing Books, Immigrant	-0.068	1.110	-0.162	1.207	-0.091	1.547
Parental Education*Immigrant						
Medium Education, Immigrant	0.114	0.352	0.172	0.375	0.606	0.407
High Education, Immigrant	0.431	0.345	0.534	0.344	0.203	0.344
Missing Education, Immigrant	-	-	0.000	0.000	0.000	0.000
Occupation 0-10 * Immigrant	-0.186	0.153	-0.092	0.155	-0.094	0.152
Occupation 11-25 * Immigrant	-0.048	0.058	-0.041	0.057	-0.029	0.057
Occupation 26-50 * Immigrant	0.086	0.064	0.053	0.064	0.061	0.066
Occupation 51-75 * Immigrant	-0.027	0.035	-0.036	0.035	-0.028	0.035
Occupation 76-90 * Immigrant	0.150	0.154	0.204	0.168	0.163	0.160
Occupation 91-100 * Immigrant	-0.065	0.036	-0.075	0.037	-0.049	0.042
Ability spline 0-10th percentile	-	-	0.012	0.005	0.010	0.004
Ability spline 11-25th percentile	-	-	0.005	0.005	0.006	0.005
Ability spline 26-50th percentile	-	-	0.010	0.003	0.010	0.003
Ability spline 50-75th percentile	-	-	0.010	0.003	0.012	0.003
Ability spline 76-90th percentile	-	-	0.004	0.004	0.002	0.004
Ability spline 91-100th percentile	-	-	0.008	0.004	0.011	0.004
School FE	-	-			Yes	Yes

1 Source: Author's calculations using PISA 2003 data. Sample size = 3,324 in models 1,2, 4 and 5. Sample size = 3,223 in model 3 where some observations are dropped due to the fixed effect perfectly predicting the response. Dependent variable in all models was whether respondent expected to obtain a degree (coded 0 if they did not and 1 if they did)

	Model 4	
	Beta	SE
Gender (Ref: Girl)		
Boy	-0.535	0.091
ESCS (Ref: Lowest Quartile "Disadvantaged")		
Second Quartile	0.558	0.134
Third Quartile	0.935	0.147
Highest Quartile "Advantaged"	1.638	0.148
Maths Ability (Ref: Lowest Ability Quintile)		
Second Quintile	0.609	0.169
Third Quintile	1.026	0.156
Fourth Quintile	1.544	0.145
Highest Quintile	1.971	0.130
Immigrant Status (Ref: Native)		
Immigrant	0.685	0.302
ESCS*Immigrant		
Second Quartile, Immigrant	-1.289	0.531
Third Quartile, Immigrant	-0.283	0.343
Highest Quartile, Immigrant	-0.451	0.384
Constant	-2.271	0.144

	Model 5	
	Beta	SE
Gender (Ref: Girl)		
Boy	-0.238	0.096
ESCS (Ref: Lowest Quartile "Disadvantaged")		
Second Quartile	0.626	0.143
Third Quartile	1.054	0.162
Highest Quartile	1.827	0.158
Maths Ability	1.099	0.143
ESCS*Maths Ability		
Second ESCS Quartile*Maths Ability	-0.213	0.172
Third ESCS Quartile*Maths Ability	-0.464	0.187
Highest ESCS Quartile*Maths Ability	-0.508	0.150
Constant	-1.532	0.115

Ireland

	Model 1		Model 2		Model 3	
	Beta	**SE**	**Beta**	**SE**	**Beta**	**SE**
Gender (Ref: Girl)						
Boy	-0.752	0.074	-0.969	0.081	-1.086	0.081
Books (Ref: Low)						
High	0.786	0.083	0.543	0.090	0.552	0.093
Missing	-1.770	0.522	-1.673	0.540	-1.921	0.624
Parental Education (Ref: Low)						
Medium	0.383	0.092	0.303	0.098	0.299	0.108
High	1.037	0.122	0.993	0.130	0.466	0.111
Missing	-0.584	0.550	-0.354	0.505	-0.446	0.513
Occupation spline 0-10th percentile	-0.014	0.035	-0.016	0.035	-0.016	0.032
Occupation spline 11-25th percentile	-0.069	0.037	-0.081	0.038	-0.095	0.036
Occupation spline 26-50th percentile	0.057	0.012	0.046	0.013	0.047	0.013
Occupation spline 50-75th percentile	-0.025	0.025	-0.032	0.029	-0.025	0.029
Occupation spline 76-90th percentile	0.030	0.012	0.021	0.014	0.029	0.014
Occupation spline 91-100th percentile	-0.003	0.018	-0.012	0.019	-0.019	0.017
Immigrant Status (Ref: Native)						
Immigrant	-2.751	1.605	-2.454	1.662	-2.199	1.708
Books*Immigrant						
High Books, Immigrant	-0.293	0.188	-0.235	0.193	-0.219	0.206
Missing Books, Immigrant	2.206	0.857	2.131	1.010	2.360	1.062
Parental Education*Immigrant						
Medium Education, Immigrant	0.052	0.231	0.070	0.237	0.238	0.275
High Education, Immigrant	0.182	0.335	0.258	0.344	0.231	0.253
Missing Education, Immigrant	-0.552	0.875	-0.565	0.917	-0.261	0.861
Occupation 0-10 * Immigrant	0.114	0.060	0.101	0.063	0.085	0.064
Occupation 11-25 * Immigrant	0.016	0.095	0.032	0.099	0.052	0.099
Occupation 26-50 * Immigrant	-0.054	0.034	-0.055	0.036	-0.058	0.037
Occupation 51-75 * Immigrant	0.113	0.067	0.118	0.072	0.137	0.070
Occupation 76-90 * Immigrant	-0.065	0.037	-0.073	0.041	-0.076	0.039
Occupation 91-100 * Immigrant	0.037	0.036	0.055	0.035	0.064	0.036
Ability spline 0-10th percentile	-	-	0.011	0.005	0.011	0.005
Ability spline 11-25th percentile	-	-	0.005	0.004	0.004	0.004
Ability spline 26-50th percentile	-	-	0.017	0.003	0.017	0.003
Ability spline 50-75th percentile	-	-	0.005	0.003	0.005	0.003
Ability spline 76-90th percentile	-	-	0.007	0.004	0.007	0.004
Ability spline 91-100th percentile	-	-	0.004	0.004	0.004	0.004
School FE	-	-			Yes	Yes

1 Source: Author’s calculations using PISA 2003 data. Sample size = 3,848 in models 1,2, 4 and 5. Sample size = 3,698 in model 3 where some observations are dropped due to the fixed effect perfectly predicting the response. Dependent variable in all models was whether respondent expected to obtain a degree (coded 0 if they did not and 1 if they did)

	Model 4	
	Beta	**SE**
Gender (Ref: Girl)		
Boy	-0.929	0.081
ESCS (Ref: Lowest Quartile "Disadvantaged")		
Second Quartile	0.355	0.116
Third Quartile	0.785	0.115
Highest Quartile "Advantaged"	1.430	0.132
Maths Ability (Ref: Lowest Ability Quintile)		
Second Quintile	0.499	0.133
Third Quintile	1.282	0.126
Fourth Quintile	1.584	0.135
Highest Quintile	1.959	0.137
Immigrant Status (Ref: Native)		
Immigrant	0.293	0.211
ESCS*Immigrant		
Second Quartile, Immigrant	-0.014	0.276
Third Quartile, Immigrant	-0.151	0.259
Highest Quartile, Immigrant	-0.364	0.301
Constant	-1.109	0.124

	Model 5	
	Beta	**SE**
Gender (Ref: Girl)		
Boy	-0.560	0.073
ESCS (Ref: Lowest Quartile "Disadvantaged")		
Second Quartile	0.294	0.114
Third Quartile	0.684	0.115
Highest Quartile	1.349	0.126
Maths Ability	0.823	0.101
ESCS*Maths Ability		
Second ESCS Quartile*Maths Ability	0.025	0.151
Third ESCS Quartile*Maths Ability	0.068	0.148
Highest ESCS Quartile*Maths Ability	-0.144	0.138
Constant	-0.178	0.092

Italy

	Model 1		Model 2		Model 3	
	Beta	**SE**	**Beta**	**SE**	**Beta**	**SE**
Gender (Ref: Girl)						
Boy	-0.812	0.092	-0.975	0.096	-0.592	0.097
Books (Ref: Low)						
High	0.678	0.075	0.486	0.076	0.346	0.084
Missing	-0.275	0.240	-0.073	0.244	0.205	0.263
Parental Education (Ref: Low)						
Medium	0.157	0.070	0.241	0.073	0.187	0.105
High	0.468	0.107	0.522	0.111	0.276	0.089
Missing	-1.501	0.689	-1.422	0.579	-1.437	0.779
Occupation spline 0-10th percentile	-0.022	0.022	-0.041	0.024	-0.061	0.028
Occupation spline 11-25th percentile	-0.032	0.024	-0.038	0.024	0.011	0.030
Occupation spline 26-50th percentile	0.067	0.010	0.060	0.011	0.023	0.012
Occupation spline 50-75th percentile	0.012	0.014	0.005	0.015	0.021	0.016
Occupation spline 76-90th percentile	0.064	0.010	0.059	0.010	0.026	0.012
Occupation spline 91-100th percentile	0.025	0.013	0.024	0.013	0.013	0.016
Immigrant Status (Ref: Native)						
Immigrant	-3.618	1.888	-4.349	1.936	-4.407	1.961
Books*Immigrant						
High Books, Immigrant	0.317	0.215	0.386	0.224	0.372	0.239
Missing Books, Immigrant	-0.229	0.756	-0.265	0.795	-0.167	0.822
Parental Education*Immigrant						
Medium Education, Immigrant	-0.482	0.297	-0.490	0.308	0.001	0.377
High Education, Immigrant	-0.284	0.330	-0.325	0.335	-0.318	0.308
Missing Education, Immigrant	-0.201	1.388	0.264	1.310	0.740	1.437
Occupation 0-10 * Immigrant	0.168	0.081	0.193	0.082	0.192	0.085
Occupation 11-25 * Immigrant	-0.200	0.098	-0.203	0.102	-0.139	0.111
Occupation 26-50 * Immigrant	0.080	0.048	0.086	0.049	0.100	0.057
Occupation 51-75 * Immigrant	-0.013	0.056	-0.018	0.059	-0.083	0.052
Occupation 76-90 * Immigrant	-0.068	0.035	-0.056	0.037	-0.005	0.032
Occupation 91-100 * Immigrant	-0.029	0.034	-0.014	0.034	-0.058	0.048
Ability spline 0-10th percentile	-	-	0.008	0.003	0.005	0.004
Ability spline 11-25th percentile	-	-	0.004	0.004	0.007	0.003
Ability spline 26-50th percentile	-	-	0.007	0.002	0.003	0.002
Ability spline 50-75th percentile	-	-	0.004	0.002	0.005	0.002
Ability spline 76-90th percentile	-	-	0.006	0.002	0.005	0.003
Ability spline 91-100th percentile	-	-	0.007	0.002	0.005	0.002
School FE	-	-			Yes	Yes

1 Source: Author's calculations using PISA 2003 data. Sample size = 11,631 in models 1,2, 4 and 5. Sample size = 10,783 in model 3 where some observations are dropped due to the fixed effect perfectly predicting the response. Dependent variable in all models was whether respondent expected to obtain a degree (coded 0 if they did not and 1 if they did)

	Model 4	
	Beta	SE
Gender (Ref: Girl)		
Boy	-1.032	0.095
ESCS (Ref: Lowest Quartile "Disadvantaged")		
Second Quartile	0.389	0.092
Third Quartile	1.081	0.102
Highest Quartile "Advantaged"	2.110	0.105
Maths Ability (Ref: Lowest Ability Quintile)		
Second Quintile	0.473	0.157
Third Quintile	0.793	0.156
Fourth Quintile	0.992	0.164
Highest Quintile	1.472	0.173
Immigrant Status (Ref: Native)		
Immigrant	0.192	0.268
ESCS*Immigrant		
Second Quartile, Immigrant	-0.316	0.327
Third Quartile, Immigrant	-0.326	0.335
Highest Quartile, Immigrant	-0.760	0.350
Constant	-1.033	0.190

	Model 5	
	Beta	SE
Gender (Ref: Girl)		
Boy	-0.718	0.099
ESCS (Ref: Lowest Quartile "Disadvantaged")		
Second Quartile	0.407	0.095
Third Quartile	1.017	0.100
Highest Quartile	2.050	0.103
Maths Ability	0.527	0.082
ESCS*Maths Ability		
Second ESCS Quartile*Maths Ability	0.175	0.103
Third ESCS Quartile*Maths Ability	-0.017	0.105
Highest ESCS Quartile*Maths Ability	0.060	0.104
Constant	-0.268	0.118

Japan

	Model 1		Model 2		Model 3	
	Beta	**SE**	**Beta**	**SE**	**Beta**	**SE**
Gender (Ref: Girl)						
Boy	0.264	0.134	0.224	0.136	0.813	0.115
Books (Ref: Low)						
High	0.497	0.075	0.293	0.078	0.309	0.090
Missing	-0.942	0.528	-0.822	0.476	-1.207	0.920
Parental Education (Ref: Low)						
Medium	0.362	0.103	0.315	0.109	0.000	0.000
High	1.189	0.099	1.064	0.101	0.351	0.000
Missing			0.000	0.000	0.000	0.000
Occupation spline 0-10th percentile	-0.008	0.024	-0.028	0.026	-0.036	0.032
Occupation spline 11-25th percentile	0.039	0.035	0.051	0.035	0.032	0.041
Occupation spline 26-50th percentile	0.103	0.022	0.073	0.021	0.042	0.024
Occupation spline 50-75th percentile	-0.004	0.017	0.006	0.019	0.007	0.019
Occupation spline 76-90th percentile	0.043	0.011	0.028	0.012	0.015	0.013
Occupation spline 91-100th percentile	-0.044	0.015	-0.024	0.017	-0.011	0.019
Immigrant Status (Ref: Native)						
Immigrant			1.170	0.582	1.346	0.515
Books*Immigrant						
High Books, Immigrant			0.000	0.000	0.000	0.000
Missing Books, Immigrant			0.000	0.000	0.000	0.000
Parental Education*Immigrant						
Medium Education, Immigrant			0.000	0.000	0.000	0.000
High Education, Immigrant			0.000	0.000	0.000	0.000
Missing Education, Immigrant			0.000	0.000	0.000	0.000
Occupation 0-10 * Immigrant			0.000	0.000	0.000	0.000
Occupation 11-25 * Immigrant			0.000	0.000	0.000	0.000
Occupation 26-50 * Immigrant			0.000	0.000	0.000	0.000
Occupation 51-75 * Immigrant			0.000	0.000	0.000	0.000
Occupation 76-90 * Immigrant			0.000	0.000	0.000	0.000
Occupation 91-100 * Immigrant			0.000	0.000	0.000	0.000
Ability spline 0-10th percentile	-	-	0.008	0.003	0.008	0.004
Ability spline 11-25th percentile	-	-	0.016	0.004	0.012	0.004
Ability spline 26-50th percentile	-	-	0.006	0.002	0.004	0.003
Ability spline 50-75th percentile	-	-	0.015	0.002	0.006	0.003
Ability spline 76-90th percentile	-	-	0.008	0.005	-0.003	0.005
Ability spline 91-100th percentile	-	-	0.034	0.007	0.030	0.009
School FE	-	-			Yes	Yes

Notes: 1 Immigrant interaction figures not given due to very small numbers (less than 1% of the sample) reporting that they area first or second generation immigrant. Hence no data to support these estimations

Source: Author's calculations using PISA 2003 data. Sample size = 4,700. Dependent variable in all models was whether respondent expected to obtain a degree (coded 0 if they did not and 1 if they did)

	Model 4	
	Beta	**SE**
Gender (Ref: Girl)		
Boy	0.284	0.132
ESCS (Ref: Lowest Quartile "Disadvantaged")		
Second Quartile	0.773	0.100
Third Quartile	1.514	0.107
Highest Quartile "Advantaged"	1.945	0.140
Maths Ability (Ref: Lowest Ability Quintile)		
Second Quintile	0.983	0.141
Third Quintile	1.501	0.145
Fourth Quintile	2.108	0.156
Highest Quintile	3.090	0.205
Immigrant Status (Ref: Native)		
Immigrant	1.048	0.761
ESCS*Immigrant		
Second Quartile, Immigrant	0.892	1.065
Third Quartile, Immigrant	-1.575	1.185
Highest Quartile, Immigrant	0.527	0.865
Constant	-2.746	0.164

	Model 5	
	Beta	**SE**
Gender (Ref: Girl)		
Boy	0.548	0.136
ESCS (Ref: Lowest Quartile "Disadvantaged")		
Second Quartile	0.713	0.093
Third Quartile	1.480	0.109
Highest Quartile	1.998	0.137
Maths Ability	0.925	0.109
ESCS*Maths Ability		
Second ESCS Quartile*Maths Ability	0.055	0.117
Third ESCS Quartile*Maths Ability	0.168	0.131
Highest ESCS Quartile*Maths Ability	-0.143	0.141
Constant	-1.355	0.125

Korea

	Model 1		Model 2		Model 3	
	Beta	SE	Beta	SE	Beta	SE
Gender (Ref: Girl)						
Boy	0.065	0.176	-0.174	0.162	-0.332	0.133
Books (Ref: Low)						
High	0.906	0.080	0.459	0.073	0.379	0.090
Missing	-0.248	0.673	0.371	0.964	0.686	1.173
Parental Education (Ref: Low)						
Medium	0.270	0.144	0.448	0.154	0.000	0.000
High	1.250	0.121	1.182	0.113	0.664	0.115
Missing			0.000	0.000	0.000	0.000
Occupation spline 0-10th percentile	0.008	0.024	0.012	0.023	0.020	0.027
Occupation spline 11-25th percentile	0.020	0.029	-0.017	0.031	-0.014	0.033
Occupation spline 26-50th percentile	0.044	0.025	0.053	0.026	0.031	0.028
Occupation spline 50-75th percentile	0.031	0.024	0.029	0.025	0.024	0.027
Occupation spline 76-90th percentile	-0.010	0.010	-0.013	0.010	-0.009	0.012
Occupation spline 91-100th percentile	-0.001	0.028	-0.018	0.032	-0.017	0.035
Immigrant Status (Ref: Native)						
Immigrant			0.215	1.278	0.783	1.007
Books*Immigrant						
High Books, Immigrant			0.000	0.000	0.000	0.000
Missing Books, Immigrant			0.000	0.000	0.000	0.000
Parental Education*Immigrant						
Medium Education, Immigrant			0.000	0.000	0.000	0.000
High Education, Immigrant			0.000	0.000	0.000	0.000
Missing Education, Immigrant			0.000	0.000	0.000	0.000
Occupation 0-10 * Immigrant			0.000	0.000	0.000	0.000
Occupation 11-25 * Immigrant			0.000	0.000	0.000	0.000
Occupation 26-50 * Immigrant			0.000	0.000	0.000	0.000
Occupation 51-75 * Immigrant			0.000	0.000	0.000	0.000
Occupation 76-90 * Immigrant			0.000	0.000	0.000	0.000
Occupation 91-100 * Immigrant			0.000	0.000	0.000	0.000
Ability spline 0-10th percentile	-	-	0.013	0.003	0.012	0.003
Ability spline 11-25th percentile	-	-	0.008	0.003	0.004	0.003
Ability spline 26-50th percentile	-	-	0.013	0.002	0.008	0.002
Ability spline 50-75th percentile	-	-	0.014	0.003	0.007	0.003
Ability spline 76-90th percentile	-	-	0.009	0.006	0.007	0.007
Ability spline 91-100th percentile	-	-	0.015	0.008	0.010	0.008
School FE	-	-			Yes	Yes

Notes: 1 Immigrant interaction figures not given due to very small numbers (less than 1% of the sample) reporting that they area first or second generation immigrant. Hence no data to support these estimations

Source: Author's calculations using PISA 2003 data. Sample size = 5,440. Dependent variable in all models was whether respondent expected to obtain a degree (coded 0 if they did not and 1 if they did)

	Model 4	
	Beta	SE
Gender (Ref: Girl)		
Boy	-0.184	0.159
ESCS (Ref: Lowest Quartile "Disadvantaged")		
Second Quartile	0.530	0.096
Third Quartile	1.066	0.110
Highest Quartile "Advantaged"	1.696	0.137
Maths Ability (Ref: Lowest Ability Quintile)		
Second Quintile	0.736	0.116
Third Quintile	1.470	0.137
Fourth Quintile	2.171	0.166
Highest Quintile	2.955	0.216
Immigrant Status (Ref: Native)		
Immigrant	19.410	1.174
ESCS*Immigrant		
Second Quartile, Immigrant	-20.401	1.506
Third Quartile, Immigrant	-22.009	0.000
Highest Quartile, Immigrant	0.000	0.000
Constant	-0.426	0.132

	Model 5	
	Beta	SE
Gender (Ref: Girl)		
Boy	0.344	0.163
ESCS (Ref: Lowest Quartile "Disadvantaged")		
Second Quartile	0.655	0.098
Third Quartile	1.222	0.115
Highest Quartile	1.990	0.152
Maths Ability	0.787	0.090
ESCS*Maths Ability		
Second ESCS Quartile*Maths Ability	0.257	0.115
Third ESCS Quartile*Maths Ability	0.284	0.118
Highest ESCS Quartile*Maths Ability	0.567	0.157
Constant	0.511	0.140

Luxemburg

	Model 1		Model 2		Model 3	
	Beta	**SE**	**Beta**	**SE**	**Beta**	**SE**
Gender (Ref: Girl)						
Boy	-0.102	0.105	-0.348	0.103	-0.156	0.094
Books (Ref: Low)						
High	0.907	0.141	0.618	0.136	0.514	0.127
Missing	0.435	0.618	0.501	0.729	0.552	0.759
Parental Education (Ref: Low)						
Medium	0.121	0.113	0.287	0.120	0.106	0.135
High	0.841	0.135	1.047	0.171	0.469	0.121
Missing	-0.358	0.204	-0.001	0.193	-0.006	0.203
Occupation spline 0-10th percentile	-0.042	0.039	-0.016	0.053	-0.011	0.054
Occupation spline 11-25th percentile	0.072	0.067	0.043	0.067	0.031	0.067
Occupation spline 26-50th percentile	0.046	0.015	0.027	0.017	0.013	0.018
Occupation spline 50-75th percentile	-0.002	0.037	-0.008	0.043	-0.023	0.046
Occupation spline 76-90th percentile	0.055	0.019	0.040	0.020	0.041	0.019
Occupation spline 91-100th percentile	0.026	0.021	0.032	0.022	0.040	0.019
Immigrant Status (Ref: Native)						
Immigrant	0.168	1.204	1.225	1.542	1.012	1.606
Books*Immigrant						
High Books, Immigrant	-0.186	0.189	-0.249	0.196	-0.344	0.190
Missing Books, Immigrant	-0.624	0.623	-0.445	0.783	-0.669	0.808
Parental Education*Immigrant						
Medium Education, Immigrant	0.088	0.153	-0.109	0.164	-0.229	0.248
High Education, Immigrant	0.199	0.232	-0.158	0.250	0.023	0.171
Missing Education, Immigrant	0.116	0.273	-0.074	0.253	-0.126	0.258
Occupation 0-10 * Immigrant	0.024	0.046	0.001	0.062	0.006	0.064
Occupation 11-25 * Immigrant	-0.141	0.081	-0.142	0.088	-0.127	0.091
Occupation 26-50 * Immigrant	0.012	0.023	0.017	0.027	0.014	0.025
Occupation 51-75 * Immigrant	-0.020	0.054	-0.024	0.065	-0.023	0.070
Occupation 76-90 * Immigrant	-0.005	0.025	0.012	0.029	0.018	0.032
Occupation 91-100 * Immigrant	-0.034	0.021	-0.056	0.022	-0.069	0.022
Ability spline 0-10th percentile	-	-	0.011	0.004	0.013	0.005
Ability spline 11-25th percentile	-	-	0.008	0.004	0.004	0.004
Ability spline 26-50th percentile	-	-	0.016	0.002	0.015	0.003
Ability spline 50-75th percentile	-	-	0.013	0.002	0.009	0.002
Ability spline 76-90th percentile	-	-	0.007	0.004	0.006	0.004
Ability spline 91-100th percentile	-	-	0.009	0.004	0.005	0.004
School FE	-	-			Yes	Yes

1 Source: Author's calculations using PISA 2003 data. Sample size = 3,892 in models 1,2, 4 and 5. Sample size = 3,748 in model 3 where some observations are dropped due to the fixed effect perfectly predicting the response. Dependent variable in all models was whether respondent expected to obtain a degree (coded 0 if they did not and 1 if they did)

	Model 4	
	Beta	SE
Gender (Ref: Girl)		
Boy	-0.308	0.097
ESCS (Ref: Lowest Quartile "Disadvantaged")		
Second Quartile	0.843	0.227
Third Quartile	1.321	0.259
Highest Quartile "Advantaged"	2.207	0.248
Maths Ability (Ref: Lowest Ability Quintile)		
Second Quintile	0.821	0.141
Third Quintile	1.539	0.135
Fourth Quintile	2.176	0.125
Highest Quintile	2.795	0.153
Immigrant Status (Ref: Native)		
Immigrant	1.226	0.240
ESCS*Immigrant		
Second Quartile, Immigrant	-0.878	0.260
Third Quartile, Immigrant	-0.556	0.312
Highest Quartile, Immigrant	-0.762	0.301
Constant	-3.302	0.249

	Model 5	
	Beta	SE
Gender (Ref: Girl)		
Boy	0.352	0.160
ESCS (Ref: Lowest Quartile "Disadvantaged")		
Second Quartile	0.600	0.215
Third Quartile	1.153	0.176
Highest Quartile	2.002	0.194
Maths Ability	1.152	0.218
ESCS*Maths Ability		
Second ESCS Quartile*Maths Ability	-0.352	0.239
Third ESCS Quartile*Maths Ability	-0.283	0.291
Highest ESCS Quartile*Maths Ability	-0.189	0.267
Constant	-1.652	0.210

Mexico

	Model 1		Model 2		Model 3	
	Beta	**SE**	**Beta**	**SE**	**Beta**	**SE**
Gender (Ref: Girl)						
Boy	-0.674	0.057	-0.811	0.060	-0.733	0.067
Books (Ref: Low)						
High	0.532	0.122	0.376	0.132	0.283	0.149
Missing	-0.311	0.194	-0.086	0.198	0.118	0.203
Parental Education (Ref: Low)						
Medium	0.747	0.100	0.571	0.104	0.000	0.000
High	0.489	0.089	0.572	0.097	0.462	0.084
Missing	-0.678	0.443	-0.445	0.397	-0.344	0.455
Occupation spline 0-10th percentile	-0.034	0.029	-0.005	0.030	-0.047	0.028
Occupation spline 11-25th percentile	0.459	0.167	0.111	0.178	0.125	0.170
Occupation spline 26-50th percentile	-0.010	0.016	-0.010	0.017	0.004	0.018
Occupation spline 50-75th percentile	0.020	0.005	0.013	0.005	0.008	0.005
Occupation spline 76-90th percentile	0.027	0.009	0.023	0.009	0.010	0.010
Occupation spline 91-100th percentile	0.022	0.013	0.011	0.013	0.010	0.013
Immigrant Status (Ref: Native)						
Immigrant	1.871	2.157	2.761	2.257	0.597	2.395
Books*Immigrant						
High Books, Immigrant	-0.079	0.457	-0.175	0.454	-0.219	0.510
Missing Books, Immigrant	0.064	0.849	-0.171	0.860	-0.738	0.813
Parental Education*Immigrant						
Medium Education, Immigrant	1.643	0.549	1.512	0.526	0.000	0.000
High Education, Immigrant	-0.814	0.346	-0.990	0.354	-0.754	0.273
Missing Education, Immigrant	-1.631	1.170	-1.855	1.191	-2.025	1.082
Occupation 0-10 * Immigrant	-0.131	0.110	-0.155	0.112	-0.059	0.120
Occupation 11-25 * Immigrant	-0.619	0.620	-0.415	0.642	-0.329	0.766
Occupation 26-50 * Immigrant	0.280	0.086	0.301	0.094	0.285	0.106
Occupation 51-75 * Immigrant	-0.006	0.025	-0.023	0.026	-0.032	0.029
Occupation 76-90 * Immigrant	-0.041	0.049	-0.021	0.045	0.058	0.061
Occupation 91-100 * Immigrant	0.069	0.048	0.072	0.043	-0.001	0.051
Ability spline 0-10th percentile	-	-	0.008	0.003	0.006	0.003
Ability spline 11-25th percentile	-	-	0.002	0.004	0.000	0.003
Ability spline 26-50th percentile	-	-	0.012	0.002	0.008	0.002
Ability spline 50-75th percentile	-	-	0.008	0.002	0.007	0.002
Ability spline 76-90th percentile	-	-	0.011	0.003	0.008	0.003
Ability spline 91-100th percentile	-	-	0.005	0.003	0.004	0.003
School FE	-	-			Yes	Yes

1 Source: Author's calculations using PISA 2003 data. Sample size = 29,845 in models 1,2, 4 and 5. Sample size = 28,336 in model 3 where some observations are dropped due to the fixed effect perfectly predicting the response. Dependent variable in all models was whether respondent expected to obtain a degree (coded 0 if they did not and 1 if they did)

	Model 4	
	Beta	**SE**
Gender (Ref: Girl)		
Boy	-0.797	0.058
ESCS (Ref: Lowest Quartile "Disadvantaged")		
Second Quartile	0.332	0.112
Third Quartile	0.920	0.102
Highest Quartile "Advantaged"	1.564	0.127
Maths Ability (Ref: Lowest Ability Quintile)		
Second Quintile	0.317	0.120
Third Quintile	0.756	0.118
Fourth Quintile	1.130	0.113
Highest Quintile	1.748	0.122
Immigrant Status (Ref: Native)		
Immigrant	-0.160	0.298
ESCS*Immigrant		
Second Quartile, Immigrant	0.173	0.424
Third Quartile, Immigrant	0.029	0.464
Highest Quartile, Immigrant	-0.222	0.367
Constant	-1.153	0.147

	Model 5	
	Beta	**SE**
Gender (Ref: Girl)		
Boy	-0.592	0.056
ESCS (Ref: Lowest Quartile "Disadvantaged")		
Second Quartile	0.302	0.102
Third Quartile	0.924	0.093
Highest Quartile	1.540	0.119
Maths Ability	0.528	0.087
ESCS*Maths Ability		
Second ESCS Quartile*Maths Ability	0.054	0.117
Third ESCS Quartile*Maths Ability	-0.038	0.110
Highest ESCS Quartile*Maths Ability	0.101	0.121
Constant	-0.304	0.091

Netherlands

	Model 1		Model 2		Model 3	
	Beta	**SE**	**Beta**	**SE**	**Beta**	**SE**
Gender (Ref: Girl)						
Boy	-0.201	0.084	-0.369	0.090	-0.212	0.096
Books (Ref: Low)						
High	0.855	0.097	0.392	0.096	0.249	0.113
Missing	0.039	0.433	-0.202	0.434	-0.226	0.547
Parental Education (Ref: Low)						
Medium	0.186	0.157	0.288	0.172	0.542	0.169
High	0.624	0.156	0.734	0.157	0.829	0.168
Missing	-1.185	0.389	-0.490	0.433	-0.669	0.563
Occupation spline 0-10th percentile	0.035	0.036	0.045	0.040	0.045	0.040
Occupation spline 11-25th percentile	0.072	0.029	0.049	0.028	0.028	0.032
Occupation spline 26-50th percentile	0.023	0.017	0.010	0.019	0.010	0.021
Occupation spline 50-75th percentile	0.019	0.010	0.000	0.011	-0.007	0.011
Occupation spline 76-90th percentile	0.107	0.067	0.102	0.076	0.084	0.085
Occupation spline 91-100th percentile	0.003	0.017	0.014	0.025	0.014	0.023
Immigrant Status (Ref: Native)						
Immigrant	2.115	1.609	3.982	1.669	3.952	1.812
Books*Immigrant						
High Books, Immigrant	-0.230	0.198	-0.283	0.259	-0.042	0.293
Missing Books, Immigrant	0.405	0.788	1.279	0.791	1.061	0.765
Parental Education*Immigrant						
Medium Education, Immigrant	-0.393	0.284	-0.489	0.332	-0.607	0.345
High Education, Immigrant	-0.540	0.302	-0.683	0.324	-0.857	0.325
Missing Education, Immigrant	1.078	0.662	0.413	0.636	0.469	0.801
Occupation 0-10 * Immigrant	-0.031	0.060	-0.069	0.062	-0.093	0.065
Occupation 11-25 * Immigrant	-0.031	0.061	-0.033	0.058	0.024	0.058
Occupation 26-50 * Immigrant	-0.024	0.032	-0.012	0.034	-0.016	0.036
Occupation 51-75 * Immigrant	-0.007	0.027	-0.028	0.027	-0.016	0.033
Occupation 76-90 * Immigrant	-0.129	0.174	-0.145	0.216	-0.120	0.239
Occupation 91-100 * Immigrant	0.082	0.044	0.079	0.054	0.038	0.056
Ability spline 0-10th percentile	-	-	-0.001	0.005	-0.001	0.007
Ability spline 11-25th percentile	-	-	0.012	0.005	0.010	0.006
Ability spline 26-50th percentile	-	-	0.012	0.003	0.004	0.003
Ability spline 50-75th percentile	-	-	0.026	0.003	0.018	0.003
Ability spline 76-90th percentile	-	-	0.011	0.004	0.003	0.005
Ability spline 91-100th percentile	-	-	0.006	0.005	0.005	0.005
School FE	-	-			Yes	Yes

1 Source: Author’s calculations using PISA 2003 data. Sample size = 3,902 in models 1,2, 4 and 5. Sample size = 3,645 in model 3 where some observations are dropped due to the fixed effect perfectly predicting the response. Dependent variable in all models was whether respondent expected to obtain a degree (coded 0 if they did not and 1 if they did)

	Model 4	
	Beta	**SE**
Gender (Ref: Girl)		
Boy	-0.364	0.084
ESCS (Ref: Lowest Quartile "Disadvantaged")		
Second Quartile	0.542	0.153
Third Quartile	1.064	0.169
Highest Quartile "Advantaged"	1.598	0.170
Maths Ability (Ref: Lowest Ability Quintile)		
Second Quintile	0.653	0.178
Third Quintile	1.479	0.183
Fourth Quintile	2.716	0.187
Highest Quintile	3.585	0.192
Immigrant Status (Ref: Native)		
Immigrant	1.536	0.225
ESCS*Immigrant		
Second Quartile, Immigrant	-0.567	0.284
Third Quartile, Immigrant	-0.905	0.308
Highest Quartile, Immigrant	-1.167	0.321
Constant	-3.082	0.223

	Model 5	
	Beta	**SE**
Gender (Ref: Girl)		
Boy	0.066	0.113
ESCS (Ref: Lowest Quartile "Disadvantaged")		
Second Quartile	0.534	0.171
Third Quartile	1.068	0.173
Highest Quartile	1.616	0.194
Maths Ability	1.621	0.224
ESCS*Maths Ability		
Second ESCS Quartile*Maths Ability	0.067	0.253
Third ESCS Quartile*Maths Ability	-0.175	0.260
Highest ESCS Quartile*Maths Ability	0.098	0.282
Constant	-1.467	0.171

New Zealand

	Model 1		Model 2		Model 3	
	Beta	**SE**	**Beta**	**SE**	**Beta**	**SE**
Gender (Ref: Girl)						
Boy	-0.266	0.086	-0.385	0.090	-0.378	0.105
Books (Ref: Low)						
High	0.770	0.103	0.530	0.107	0.478	0.115
Missing	0.116	0.381	0.050	0.429	0.091	0.445
Parental Education (Ref: Low)						
Medium	0.362	0.116	0.312	0.119	0.359	0.149
High	1.455	0.138	1.353	0.139	0.804	0.119
Missing	-0.004	0.180	0.149	0.180	0.255	0.204
Occupation spline 0-10th percentile	-0.041	0.038	-0.033	0.036	-0.048	0.037
Occupation spline 11-25th percentile	0.041	0.026	0.030	0.025	0.034	0.027
Occupation spline 26-50th percentile	0.018	0.017	0.006	0.017	0.003	0.017
Occupation spline 50-75th percentile	0.013	0.012	0.006	0.013	0.004	0.013
Occupation spline 76-90th percentile	-0.014	0.065	-0.024	0.065	0.005	0.066
Occupation spline 91-100th percentile	0.043	0.018	0.046	0.019	0.044	0.018
Immigrant Status (Ref: Native)						
Immigrant	0.827	1.622	0.853	1.729	0.513	1.808
Books*Immigrant						
High Books, Immigrant	-0.369	0.154	-0.430	0.162	-0.278	0.165
Missing Books, Immigrant	0.177	0.656	0.344	0.661	0.590	0.736
Parental Education*Immigrant						
Medium Education, Immigrant	-0.188	0.173	-0.115	0.178	0.022	0.231
High Education, Immigrant	-0.151	0.218	-0.095	0.235	-0.078	0.188
Missing Education, Immigrant	0.117	0.277	0.069	0.295	-0.048	0.305
Occupation 0-10 * Immigrant	0.013	0.061	0.014	0.065	0.012	0.067
Occupation 11-25 * Immigrant	-0.043	0.037	-0.039	0.038	-0.033	0.041
Occupation 26-50 * Immigrant	0.001	0.026	0.001	0.027	0.011	0.028
Occupation 51-75 * Immigrant	-0.001	0.023	0.009	0.024	0.001	0.024
Occupation 76-90 * Immigrant	0.025	0.101	-0.053	0.111	-0.003	0.110
Occupation 91-100 * Immigrant	-0.029	0.023	-0.034	0.025	-0.024	0.024
Ability spline 0-10th percentile	-	-	0.004	0.003	0.005	0.004
Ability spline 11-25th percentile	-	-	0.004	0.004	0.005	0.004
Ability spline 26-50th percentile	-	-	0.006	0.002	0.007	0.003
Ability spline 50-75th percentile	-	-	0.010	0.003	0.010	0.003
Ability spline 76-90th percentile	-	-	0.005	0.003	0.006	0.003
Ability spline 91-100th percentile	-	-	0.015	0.004	0.014	0.005
School FE	-	-			Yes	Yes

1 Source: Author’s calculations using PISA 2003 data. Sample size = 4,447 in models 1,2, 4 and 5. Sample size = 3,759 in model 3 where some observations are dropped due to the fixed effect perfectly predicting the response. Dependent variable in all models was whether respondent expected to obtain a degree (coded 0 if they did not and 1 if they did)

	Model 4	
	Beta	**SE**
Gender (Ref: Girl)		
Boy	-0.273	0.080
ESCS (Ref: Lowest Quartile "Disadvantaged")		
Second Quartile	0.282	0.127
Third Quartile	0.711	0.134
Highest Quartile "Advantaged"	1.402	0.140
Maths Ability (Ref: Lowest Ability Quintile)		
Second Quintile	0.408	0.128
Third Quintile	0.700	0.128
Fourth Quintile	1.158	0.123
Highest Quintile	1.960	0.124
Immigrant Status (Ref: Native)		
Immigrant	0.863	0.162
ESCS*Immigrant		
Second Quartile, Immigrant	-0.083	0.201
Third Quartile, Immigrant	-0.287	0.205
Highest Quartile, Immigrant	-0.560	0.198
Constant	-2.082	0.128

	Model 5	
	Beta	**SE**
Gender (Ref: Girl)		
Boy	0.053	0.100
ESCS (Ref: Lowest Quartile "Disadvantaged")		
Second Quartile	0.263	0.132
Third Quartile	0.646	0.135
Highest Quartile	1.278	0.152
Maths Ability	0.571	0.113
ESCS*Maths Ability		
Second ESCS Quartile*Maths Ability	0.198	0.158
Third ESCS Quartile*Maths Ability	0.108	0.153
Highest ESCS Quartile*Maths Ability	0.405	0.163
Constant	-1.427	0.109

Northern Ireland

	Model 1		Model 2		Model 3	
	Beta	**SE**	**Beta**	**SE**	**Beta**	**SE**
Gender (Ref: Girl)						
Boy	-0.4921	0.1041	-0.6963	0.1003	-0.8667	0.1272
Books (Ref: Low)						
High	0.7221	0.0967	0.3711	0.1110	0.4054	0.1161
Missing	0.4382	0.4811	0.4365	0.4477	0.2688	0.4998
Parental Education (Ref: Low)						
Medium	0.2116	0.1187	0.3334	0.1227	0.2923	0.1556
High	1.1619	0.1617	1.1871	0.1735	0.7528	0.1325
Missing	-0.9841	0.3254	-0.4590	0.3455	-0.3149	0.3736
Occupation spline 0-10th percentile	0.0427	0.0413	0.0451	0.0482	0.0367	0.0471
Occupation spline 11-25th percentile	0.0283	0.0533	-0.0144	0.0601	-0.0073	0.0617
Occupation spline 26-50th percentile	0.0046	0.0264	-0.0037	0.0271	-0.0001	0.0276
Occupation spline 50-75th percentile	0.0425	0.0206	0.0250	0.0214	0.0226	0.0230
Occupation spline 76-90th percentile	-0.0081	0.0217	-0.0088	0.0230	-0.0107	0.0234
Occupation spline 91-100th percentile	0.0074	0.0131	-0.0085	0.0135	0.0042	0.0131
Immigrant Status (Ref: Native)						
Immigrant	1.2150	2.2502	0.7569	2.5032	1.0527	2.5710
Books*Immigrant						
High Books, Immigrant	-0.2614	0.3212	-0.3487	0.3546	-0.3103	0.3458
Missing Books, Immigrant	-	-			0.0000	0.0000
Parental Education*Immigrant						
Medium Education, Immigrant	-0.2544	0.3600	-0.2493	0.3613	-0.3209	0.4595
High Education, Immigrant	-0.2433	0.5781	-0.5712	0.6461	-0.2634	0.4327
Missing Education, Immigrant	0.3244	0.8291	0.2908	0.8435	0.3580	1.0803
Occupation 0-10 * Immigrant	-0.0153	0.0836	-0.0074	0.0933	-0.0121	0.0963
Occupation 11-25 * Immigrant	0.2021	0.1455	0.3285	0.1513	0.2341	0.1550
Occupation 26-50 * Immigrant	-0.0793	0.0769	-0.1075	0.0750	-0.0834	0.0740
Occupation 51-75 * Immigrant	-0.0821	0.0657	-0.0732	0.0629	-0.0868	0.0636
Occupation 76-90 * Immigrant	0.2433	0.0704	0.2720	0.0745	0.2639	0.0745
Occupation 91-100 * Immigrant	-0.1185	0.0428	-0.1044	0.0499	-0.0944	0.0448
Ability spline 0-10th percentile	-	-	0.0181	0.0072	0.0223	0.0083
Ability spline 11-25th percentile	-	-	0.0019	0.0065	0.0007	0.0066
Ability spline 26-50th percentile	-	-	0.0172	0.0040	0.0160	0.0040
Ability spline 50-75th percentile	-	-	0.0075	0.0031	0.0054	0.0033
Ability spline 76-90th percentile	-	-	0.0111	0.0038	0.0097	0.0038
Ability spline 91-100th percentile	-	-	0.0161	0.0047	0.0155	0.0048
School FE	-	-			Yes	Yes

1 Source: Author's calculations using PISA 2003 data. Sample size = 2,829 in models 1,2, 4 and 5. Sample size = 2,614 in model 3 where some observations are dropped due to the fixed effect perfectly predicting the response. Dependent variable in all models was whether respondent expected to obtain a degree (coded 0 if they did not and 1 if they did)

	Model 4	
	Beta	SE
Gender (Ref: Girl)		
Boy	-0.641	0.093
ESCS (Ref: Lowest Quartile "Disadvantaged")		
Second Quartile	0.320	0.195
Third Quartile	1.030	0.166
Highest Quartile "Advantaged"	1.394	0.183
Maths Ability (Ref: Lowest Ability Quintile)		
Second Quintile	0.688	0.254
Third Quintile	1.347	0.222
Fourth Quintile	1.856	0.199
Highest Quintile	2.624	0.222
Immigrant Status (Ref: Native)		
Immigrant	0.176	0.367
ESCS*Immigrant		
Second Quartile, Immigrant	1.323	0.513
Third Quartile, Immigrant	0.677	0.457
Highest Quartile, Immigrant	0.344	0.455
Constant	-2.735	0.197

	Model 5	
	Beta	SE
Gender (Ref: Girl)		
Boy	-0.355	0.103
ESCS (Ref: Lowest Quartile "Disadvantaged")		
Second Quartile	0.310	0.199
Third Quartile	1.059	0.173
Highest Quartile	1.313	0.195
Maths Ability	1.262	0.178
ESCS*Maths Ability		
Second ESCS Quartile*Maths Ability	-0.259	0.251
Third ESCS Quartile*Maths Ability	-0.399	0.227
Highest ESCS Quartile*Maths Ability	-0.039	0.210
Constant	-1.615	0.149

Norway

	Model 1		Model 2		Model 3	
	Beta	**SE**	**Beta**	**SE**	**Beta**	**SE**
Gender (Ref: Girl)						
Boy	-0.415	0.081	-0.549	0.085	-0.516	0.088
Books (Ref: Low)						
High	0.718	0.105	0.460	0.106	0.473	0.115
Missing	0.890	0.352	0.817	0.377	1.028	0.402
Parental Education (Ref: Low)						
Medium	0.605	0.167	0.514	0.169	0.305	0.211
High	1.345	0.178	1.389	0.186	0.841	0.176
Missing	-0.110	0.328	0.000	0.332	0.004	0.338
Occupation spline 0-10th percentile	0.029	0.033	0.033	0.034	0.011	0.035
Occupation spline 11-25th percentile	0.044	0.030	0.023	0.031	0.034	0.032
Occupation spline 26-50th percentile	0.047	0.017	0.043	0.017	0.033	0.018
Occupation spline 50-75th percentile	0.003	0.009	-0.009	0.010	-0.009	0.010
Occupation spline 76-90th percentile	0.251	0.082	0.219	0.083	0.217	0.090
Occupation spline 91-100th percentile	0.010	0.016	0.010	0.017	0.033	0.016
Immigrant Status (Ref: Native)						
Immigrant	3.380	1.767	3.843	1.721	3.544	1.787
Books*Immigrant						
High Books, Immigrant	-0.435	0.243	-0.486	0.252	-0.343	0.265
Missing Books, Immigrant			0.000	0.000	0.000	0.000
Parental Education*Immigrant						
Medium Education, Immigrant	-0.871	0.384	-0.838	0.382	-0.970	0.434
High Education, Immigrant	-0.784	0.418	-0.759	0.420	-0.675	0.399
Missing Education, Immigrant	-0.385	1.000	-0.369	0.918	-0.207	0.927
Occupation 0-10 * Immigrant	-0.059	0.058	-0.075	0.057	-0.070	0.062
Occupation 11-25 * Immigrant	0.040	0.069	0.065	0.073	0.037	0.075
Occupation 26-50 * Immigrant	-0.073	0.052	-0.080	0.054	-0.070	0.052
Occupation 51-75 * Immigrant	0.059	0.027	0.068	0.027	0.071	0.027
Occupation 76-90 * Immigrant	-0.647	0.220	-0.666	0.249	-0.609	0.255
Occupation 91-100 * Immigrant	0.057	0.039	0.053	0.042	0.043	0.042
Ability spline 0-10th percentile	-	-	0.000	0.004	-0.001	0.004
Ability spline 11-25th percentile	-	-	0.002	0.004	0.001	0.005
Ability spline 26-50th percentile	-	-	0.009	0.003	0.010	0.003
Ability spline 50-75th percentile	-	-	0.011	0.003	0.009	0.003
Ability spline 76-90th percentile	-	-	0.001	0.003	0.003	0.003
Ability spline 91-100th percentile	-	-	0.010	0.003	0.010	0.003
School FE	-	-			Yes	Yes

1 Source: Author’s calculations using PISA 2003 data. Sample size = 4,023 in models 1,2, 4 and 5. Sample size = 3,888 in model 3 where some observations are dropped due to the fixed effect perfectly predicting the response. Dependent variable in all models was whether respondent expected to obtain a degree (coded 0 if they did not and 1 if they did)

	Model 4	
	Beta	**SE**
Gender (Ref: Girl)		
Boy	-0.502	0.084
ESCS (Ref: Lowest Quartile "Disadvantaged")		
Second Quartile	0.695	0.159
Third Quartile	1.058	0.162
Highest Quartile "Advantaged"	1.808	0.160
Maths Ability (Ref: Lowest Ability Quintile)		
Second Quintile	0.367	0.138
Third Quintile	0.801	0.151
Fourth Quintile	1.161	0.148
Highest Quintile	1.603	0.146
Immigrant Status (Ref: Native)		
Immigrant	0.950	0.231
ESCS*Immigrant		
Second Quartile, Immigrant	-0.743	0.297
Third Quartile, Immigrant	-0.377	0.314
Highest Quartile, Immigrant	-0.556	0.319
Constant	-2.799	0.169

	Model 5	
	Beta	**SE**
Gender (Ref: Girl)		
Boy	-0.146	0.087
ESCS (Ref: Lowest Quartile "Disadvantaged")		
Second Quartile	0.737	0.168
Third Quartile	1.066	0.159
Highest Quartile	1.844	0.163
Maths Ability	0.948	0.187
ESCS*Maths Ability		
Second ESCS Quartile*Maths Ability	-0.347	0.216
Third ESCS Quartile*Maths Ability	-0.380	0.220
Highest ESCS Quartile*Maths Ability	-0.326	0.201
Constant	-2.193	0.145

Poland

	Model 1		Model 2		Model 3	
	Beta	**SE**	**Beta**	**SE**	**Beta**	**SE**
Gender (Ref: Girl)						
Boy	-0.794	0.069	-0.986	0.081	-0.979	0.083
Books (Ref: Low)						
High	0.814	0.075	0.546	0.082	0.638	0.085
Missing	0.119	0.486	0.682	0.537	1.078	0.499
Parental Education (Ref: Low)						
Medium	0.259	0.102	0.237	0.102	0.278	0.120
High	1.368	0.121	1.432	0.130	0.805	0.128
Missing			0.000	0.000	0.000	0.000
Occupation spline 0-10th percentile	-0.130	0.061	-0.180	0.059	-0.148	0.064
Occupation spline 11-25th percentile	0.092	0.025	0.090	0.026	0.085	0.028
Occupation spline 26-50th percentile	0.009	0.013	0.003	0.014	0.005	0.015
Occupation spline 50-75th percentile	0.060	0.017	0.039	0.019	0.033	0.020
Occupation spline 76-90th percentile	0.004	0.012	-0.009	0.013	0.015	0.013
Occupation spline 91-100th percentile	0.012	0.015	0.002	0.015	0.016	0.016
Immigrant Status (Ref: Native)						
Immigrant	1.206	0.656	1.521	0.745	1.301	0.861
Books*Immigrant						
High Books, Immigrant			0.000	0.000	0.000	0.000
Missing Books, Immigrant			0.000	0.000	0.000	0.000
Parental Education*Immigrant						
Medium Education, Immigrant			0.000	0.000	0.000	0.000
High Education, Immigrant			0.000	0.000	0.000	0.000
Missing Education, Immigrant			0.000	0.000	0.000	0.000
Occupation 0-10 * Immigrant			0.000	0.000	0.000	0.000
Occupation 11-25 * Immigrant			0.000	0.000	0.000	0.000
Occupation 26-50 * Immigrant			0.000	0.000	0.000	0.000
Occupation 51-75 * Immigrant			0.000	0.000	0.000	0.000
Occupation 76-90 * Immigrant			0.000	0.000	0.000	0.000
Occupation 91-100 * Immigrant			0.000	0.000	0.000	0.000
Ability spline 0-10th percentile	-	-	0.003	0.007	0.003	0.007
Ability spline 11-25th percentile	-	-	0.015	0.004	0.015	0.004
Ability spline 26-50th percentile	-	-	0.011	0.003	0.011	0.003
Ability spline 50-75th percentile	-	-	0.011	0.003	0.012	0.003
Ability spline 76-90th percentile	-	-	0.007	0.003	0.008	0.003
Ability spline 91-100th percentile	-	-	0.006	0.003	0.007	0.003
School FE	-	-			Yes	Yes

1 Source: Author’s calculations using PISA 2003 data. Sample size = 4,381 in models 1,2, 4 and 5. Sample size = 4,281 in model 3 where some observations are dropped due to the fixed effect perfectly predicting the response. Dependent variable in all models was whether respondent expected to obtain a degree (coded 0 if they did not and 1 if they did)

2 Immigrant interaction figures not given due to very small numbers (less than 1% of the sample) reporting that they area first or second generation immigrant. Hence no data to support these estimation

	Model 4	
	Beta	SE
Gender (Ref: Girl)		
Boy	-0.950	0.076
ESCS (Ref: Lowest Quartile "Disadvantaged")		
Second Quartile	0.868	0.146
Third Quartile	1.385	0.143
Highest Quartile "Advantaged"	2.247	0.137
Maths Ability (Ref: Lowest Ability Quintile)		
Second Quintile	0.663	0.155
Third Quintile	1.240	0.165
Fourth Quintile	1.805	0.163
Highest Quintile	2.342	0.162
Immigrant Status (Ref: Native)		
Immigrant	0.000	0.000
ESCS*Immigrant		
Second Quartile, Immigrant	0.000	0.000
Third Quartile, Immigrant	0.000	0.000
Highest Quartile, Immigrant	0.000	0.000
Constant	-3.049	0.172

	Model 5	
	Beta	SE
Gender (Ref: Girl)		
Boy	-0.591	0.075
ESCS (Ref: Lowest Quartile "Disadvantaged")		
Second Quartile	0.930	0.144
Third Quartile	1.483	0.143
Highest Quartile	2.342	0.142
Maths Ability	1.013	0.132
ESCS*Maths Ability		
Second ESCS Quartile*Maths Ability	-0.125	0.178
Third ESCS Quartile*Maths Ability	-0.346	0.164
Highest ESCS Quartile*Maths Ability	-0.306	0.161
Constant	-2.071	0.125

Portugal

	Model 1		Model 2		Model 3	
	Beta	**SE**	**Beta**	**SE**	**Beta**	**SE**
Gender (Ref: Girl)						
Boy	-0.825	0.072	-1.142	0.081	-1.082	0.089
Books (Ref: Low)						
High	0.903	0.095	0.590	0.109	0.633	0.109
Missing	-0.343	0.328	-0.005	0.468	0.085	0.505
Parental Education (Ref: Low)						
Medium	-0.401	0.174	-0.185	0.188	0.000	0.000
High	0.329	0.155	0.455	0.165	0.172	0.130
Missing	-1.057	0.323	-0.612	0.351	-0.715	0.414
Occupation spline 0-10th percentile	0.035	0.037	0.005	0.039	-0.001	0.040
Occupation spline 11-25th percentile	0.001	0.051	0.011	0.053	0.025	0.056
Occupation spline 26-50th percentile	0.056	0.016	0.033	0.017	0.033	0.018
Occupation spline 50-75th percentile	0.036	0.011	0.027	0.012	0.023	0.012
Occupation spline 76-90th percentile	0.042	0.012	0.012	0.014	0.020	0.014
Occupation spline 91-100th percentile	-0.022	0.015	-0.017	0.016	-0.012	0.017
Immigrant Status (Ref: Native)						
Immigrant	-2.722	1.879	-2.201	2.050	-2.137	1.963
Books*Immigrant						
High Books, Immigrant	-0.147	0.196	-0.437	0.256	-0.582	0.258
Missing Books, Immigrant	-	-	0.000	0.000	0.000	0.000
Parental Education*Immigrant						
Medium Education, Immigrant	0.512	0.374	0.645	0.399	0.000	0.000
High Education, Immigrant	-0.022	0.284	0.121	0.340	0.186	0.300
Missing Education, Immigrant	-0.907	0.982	-0.975	1.271	-0.959	1.285
Occupation 0-10 * Immigrant	0.115	0.082	0.098	0.090	0.104	0.087
Occupation 11-25 * Immigrant	-0.071	0.147	0.057	0.153	0.044	0.153
Occupation 26-50 * Immigrant	-0.054	0.046	-0.078	0.056	-0.074	0.057
Occupation 51-75 * Immigrant	0.055	0.030	0.034	0.037	0.029	0.039
Occupation 76-90 * Immigrant	-0.060	0.024	-0.035	0.029	-0.029	0.030
Occupation 91-100 * Immigrant	0.090	0.060	0.097	0.109	0.082	0.095
Ability spline 0-10th percentile	-	-	0.004	0.005	-0.001	0.005
Ability spline 11-25th percentile	-	-	0.021	0.004	0.019	0.004
Ability spline 26-50th percentile	-	-	0.014	0.002	0.011	0.002
Ability spline 50-75th percentile	-	-	0.012	0.003	0.010	0.003
Ability spline 76-90th percentile	-	-	0.012	0.005	0.011	0.005
Ability spline 91-100th percentile	-	-	0.017	0.006	0.016	0.006
School FE	-	-			Yes	Yes

1 Source: Author's calculations using PISA 2003 data. Sample size = 4,594 in models 1,2, 4 and 5. Sample size = 4,454 in model 3 where some observations are dropped due to the fixed effect perfectly predicting the response. Dependent variable in all models was whether respondent expected to obtain a degree (coded 0 if they did not and 1 if they did)

	Model 4	
	Beta	**SE**
Gender (Ref: Girl)		
Boy	-1.190	0.079
ESCS (Ref: Lowest Quartile "Disadvantaged")		
Second Quartile	0.662	0.113
Third Quartile	1.111	0.113
Highest Quartile "Advantaged"	1.841	0.139
Maths Ability (Ref: Lowest Ability Quintile)		
Second Quintile	1.223	0.100
Third Quintile	1.835	0.120
Fourth Quintile	2.362	0.131
Highest Quintile	3.278	0.129
Immigrant Status (Ref: Native)		
Immigrant	0.443	0.251
ESCS*Immigrant		
Second Quartile, Immigrant	-0.609	0.310
Third Quartile, Immigrant	-0.294	0.313
Highest Quartile, Immigrant	-0.771	0.364
Constant	-2.049	0.122

	Model 5	
	Beta	**SE**
Gender (Ref: Girl)		
Boy	-0.632	0.081
ESCS (Ref: Lowest Quartile "Disadvantaged")		
Second Quartile	0.668	0.117
Third Quartile	1.052	0.119
Highest Quartile	1.911	0.134
Maths Ability	1.145	0.149
ESCS*Maths Ability		
Second ESCS Quartile*Maths Ability	-0.051	0.176
Third ESCS Quartile*Maths Ability	0.234	0.171
Highest ESCS Quartile*Maths Ability	-0.163	0.195
Constant	-0.603	0.102

Scotland

	Model 1		Model 2		Model 3	
	Beta	**SE**	**Beta**	**SE**	**Beta**	**SE**
Gender (Ref: Girl)						
Boy	-0.4920	0.0950	-0.7187	0.1061	-0.6928	0.1053
Books (Ref: Low)						
High	0.9840	0.1287	0.6296	0.1395	0.7088	0.1350
Missing	-0.5581	0.4914	-0.3666	0.5088	-0.5102	0.5454
Parental Education (Ref: Low)						
Medium	0.0551	0.1373	0.0571	0.1544	0.0000	0.0000
High	0.7051	0.1448	0.7619	0.1605	0.2692	0.1364
Missing	-0.5504	0.2381	-0.2084	0.2231	-0.2948	0.2386
Occupation spline 0-10th percentile	0.0655	0.0373	0.0684	0.0406	0.0853	0.0392
Occupation spline 11-25th percentile	-0.0440	0.0268	-0.0546	0.0256	-0.0640	0.0258
Occupation spline 26-50th percentile	0.0863	0.0195	0.0716	0.0197	0.0713	0.0207
Occupation spline 50-75th percentile	-0.0160	0.0182	-0.0194	0.0194	-0.0094	0.0196
Occupation spline 76-90th percentile	0.1318	0.0748	0.0872	0.0766	0.0729	0.0797
Occupation spline 91-100th percentile	0.0223	0.0242	0.0075	0.0244	-0.0003	0.0258
Immigrant Status (Ref: Native)						
Immigrant	5.2185	1.9042	5.6132	2.0079	6.3493	2.0849
Books*Immigrant						
High Books, Immigrant	-0.3205	0.2411	-0.5329	0.2573	-0.5193	0.2704
Missing Books, Immigrant	0.2073	0.7220	-0.2553	0.6556	-0.2264	0.6524
Parental Education*Immigrant						
Medium Education, Immigrant	-0.2308	0.2814	-0.4335	0.3089	0.0000	0.0000
High Education, Immigrant	0.0989	0.2785	-0.2064	0.3119	-0.1909	0.2558
Missing Education, Immigrant	-0.3703	0.5001	-0.7655	0.5085	-0.7769	0.5788
Occupation 0-10 * Immigrant	-0.1746	0.0691	-0.1828	0.0743	-0.2139	0.0773
Occupation 11-25 * Immigrant	0.0535	0.0621	0.0480	0.0671	0.0800	0.0694
Occupation 26-50 * Immigrant	0.0036	0.0356	0.0235	0.0404	0.0128	0.0410
Occupation 51-75 * Immigrant	0.0304	0.0347	0.0087	0.0397	0.0036	0.0382
Occupation 76-90 * Immigrant	-0.1596	0.1432	-0.0986	0.1619	-0.0706	0.1637
Occupation 91-100 * Immigrant	-0.0178	0.0471	0.0043	0.0562	0.0122	0.0588
Ability spline 0-10th percentile	-	-	0.0151	0.0066	0.0136	0.0070
Ability spline 11-25th percentile	-	-	0.0099	0.0051	0.0100	0.0052
Ability spline 26-50th percentile	-	-	0.0117	0.0031	0.0127	0.0031
Ability spline 50-75th percentile	-	-	0.0168	0.0038	0.0168	0.0038
Ability spline 76-90th percentile	-	-	0.0160	0.0057	0.0166	0.0058
Ability spline 91-100th percentile	-	-	0.0101	0.0062	0.0114	0.0060
School FE	-	-			Yes	Yes

1 Source: Author's calculations using PISA 2003 data. Sample size = 2,707 in models 1,2, 4 and 5. Sample size = 2,602 in model 3 where some observations are dropped due to the fixed effect perfectly predicting the response. Dependent variable in all models was whether respondent expected to obtain a degree (coded 0 if they did not and 1 if they did)

	Model 4	
	Beta	**SE**
Gender (Ref: Girl)		
Boy	-0.722	0.104
ESCS (Ref: Lowest Quartile "Disadvantaged")		
Second Quartile	0.401	0.159
Third Quartile	0.875	0.145
Highest Quartile "Advantaged"	1.746	0.184
Maths Ability (Ref: Lowest Ability Quintile)		
Second Quintile	0.678	0.184
Third Quintile	1.264	0.164
Fourth Quintile	1.964	0.172
Highest Quintile	3.012	0.194
Immigrant Status (Ref: Native)		
Immigrant	0.254	0.241
ESCS*Immigrant		
Second Quartile, Immigrant	0.399	0.336
Third Quartile, Immigrant	0.210	0.357
Highest Quartile, Immigrant	-0.181	0.326
Constant	-1.777	0.166

	Model 5	
	Beta	**SE**
Gender (Ref: Girl)		
Boy	-0.366	0.107
ESCS (Ref: Lowest Quartile "Disadvantaged")		
Second Quartile	0.230	0.155
Third Quartile	0.740	0.155
Highest Quartile	1.533	0.172
Maths Ability	0.987	0.174
ESCS*Maths Ability		
Second ESCS Quartile*Maths Ability	0.047	0.218
Third ESCS Quartile*Maths Ability	0.037	0.227
Highest ESCS Quartile*Maths Ability	0.382	0.245
Constant	-0.641	0.128

Slovakia

	Model 1		Model 2		Model 3	
	Beta	**SE**	**Beta**	**SE**	**Beta**	**SE**
Gender (Ref: Girl)						
Boy	-0.536	0.088	-0.943	0.085	-0.799	0.087
Books (Ref: Low)						
High	0.948	0.074	0.562	0.080	0.449	0.089
Missing	0.600	0.488	1.042	0.620	0.751	0.659
Parental Education (Ref: Low)						
Medium	-0.095	0.080	0.083	0.103	-0.035	0.110
High	0.858	0.096	0.709	0.098	0.431	0.102
Missing	-1.282	0.635	-1.122	0.759	-0.883	0.675
Occupation spline 0-10th percentile	0.040	0.022	0.036	0.023	0.031	0.026
Occupation spline 11-25th percentile	0.080	0.040	0.043	0.044	0.066	0.044
Occupation spline 26-50th percentile	0.028	0.011	0.007	0.013	-0.004	0.014
Occupation spline 50-75th percentile	0.063	0.014	0.055	0.016	0.053	0.016
Occupation spline 76-90th percentile	-0.010	0.016	-0.009	0.018	-0.007	0.018
Occupation spline 91-100th percentile	0.018	0.013	0.015	0.013	0.008	0.014
Immigrant Status (Ref: Native)						
Immigrant	0.027	2.771	-0.640	3.144	-1.344	2.755
Books*Immigrant						
High Books, Immigrant	0.306	0.257	0.249	0.286	0.174	0.277
Missing Books, Immigrant	-	-	0.000	0.000	0.000	0.000
Parental Education*Immigrant						
Medium Education, Immigrant	-0.566	0.389	-0.803	0.398	-0.667	0.427
High Education, Immigrant	-0.246	0.333	-0.323	0.395	-0.134	0.373
Missing Education, Immigrant	-	-	0.000	0.000	0.000	0.000
Occupation 0-10 * Immigrant	-0.008	0.097	0.018	0.110	0.041	0.100
Occupation 11-25 * Immigrant	-0.188	0.137	-0.166	0.158	-0.124	0.169
Occupation 26-50 * Immigrant	0.093	0.043	0.081	0.051	0.086	0.049
Occupation 51-75 * Immigrant	-0.016	0.062	0.012	0.072	-0.030	0.065
Occupation 76-90 * Immigrant	-0.045	0.057	-0.076	0.079	-0.049	0.073
Occupation 91-100 * Immigrant	0.054	0.046	0.066	0.041	0.104	0.045
Ability spline 0-10th percentile	-	-	0.016	0.007	0.015	0.007
Ability spline 11-25th percentile	-	-	0.014	0.005	0.014	0.005
Ability spline 26-50th percentile	-	-	0.012	0.003	0.008	0.003
Ability spline 50-75th percentile	-	-	0.019	0.003	0.017	0.003
Ability spline 76-90th percentile	-	-	0.010	0.003	0.006	0.004
Ability spline 91-100th percentile	-	-	0.010	0.003	0.006	0.003
School FE	-	-			Yes	Yes

1 Source: Author's calculations using PISA 2003 data. Sample size = 7,328 in models 1,2, 4 and 5. Sample size = 6,661 in model 3 where some observations are dropped due to the fixed effect perfectly predicting the response. Dependent variable in all models was whether respondent expected to obtain a degree (coded 0 if they did not and 1 if they did)

	Model 4	
	Beta	**SE**
Gender (Ref: Girl)		
Boy	-0.955	0.088
ESCS (Ref: Lowest Quartile "Disadvantaged")		
Second Quartile	0.797	0.105
Third Quartile	1.371	0.107
Highest Quartile "Advantaged"	2.149	0.123
Maths Ability (Ref: Lowest Ability Quintile)		
Second Quintile	1.025	0.184
Third Quintile	1.538	0.169
Fourth Quintile	2.465	0.182
Highest Quintile	3.386	0.201
Immigrant Status (Ref: Native)		
Immigrant	-0.296	0.350
ESCS*Immigrant		
Second Quartile, Immigrant	0.326	0.432
Third Quartile, Immigrant	0.392	0.444
Highest Quartile, Immigrant	0.185	0.405
Constant	-2.735	0.177

	Model 5	
	Beta	**SE**
Gender (Ref: Girl)		
Boy	-0.282	0.087
ESCS (Ref: Lowest Quartile "Disadvantaged")		
Second Quartile	0.728	0.110
Third Quartile	1.237	0.111
Highest Quartile	2.134	0.127
Maths Ability	1.384	0.167
ESCS*Maths Ability		
Second ESCS Quartile*Maths Ability	-0.123	0.194
Third ESCS Quartile*Maths Ability	-0.103	0.197
Highest ESCS Quartile*Maths Ability	-0.207	0.187
Constant	-1.293	0.112

Spain

	Model 1		Model 2		Model 3	
	Beta	**SE**	**Beta**	**SE**	**Beta**	**SE**
Gender (Ref: Girl)						
Boy	-0.806	0.074	-1.094	0.081	-1.145	0.086
Books (Ref: Low)						
High	0.955	0.079	0.579	0.086	0.555	0.087
Missing	-0.167	0.368	0.094	0.429	0.186	0.460
Parental Education (Ref: Low)						
Medium	0.280	0.085	0.279	0.092	0.259	0.132
High	0.935	0.107	0.869	0.116	0.617	0.110
Missing	-0.979	0.194	-0.718	0.210	-0.569	0.248
Occupation spline 0-10th percentile	0.008	0.025	-0.001	0.027	-0.016	0.030
Occupation spline 11-25th percentile	0.026	0.060	0.068	0.067	0.107	0.067
Occupation spline 26-50th percentile	0.024	0.011	0.013	0.012	0.007	0.012
Occupation spline 50-75th percentile	0.038	0.013	0.032	0.014	0.029	0.014
Occupation spline 76-90th percentile	0.025	0.009	0.026	0.010	0.031	0.010
Occupation spline 91-100th percentile	0.012	0.013	0.006	0.016	0.000	0.017
Immigrant Status (Ref: Native)						
Immigrant	1.598	1.818	2.936	2.062	2.263	1.991
Books*Immigrant						
High Books, Immigrant	-0.428	0.262	-0.770	0.309	-0.587	0.330
Missing Books, Immigrant	-0.653	0.852	-1.342	0.862	-1.586	0.801
Parental Education*Immigrant						
Medium Education, Immigrant	-0.426	0.298	-0.167	0.321	-0.452	0.484
High Education, Immigrant	-0.129	0.286	0.085	0.365	-0.079	0.329
Missing Education, Immigrant	-0.909	0.677	-0.963	0.685	-1.419	0.669
Occupation 0-10 * Immigrant	-0.020	0.075	-0.064	0.085	-0.038	0.084
Occupation 11-25 * Immigrant	-0.340	0.199	-0.232	0.221	-0.248	0.250
Occupation 26-50 * Immigrant	0.053	0.044	0.052	0.048	0.047	0.051
Occupation 51-75 * Immigrant	-0.044	0.052	-0.057	0.060	-0.059	0.065
Occupation 76-90 * Immigrant	-0.017	0.029	-0.028	0.033	-0.031	0.035
Occupation 91-100 * Immigrant	0.007	0.039	0.005	0.052	0.027	0.062
Ability spline 0-10th percentile	-	-	0.013	0.004	0.015	0.005
Ability spline 11-25th percentile	-	-	0.012	0.003	0.015	0.003
Ability spline 26-50th percentile	-	-	0.016	0.002	0.017	0.002
Ability spline 50-75th percentile	-	-	0.013	0.002	0.014	0.003
Ability spline 76-90th percentile	-	-	0.014	0.004	0.014	0.004
Ability spline 91-100th percentile	-	-	0.008	0.005	0.007	0.005
School FE	-	-			Yes	Yes

1 Source: Author's calculations using PISA 2003 data. Sample size = 10,776 in models 1,2, 4 and 5. Sample size = 10,365 in model 3 where some observations are dropped due to the fixed effect perfectly predicting the response. Dependent variable in all models was whether respondent expected to obtain a degree (coded 0 if they did not and 1 if they did)

	Model 4	
	Beta	**SE**
Gender (Ref: Girl)		
Boy	-1.056	0.078
ESCS (Ref: Lowest Quartile "Disadvantaged")		
Second Quartile	0.443	0.103
Third Quartile	1.028	0.106
Highest Quartile "Advantaged"	1.959	0.116
Maths Ability (Ref: Lowest Ability Quintile)		
Second Quintile	1.031	0.139
Third Quintile	1.689	0.133
Fourth Quintile	2.296	0.130
Highest Quintile	3.215	0.140
Immigrant Status (Ref: Native)		
Immigrant	0.813	0.242
ESCS*Immigrant		
Second Quartile, Immigrant	0.140	0.278
Third Quartile, Immigrant	-0.627	0.321
Highest Quartile, Immigrant	-1.148	0.373
Constant	-2.143	0.128

	Model 5	
	Beta	**SE**
Gender (Ref: Girl)		
Boy	-0.492	0.076
ESCS (Ref: Lowest Quartile "Disadvantaged")		
Second Quartile	0.418	0.103
Third Quartile	1.052	0.106
Highest Quartile	2.065	0.116
Maths Ability	1.148	0.103
ESCS*Maths Ability		
Second ESCS Quartile*Maths Ability	0.030	0.141
Third ESCS Quartile*Maths Ability	0.045	0.132
Highest ESCS Quartile*Maths Ability	0.209	0.141
Constant	-0.684	0.092

Sweden

	Model 1		Model 2		Model 3	
	Beta	**SE**	**Beta**	**SE**	**Beta**	**SE**
Gender (Ref: Girl)						
Boy	-0.432	0.068	-0.524	0.071	-0.468	0.073
Books (Ref: Low)						
High	0.602	0.096	0.357	0.103	0.389	0.099
Missing	0.290	0.301	0.192	0.290	0.340	0.293
Parental Education (Ref: Low)						
Medium	0.419	0.121	0.370	0.126	0.000	0.000
High	1.120	0.106	1.182	0.109	0.811	0.100
Missing	-0.348	0.278	-0.084	0.293	-0.113	0.318
Occupation spline 0-10th percentile	-0.027	0.041	-0.046	0.043	-0.023	0.041
Occupation spline 11-25th percentile	0.021	0.026	0.010	0.026	-0.004	0.026
Occupation spline 26-50th percentile	0.031	0.013	0.027	0.013	0.028	0.013
Occupation spline 50-75th percentile	0.016	0.010	0.007	0.011	0.015	0.011
Occupation spline 76-90th percentile	0.054	0.045	0.036	0.046	0.015	0.049
Occupation spline 91-100th percentile	0.022	0.014	0.018	0.014	0.032	0.015
Immigrant Status (Ref: Native)						
Immigrant	-1.077	1.604	-1.768	1.742	-1.568	1.616
Books*Immigrant						
High Books, Immigrant	-0.619	0.176	-0.675	0.192	-0.618	0.190
Missing Books, Immigrant	0.359	0.590	0.349	0.643	0.488	0.614
Parental Education*Immigrant						
Medium Education, Immigrant	-0.680	0.264	-0.592	0.263	0.000	0.000
High Education, Immigrant	-0.472	0.190	-0.405	0.208	-0.450	0.185
Missing Education, Immigrant	-0.138	0.507	-0.160	0.520	-0.083	0.527
Occupation 0-10 * Immigrant	0.084	0.058	0.109	0.062	0.094	0.057
Occupation 11-25 * Immigrant	0.020	0.047	0.041	0.047	0.074	0.048
Occupation 26-50 * Immigrant	-0.041	0.026	-0.049	0.027	-0.051	0.027
Occupation 51-75 * Immigrant	0.023	0.022	0.029	0.023	0.022	0.023
Occupation 76-90 * Immigrant	-0.175	0.098	-0.203	0.099	-0.166	0.104
Occupation 91-100 * Immigrant	0.028	0.030	0.027	0.031	0.011	0.032
Ability spline 0-10th percentile	-	-	-0.001	0.003	0.001	0.003
Ability spline 11-25th percentile	-	-	0.004	0.003	0.003	0.004
Ability spline 26-50th percentile	-	-	0.008	0.003	0.008	0.003
Ability spline 50-75th percentile	-	-	0.006	0.003	0.005	0.003
Ability spline 76-90th percentile	-	-	0.009	0.004	0.010	0.004
Ability spline 91-100th percentile	-	-	0.002	0.003	0.001	0.003
School FE	-	-			Yes	Yes
Constant						

1 Source: Author's calculations using PISA 2003 data. Sample size = 4,605 in models 1,2, 4 and 5. Sample size = 4,488 in model 3 where some observations are dropped due to the fixed effect perfectly predicting the response. Dependent variable in all models was whether respondent expected to obtain a degree (coded 0 if they did not and 1 if they did)

	Model 4	
	Beta	**SE**
Gender (Ref: Girl)		
Boy	-0.518	0.070
ESCS (Ref: Lowest Quartile "Disadvantaged")		
Second Quartile	0.430	0.122
Third Quartile	0.836	0.131
Highest Quartile "Advantaged"	1.806	0.128
Maths Ability (Ref: Lowest Ability Quintile)		
Second Quintile	0.191	0.115
Third Quintile	0.641	0.122
Fourth Quintile	0.849	0.115
Highest Quintile	1.398	0.124
Immigrant Status (Ref: Native)		
Immigrant	1.228	0.185
ESCS*Immigrant		
Second Quartile, Immigrant	-0.490	0.229
Third Quartile, Immigrant	-0.459	0.240
Highest Quartile, Immigrant	-0.985	0.251
Constant	-2.151	0.137

	Model 5	
	Beta	**SE**
Gender (Ref: Girl)		
Boy	-0.273	0.084
ESCS (Ref: Lowest Quartile "Disadvantaged")		
Second Quartile	0.392	0.122
Third Quartile	0.879	0.122
Highest Quartile	1.787	0.128
Maths Ability	0.490	0.116
ESCS*Maths Ability		
Second ESCS Quartile*Maths Ability	0.301	0.152
Third ESCS Quartile*Maths Ability	-0.100	0.150
Highest ESCS Quartile*Maths Ability	0.079	0.139
Constant	-1.608	0.108

Switzerland

	Model 1		Model 2		Model 3	
	Beta	**SE**	**Beta**	**SE**	**Beta**	**SE**
Gender (Ref: Girl)						
Boy	-0.123	0.115	-0.397	0.104	-0.284	0.077
Books (Ref: Low)						
High	0.968	0.112	0.590	0.120	0.346	0.105
Missing	0.393	0.347	0.370	0.370	0.598	0.338
Parental Education (Ref: Low)						
Medium	0.406	0.162	0.431	0.156	0.070	0.235
High	1.333	0.166	1.311	0.177	0.573	0.109
Missing	-0.422	0.487	-0.021	0.601	-0.180	0.442
Occupation spline 0-10th percentile	0.129	0.084	0.183	0.085	0.079	0.060
Occupation spline 11-25th percentile	-0.028	0.051	-0.062	0.047	-0.009	0.036
Occupation spline 26-50th percentile	0.033	0.020	0.034	0.020	0.015	0.016
Occupation spline 50-75th percentile	0.028	0.054	0.010	0.054	0.029	0.042
Occupation spline 76-90th percentile	0.023	0.016	0.022	0.015	0.017	0.013
Occupation spline 91-100th percentile	0.006	0.017	-0.004	0.021	0.008	0.016
Immigrant Status (Ref: Native)						
Immigrant	3.544	2.359	5.735	2.433	2.712	1.779
Books*Immigrant						
High Books, Immigrant	-0.303	0.189	-0.428	0.192	-0.295	0.164
Missing Books, Immigrant	0.290	0.549	0.536	0.655	-0.021	0.534
Parental Education*Immigrant						
Medium Education, Immigrant	-0.085	0.247	0.087	0.240	0.107	0.379
High Education, Immigrant	0.044	0.238	0.084	0.282	0.407	0.175
Missing Education, Immigrant	-0.159	0.613	-0.219	0.734	0.328	0.613
Occupation 0-10 * Immigrant	-0.118	0.091	-0.175	0.092	-0.071	0.068
Occupation 11-25 * Immigrant	0.015	0.064	-0.005	0.060	-0.053	0.047
Occupation 26-50 * Immigrant	0.000	0.035	0.003	0.038	0.021	0.027
Occupation 51-75 * Immigrant	0.025	0.090	0.029	0.100	0.010	0.073
Occupation 76-90 * Immigrant	-0.012	0.027	-0.026	0.026	-0.019	0.022
Occupation 91-100 * Immigrant	0.005	0.028	0.004	0.034	-0.009	0.025
Ability spline 0-10th percentile	-	-	-0.003	0.005	0.005	0.005
Ability spline 11-25th percentile	-	-	0.007	0.006	0.007	0.004
Ability spline 26-50th percentile	-	-	0.014	0.003	0.014	0.003
Ability spline 50-75th percentile	-	-	0.010	0.003	0.008	0.002
Ability spline 76-90th percentile	-	-	0.014	0.003	0.014	0.003
Ability spline 91-100th percentile	-	-	0.012	0.003	0.010	0.003
School FE	-	-			Yes	Yes

1 Source: Author's calculations using PISA 2003 data. Sample size = 8,393 in models 1,2, 4 and 5. Sample size = 6,529 in model 3 where some observations are dropped due to the fixed effect perfectly predicting the response. Dependent variable in all models was whether respondent expected to obtain a degree (coded 0 if they did not and 1 if they did)

	Model 4	
	Beta	**SE**
Gender (Ref: Girl)		
Boy	-0.274	0.100
ESCS (Ref: Lowest Quartile "Disadvantaged")		
Second Quartile	0.784	0.283
Third Quartile	1.582	0.253
Highest Quartile "Advantaged"	2.583	0.261
Maths Ability (Ref: Lowest Ability Quintile)		
Second Quintile	0.431	0.236
Third Quintile	1.116	0.216
Fourth Quintile	1.724	0.233
Highest Quintile	2.685	0.208
Immigrant Status (Ref: Native)		
Immigrant	1.370	0.319
ESCS*Immigrant		
Second Quartile, Immigrant	-0.672	0.381
Third Quartile, Immigrant	-0.813	0.400
Highest Quartile, Immigrant	-0.903	0.342
Constant	-4.716	0.310

	Model 5	
	Beta	**SE**
Gender (Ref: Girl)		
Boy	0.250	0.155
ESCS (Ref: Lowest Quartile "Disadvantaged")		
Second Quartile	0.661	0.283
Third Quartile	1.261	0.234
Highest Quartile	2.313	0.259
Maths Ability	0.641	0.300
ESCS*Maths Ability		
Second ESCS Quartile*Maths Ability	0.274	0.338
Third ESCS Quartile*Maths Ability	0.604	0.318
Highest ESCS Quartile*Maths Ability	0.326	0.319
Constant	-3.463	0.217

USA

	Model 1		Model 2		Model 3	
	Beta	**SE**	**Beta**	**SE**	**Beta**	**SE**
Gender (Ref: Girl)						
Boy	-0.358	0.071	-0.418	0.073	-0.3915	0.0756
Books (Ref: Low)						
High	0.538	0.082	0.298	0.085	0.3317	0.0885
Missing	-0.066	0.315	-0.037	0.316	-0.2473	0.3321
Parental Education (Ref: Low)						
Medium	0.206	0.089	0.240	0.092	0.0756	0.1208
High	1.266	0.091	1.256	0.094	0.8694	0.0919
Missing	0.056	0.286	0.218	0.288	0.2054	0.2860
Occupation spline 0-10th percentile	0.009	0.028	-0.009	0.029	-0.0092	0.0278
Occupation spline 11-25th percentile	0.036	0.014	0.030	0.014	0.0235	0.0143
Occupation spline 26-50th percentile	-0.001	0.012	-0.007	0.012	-0.0115	0.0124
Occupation spline 50-75th percentile	0.005	0.013	-0.001	0.013	0.0093	0.0137
Occupation spline 76-90th percentile	0.111	0.058	0.068	0.059	0.0491	0.0603
Occupation spline 91-100th percentile	-0.013	0.017	-0.003	0.017	-0.0033	0.0178
Immigrant Status (Ref: Native)						
Immigrant	-0.389	1.467	-0.654	1.445	-0.9607	1.3500
Books*Immigrant						
High Books, Immigrant	-0.122	0.203	-0.136	0.217	-0.0693	0.2193
Missing Books, Immigrant	-	-			0.0000	0.0000
Parental Education*Immigrant						
Medium Education, Immigrant	0.281	0.187	0.262	0.194	0.3059	0.2498
High Education, Immigrant	0.117	0.265	0.103	0.273	0.2597	0.2237
Missing Education, Immigrant	-0.422	0.593	-0.578	0.620	-0.5835	0.6117
Occupation 0-10 * Immigrant	0.023	0.053	0.035	0.052	0.0380	0.0493
Occupation 11-25 * Immigrant	0.002	0.029	0.003	0.029	0.0009	0.0300
Occupation 26-50 * Immigrant	0.014	0.024	0.009	0.024	0.0102	0.0254
Occupation 51-75 * Immigrant	-0.048	0.034	-0.038	0.034	-0.0429	0.0364
Occupation 76-90 * Immigrant	0.153	0.132	0.141	0.137	0.1337	0.1445
Occupation 91-100 * Immigrant	-0.033	0.037	-0.040	0.037	-0.0300	0.0370
Ability spline 0-10th percentile	-	-	0.005	0.003	0.0035	0.0026
Ability spline 11-25th percentile	-	-	0.006	0.003	0.0084	0.0029
Ability spline 26-50th percentile	-	-	0.006	0.002	0.0063	0.0023
Ability spline 50-75th percentile	-	-	0.005	0.002	0.0061	0.0024
Ability spline 76-90th percentile	-	-	0.009	0.003	0.0099	0.0034
Ability spline 91-100th percentile	-	-	0.001	0.004	0.0031	0.0043
School FE	-	-	-	-	Yes	Yes

1 Source: Author's calculations using PISA 2003 data. Sample size = 5,419 in models 1,2, 4 and 5. Sample size = 5,117 in model 3 where some observations are dropped due to the fixed effect perfectly predicting the response. Dependent variable in all models was whether respondent expected to obtain a degree (coded 0 if they did not and 1 if they did)

	Model 4	
	Beta	SE
Gender (Ref: Girl)		
Boy	-0.3371	0.0712
ESCS (Ref: Lowest Quartile "Disadvantaged")		
Second Quartile	0.3373	0.1029
Third Quartile	0.9455	0.0982
Highest Quartile "Advantaged"	1.5095	0.1208
Maths Ability (Ref: Lowest Ability Quintile)		
Second Quintile	0.5015	0.1012
Third Quintile	0.7648	0.1043
Fourth Quintile	0.9400	0.1064
Highest Quintile	1.6460	0.1226
Immigrant Status (Ref: Native)		
Immigrant	0.1995	0.1255
ESCS*Immigrant		
Second Quartile, Immigrant	0.2918	0.2306
Third Quartile, Immigrant	0.2589	0.2186
Highest Quartile, Immigrant	0.2217	0.2934
Constant	-0.6758	0.1104

	Model 5	
	Beta	SE
Gender (Ref: Girl)		
Boy	-0.235	0.106
ESCS (Ref: Lowest Quartile "Disadvantaged")		
Second Quartile	0.420	0.176
Third Quartile	0.747	0.177
Highest Quartile	1.681	0.169
Maths Ability	1.275	0.219
ESCS*Maths Ability		
Second ESCS Quartile*Maths Ability	-0.257	0.254
Third ESCS Quartile*Maths Ability	-0.171	0.279
Highest ESCS Quartile*Maths Ability	-0.127	0.253
Constant		

Appendix 4.2. The PISA 2003 index of economic, social and cultural status (ESCS) (Source: modified versions of PISA 2004a, PISA 2004b)

The PISA 2003 index of economic, social and cultural status (ESCS) is derived from three variables related to family background: the index of highest level of parental education in number of years of education according to the ISCED classification (PARED), the index of highest parental occupation status (HISEI) and the index of home possessions (HOMEPOS). Missing values for these three variables are imputed. The available data for each variable is then transformed to an international metric with OECD averages of 0 and OECD standard deviations of 1. These OECD-standardised variables were used for a principal component analysis in order to obtain ESCS scores applying an OECD population weight giving each OECD country a weight of 1000.

Computation of ESCS

The rationale for using these three components is that socio-economic status is usually seen as based on education, occupational status and income. As no direct income measure is available from the PISA data, the existence of household items is used as an approximate measure of family wealth.

The first two of these three measures have been described in section 4.3. The other variable included in the principle components analysis is the availability of certain household possessions. Using data about household possessions as an indicator of family wealth has received much attention in recent international studies in the field of education (Buchmann, 2000). Data about household assets are believed to capture wealth better than income because they reflect a more stable source of wealth. In PISA 2003, students reported the availability of 13 different household items at home. Specifically, children were asked:

In your home, do you have:

a) A desk for study

b) A room of your own

c) A quiet place to study

d) A computer you can use for school work

e) Educational software

f) A link to the Internet

g) Your own calculator

h) Classic literature (*e.g.* <author>)

i) Books of poetry

j) Works of art (*e.g.* paintings)

k) Books to help with your school work

l) A dictionary

m) A dishwasher

n) More than 100 books (recoded)

These items were then converted into a scale of home possession, as described on page 279 of OECD (2004c).

Missing values for students with one missing response and two valid responses (out of parental education, parental occupation and the index of home possessions) were imputed with predicted values plus a random component based on a regression of the variable with missing responses on the other two variables. Variables with imputed values were then transformed to an international metric with OECD averages of 0 and OECD standard deviations of 1. These OECD standardised variables were used for a principal component analysis applying an OECD population weight giving each OECD country a weight of 1000. The ESCS scores were obtained as factor scores for the first principal component with 0 being the score of an average OECD student and 1 the standard deviation across equally weighted OECD countries.

Using principal component analysis (PCA) to derive factor loading for each participating country provides insight into the extent to which there are similar relationships between the three components. Table 17.55 in OECD (2004c) shows the PCA results for the participating countries and the scale reliabilities for the z-standardised variables. Comparing results from within-country PCA reveals that patterns of factor loadings are generally similar across countries. All three components contribute more or less equally to this index with factor loadings ranging from 0.65 to 0.85. Internal consistency ranges between 0.56 and 0.77, the scale reliability for the pooled OECD sample with equally weighted country data is 0.69.

Appendix 4.3. Alternative results for the expectations of high ability disadvantaged children compared to their average but advantaged peers

In section 4.5, I investigate the expectations of disadvantaged children scoring high marks on the PISA maths test, drawing a comparison to those scoring substantially lower marks from advantaged homes. I base this analysis on 'model 4', of which a description can be found in the main text. In this, I discuss the possibility of an Ability*Advantaged interaction, though do not include this term in the final model based on evidence provided by a likelihood ratio test. In this Appendix I test the robustness of the main conclusion drawn from section 4.5 – that high ability children from disadvantaged homes hold lower expectations than their affluent but less able peers. I do this by estimating an alternative model specification. Specifically, in contrast to model 4, I:

- Drop immigrants from the analysis and focus on natives
- Enter ability as a single continuous variable, defining "high ability" as children who sit *at* the 85th percentile of the national PISA maths distribution[162]
- Include an interaction between ability and advantage[163]

This is defined formally as 'model 5' below:

Model 5: $$\log\left(\frac{\Pi(\boldsymbol{E_{ij}})}{1-\Pi(\boldsymbol{E_{ij}})}\right) = \boldsymbol{\alpha}_1 + \boldsymbol{\beta}_1.\boldsymbol{Sex_i} + \boldsymbol{\beta_2}.\boldsymbol{SES_i} + \boldsymbol{\beta_3}.\boldsymbol{Ab_i} + \boldsymbol{\beta_3}.\boldsymbol{SES_i} * \boldsymbol{Ab_i}$$

Where:

SES = A vector of four dummy variables reflecting children's socio-economic status, based upon quartiles the ESCS measure of family background described in section 4.5 (reference = bottom quartile)

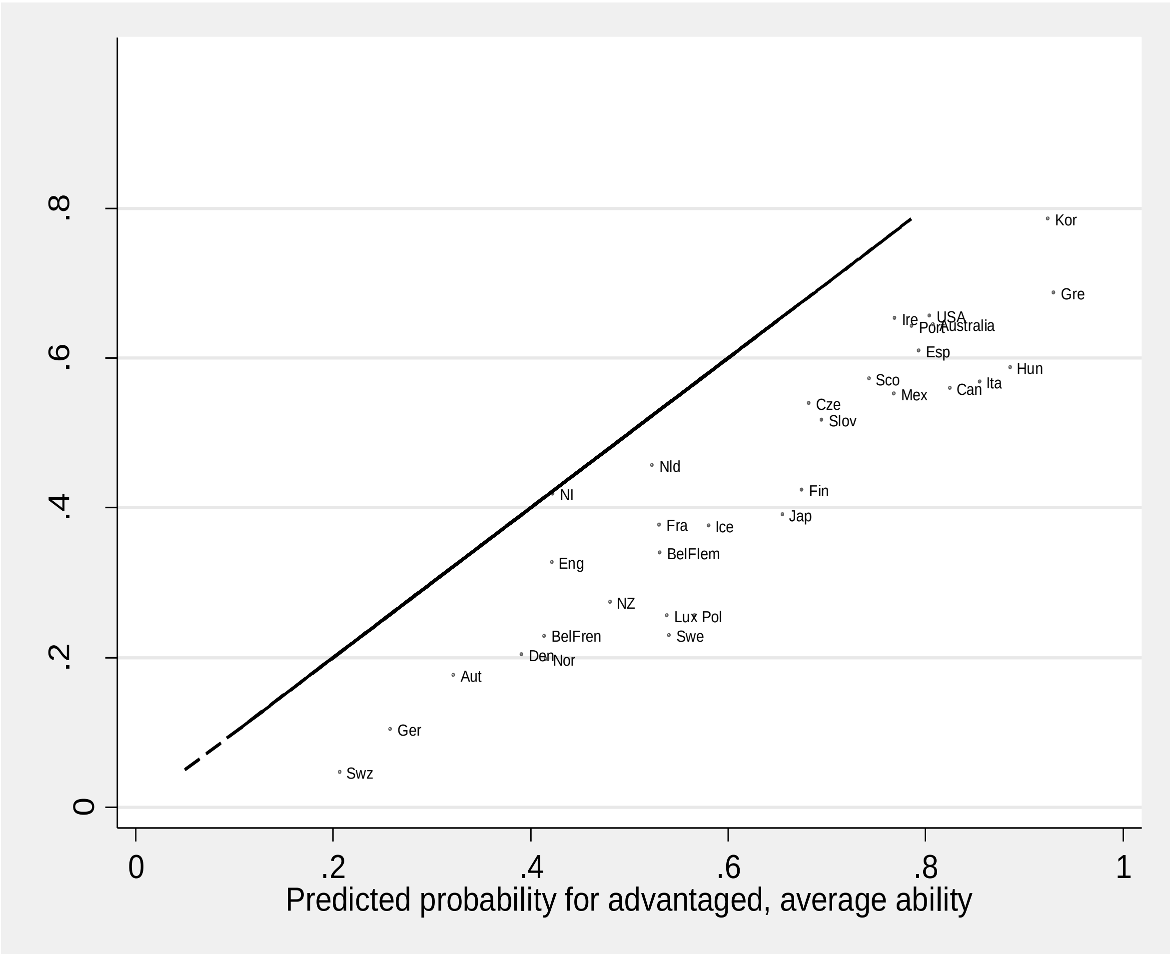
Ab = A single, continuous variable reflecting children's score on the PISA math's test. I standardise this variable within each country by subtracting from each child's score the national mean and dividing by the national standard deviation. Hence the

[162] I have also experimented with enter ability more flexibly with a series of splines. Results do not differ substantially to those presented.

[163] Although I include this interaction to test for robustness of my results, the likelihood ratio test still found only 6 countries where this substantially improved model fit at the 5% level

estimated coefficients refer to a one standard deviation increase from the mean. Using this model, I re-estimate the predicted probabilities that are presented in Figure 4.8 (full parameter estimates can be found in Appendix 4.1 as "Model 5") . Note that, as previously, all points sit below the 45 degree line (significant on 30 out of 32 occasions). In other words, I still find that high scoring disadvantaged children are less likely to believe they will complete university than lower scoring children from more affluent homes. The cross-national pattern is broadly similar to before, although England now sits closer to the 45 degree line than many other countries (infact, it is only just on the boarder of statistical significance at the 5% level). Hence it seems that this may, if anything, be *less* of a concern in England than other developed countries.

Appendix Figure 4.1. Predicted probability of a *high* ability disadvantaged girl expecting to complete university versus an advantaged girl of *average* ability



4 Source: Author's calculations using PISA 2003 data. Country by country sample sizes in Table 4.1 and Appendix 4.1. Dependent variable in regression was whether respondent expected to obtain a degree (coded 0 if they did not and 1 if they did)