Non-cognitive skills: Their formation, measurement, and interaction effects on earnings

A thesis submitted to fulfil requirements

for the degree of

Doctor of Philosophy

by

Shun Chak (Felix) Leung

School of Economics

Faculty of Arts and Social Sciences

The University of Sydney

November 2023

Contents

Abstract

My doctoral dissertation opens with an introductory chapter followed by three independent chapters on non-cognitive skills.

After the introductory chapter, the second chapter investigates the impact of university education on students' non-cognitive skills using Australian longitudinal data. To isolate the skill-building effects of tertiary education, the education decisions and non-cognitive skills - proxied by the Big Five personality traits - of 575 adolescents are followed over eight years. A standard skill production function is estimated, and it shows a robust positive relationship between university education and extraversion, and agreeableness for students from disadvantaged backgrounds. The effects are likely to operate through exposure to university life rather than through degree-specific curricula or university-specific teaching quality. As extraversion and agreeableness are associated with socially beneficial behaviours, university education may have important non-market returns.

In the third chapter, I explore the reasons why recent studies have found evidence for personality types in several large personality data sets, despite the abundance of research that has replicated the Big Five personality traits. Revisiting some of the early studies, I show that many key statistics (e.g., eigen values in a scree plot, Cronbach's alpha, the Kaiser-Meyer-Olkin index, and the Tucker-Lewis index) that are routinely used to guide factor extraction, show internal consistency, and demonstrate factor replicability do not imply the absence of clusters in the data.

Using two personality data sets, I then compare the relative model fit of a multitude of factor analysis models and Gaussian mixture models and find that a mixture model provides a better relative fit. Importantly, even the best mixture model fails to capture all the nuances in the personality data, suggesting the need for alternative latent variable models.

In the fourth chapter, I examine the interaction effects between non-cognitive and cognitive skills on labour market earnings. Earlier research has examined whether non-cognitive skills are complimentary with cognitive skills in the labour market but the findings have not been conclusive. One challenge facing researchers is that in a typical wage regression the number of two-way interactions that can be tested for is practically limited. A more systematic approach is used to find interaction effects of cognitive and non-cognitive skills on wages using Australian data. I combine the use of multivariate adaptive regression splines (MARS), a more flexible regression model, with interaction test statistics to identify any interaction effects. For females, no interaction effects on wages are found. For males, I find that the returns to Big Five conscientiousness differ by industry and between casual and permanent workers.

In the last chapter, I summarise the key findings and their implications. I then discuss the limitations and suggest potential avenues for future research.

Statement of originality

This is to certify that to the best of my knowledge, the content of this thesis is my own work. This thesis has not been submitted for any degree or other purposes.

I certify that the intellectual content of this thesis is the product of my own work and that all the assistance received in preparing this thesis and sources have been acknowledged.

Student name: Shun Chak (Felix) Leung

Signature:

Date: 30/06/2023

Authorship attribution statement

Chapter 2 of this thesis is published as Kassenboehmer, Leung, and Schurer (2018). I co-designed the study with the co-authors, analysed the data and co-wrote the paper.

Chapters 3 and 4 are entirely my own work and have not been previously published.

In addition to the statements above, in cases where I am not the corresponding author of a published item, permission to include the published material has been granted by the corresponding author.

Student name: Shun Chak (Felix) Leung

Signature:

Date: 30/06/2023

As supervisor for the candidature upon which this thesis is based, I can confirm that the authorship attribution statements above are correct.

Supervisor name: Kadir Atalay

Signature:

Date: 30/06/2023

Acknowledgements

I would like to express my deepest gratitude to my supervisors, Peter Exterkate, Stefanie Schurer, and Kadir Atalay, for their guidance and support. I also wish to thank the PhD coordinators over the years, including Stephen Whelan, Rebecca Edwards, and Kadir Atalay. Their support and encouragement have been invaluable. I am grateful to the staff at Higher Degree by Research Administration Centre and the School of Economics for their assistance and professionalism.

I would also like to thank the participants at the Microeconometrics and Public Policy Working Group, the 2016 PhD Conference in Economics and Business, and other conferences. These events have given me the opportunity to learn from the best in the field, and it was an honour to present my work at these venues.

Finally, I would like to thank my friends, family, and partner for their unwavering support throughout this journey.

Chapter 1

Introduction

Non-cognitive skills, including social and emotional skills, are an important source of human capital differences. Understanding these differences has had important implications for education, labour and health economics. Human capital theory has traditionally emphasized the importance of cognitive skills. This has changed with the abundance of research that has found associations between non-cognitive skills and a wide range of socioeconomic outcomes (Heckman, Jagelka, and Kautz, 2021; Almlund et al., 2011a).

Research on early childhood interventions has shown that investment in non-cognitive skills in the formative years has had high returns and can be an effective policy tool to address inequality. The foundational Perry Preschool Program has demonstrated long-term benefits for participants (Heckman, Jagelka, and Kautz, 2021; Heckman, Pinto, and Savelyev, 2013). While the initial gains in IQ has faded over time, the positive impact the program had on life outcomes persisted. Research has attributed these positive effects to improvement in character skills. Other examples of successful interventions include the Carolina Abecedarian Project and the Carolina Approach to Responsive Education, with an estimated rate of return of 13.7 percent per annum (García et al., 2020). These early childhood interventions have shown the effectiveness of high-quality programs on fostering socioemotional skills of children from disadvantaged families.

In this thesis, my objective is to enhance our understanding of non-cognitive skills by exploring three key questions:

1. What effects does university education have on non-cognitive skills?
2. Why have researchers found personality types despite the abundance of research replicating the Big Five personality traits?
3. Are non-cognitive skills complimentary with any cognitive skills in the labour market?

The first question is about the formation/development of non-cognitive skills during adolescence, the second about their measurement, and the third on their interaction effects, if any, on labour market earnings. Each question is addressed in turn in the next three chapters.

Early childhood interventions such as the Perry Preschool Program have shown that high-quality programs can foster social and emotional skills that have lasting beneficial effects into adulthood. While returns on investment are higher the earlier in the participants' life, research on how non-cognitive skills can be developed in adolescence is nonetheless important. Teenage years present another chance to address any disparity in both cognitive and non-cognitive skills between those from better socioeconomic families and those from disadvantaged families. Vedel (2016) found personality differences across academic majors. This suggests university education may have a direct impact on personality. Better understanding of how university education may affect skill development can provide educators with insights on how to enhance these skills.

In chapter 2, we examine the effect of university education on students' non-cognitive skills using high-quality Australian longitudinal data. To isolate the skill-building effects of tertiary education, we follow the education decisions and non-cognitive skills - proxied by the Big Five personality traits - of 575 adolescents over eight years. Estimating a standard skill production function, we demonstrate a robust positive relationship between university education and extraversion, and agreeableness for students from disadvantaged backgrounds. The effects are likely to operate through exposure to university life rather than through degree-specific curricula or university-specific teaching quality. As extraversion and agreeableness are associated with socially beneficial behaviours, we propose that university education may have important non-market returns.

In their seminal review on personality psychology and economics, Almlund et al. (2011a) developed a theoretical framework that integrates personality traits into economic models and suggested that personality psychology has the potential to deepen our understanding of conventional economic preferences such as risk aversion and time discounting. One of the challenges to this promising line of research is that the widely used and replicated Big Five personality model has several competing alternatives. Among them the most different taxonomy to the Big Five is perhaps the RUO personality types (Asendorpf et al., 2001; Caspi and Silva, 1995): resilient, undercontrolled, and overcontrolled. The fact that a recent study by Gerlach et al. (2018) has identified four personality types in several large data sets is particularly puzzling, given the body of research that has replicated the Big Five across self-ratings and peer-ratings (McCrae and Costa, 1987) and across different cultures and languages (McCrae and Terracciano, 2005; Saucier and Goldberg, 2001; Hofstee et al., 1997; McCrae and Costa Jr, 1997). Puzzling because the underlying statistical models, mixture models and factor analysis, are incompatible. It is crucial to reconcile or at least understand these differences in order to pursue research that seeks to integrate personality traits and economic preferences.

In chapter 3, we seek to understand why recent studies have found evidence for personality types in several large personality data sets (Kerber, Roth, and Herzberg, 2021; Gerlach et al., 2018). This is a puzzling finding given the abundance of research that has replicated the Big Five. More importantly, the underlying statistical models – a linear factor model for the Big Five personality traits and a Gaussian mixture model for personality types – make incompatible assumptions about the number of clusters in the data. In a linear factor model, the observed variables are assumed to be jointly normally distributed and therefore form just one cluster. In a Gaussian mixture model, however, it is implicitly assumed that there are at least two clusters within the data. We revisit some of the early studies that replicated the Big Five and show that many key statistics (e.g., eigen values in a scree plot, Cronbach's alpha, the Kaiser-Meyer-Olkin index, and the Tucker-Lewis index) that are routinely used to guide factor extraction, show internal consistency, and demonstrate factor replicability do not imply the absence of clusters in the data. Using two personality data sets, we then compare the relative model fit of a multitude of factor analysis models and Gaussian mixture models and find that a mixture model provides a better relative fit. Importantly, we show that even the best mixture model fails to capture all the nuances in the personality data, suggesting the need for alternative latent variable models.

That both cognitive and non-cognitive skills are valued in the labour market has been well documented (See e.g. Heckman, Stixrud, and Urzua, 2006a). Earlier research has looked at whether non-cognitive skills are complimentary with cognitive skills but the findings have not been conclusive about whether such interaction effects exist. For example, Nyhus and Pons (2005) found evidence of interaction between personality and tenure, whereas Heineck (2011) did not. Green and Riddell (2003) found that cognitive skills did not interact with any unobserved skills. Deming (2017) and Weinberger (2014), on the other hand, found that cognitive and social skills had an interaction effect on wages. One challenge facing researchers is that in a typical wage regression the number of two-way interactions that can be tested for is practically limited. In chapter 4, we propose a more systematic approach to finding interaction effects of cognitive and non-cognitive skills on wages using Australian data. We combine the use of multivariate adaptive regression splines (MARS), a more flexible regression model, with interaction test statistics to identify interactions. In the Australian labour market, for females we do not find any interaction effects on wages. For males, we find that the returns to Big Five conscientiousness differ by industry and between casual and permanent workers.

Chapter 2

University education and non-cognitive skill development

2.1 Introduction

The primary goal of university education is to teach students mastery of an academic subject, human capital that is needed for a growth-oriented and innovative economy (Delbanco, 2014; DeVitis, 2013). Notwithstanding the importance of academic qualifications, employers also value non-academic qualifications that are often referred to as soft or non-cognitive skills (NCS) (Almlund et al., 2011b). David Docherty, the chief executive of the *National Centre for Universities and Business* in Britain, suggested that universities must provide society with workers who do not only have the ability to "continually learn, to think critically and theoretically", but also "to innovate and break the status quo, and to navigate in the unstable waters of the global economy" (Docherty, 2012). A recent survey of Australian graduate employers revealed that employers rank "poor or inappropriate academic qualifications or results" low as an issue in graduate hiring, while they care about "interpersonal and communication skills, attitude and work ethic, and motivation" (Graduate Careers Australia, 2014, p.6). Yet, "it is not clear how actively universities develop these traits through their courses or other aspects of university life" (Norton and Cherastidtham, 2014, p. 69).

Most of the recent literature on the NCS-building effects of the education sector focuses on pre-school programs (e.g. Heckman, Pinto, and Savelyev, 2013; Heckman et al., 2010; Chetty et al., 2011), although some studies evaluate the skill-building effects of class-room interventions or institutional changes in the secondary education sector (See Schurer, 2017, for a review). Jacob (2002); Heckman, Stixrud, and Urzua (2006b); and Lundberg (2013) are among the few studies which explore the role of NCS in shaping college choices, although they do not elaborate specifically on whether university education shapes such skills.[1] We contribute to this literature by positing that university education is an important input into the skill production function of adolescents. University education has the potential to shape NCS because such skills are still malleable during late adolescence and young adulthood (Elkins, Kassenboehmer, and Schurer, 2017; Bleidorn et al., 2013; Cobb-Clark and Schurer, 2013, 2012; Hopwood et al., 2011; Specht, Egloff, and Schmukle, 2011), and because universities provide an intensive new learning and social environment for adolescents.

To identify the effect of university education, we follow the education and NCS trajectories of a cohort of 575 adolescents over eight years using nationally-representative, longitudinal data from the Household, Income, and Labour Dynamics in Australia (HILDA) survey. The data provide measures of NCS before potential university entry, and follow up measures four and eight years later. This unique data feature allows us to use value-added and fixed effects estimation models to control for the potential endogeneity in university participation. Because of the richness of our data, we are also able to control for a large number

[1]Heckman, Stixrud, and Urzua (2006b) evaluates the impact of 13+ years of schooling on a summary measure of self-esteem and self-efficacy, relative to less than 12 years of education, which could be interpreted as the impact of college eduction on non-cognitive skills.

of key inputs into the adolescent skill production function that may vary over time and affect the decision to enter or complete university education (Elkins, Kassenboehmer, and Schurer, 2017).

Our results show that university education has positive effects on extraversion, reversing a downward sloping population trend in outward orientation as people age. It also accelerates an upward-sloping population trend in agreeableness for students from low socioeconomic status, boosting agreeableness scores from the lowest levels observed at baseline to the highest levels at the eight-year follow up. The effects are robust to controlling for individual-specific heterogeneity, time-varying shocks, work experience and modeling assumptions, and they do not differ by sex, university type or field of study. This finding suggests that the causal mechanism is likely to operate through actual exposure to university life, rather than through university-specific teaching styles or quality, or subject-specific contents. Such interpretation is strengthened by the observation that length of exposure to university life is positively associated with NCS development.

The remainder of this paper is as follows. In Section 2.2 we outline the likely mechanisms through which university education affects non-cognitive skill formation. We propose a model of selection into university by non-cognitive skills, discuss our estimation strategy and the identification assumptions in Section 2.3. In Section 2.4 we describe the study design, in particular the specifics of the Australian higher education sector and the data we use. Furthermore, we provide in Section 2.4 descriptive statistics of NCS development for university students compared to non-university students. Section 2.5 presents the estimation results, while Section 2.6 concludes. Supplementary data is provided in the appendix.

2.2 University education and non-cognitive skill formation: Mechanisms

We hypothesise that university education builds non-cognitive skills (NCS) according to the human capital model of education.[2] University education coincides with the transition from adolescence into young adulthood which has been characterised by rapid change and the peak of personality maturation. The nature of this maturation process is toward increasing levels of agreeableness, conscientiousness, and emotional stability, and decreasing levels of openness to experience and extraversion (See Elkins, Kassenboehmer, and Schurer, 2017, for a full review of the literature). We suggest that university training alters this maturation process: Theoretically, it could boost, weaken, or even reverse population trends in personality trait maturation. There are at least three avenues through which trends in personality maturation could be altered.

First, NCS could be shaped by the students' exposure to new demands on their NCS through the cognitively challenging features of the curriculum. Keeping up with course work shapes attention, openness to new ideas, persistence, and the ability to manage scarce time resources, in addition to high levels of intellectual and social engagement with teachers and peers. According to academic learning theory (ALT), studying is characterised by both "effort, isolated learning, and independent competence" (Brown et al., 1983, p.79), and classroom learning that takes place "in social settings where it is guided and broken up into segments by others" (Thomas and Rohwer Jr., 1986, p. 20). ALT further acknowledges that

[2] Individuals living in an OECD country who completed tertiary education earn, on average, 55% more than individuals who did not obtain such a qualification (OECD, 2012). These high economic returns of university education suggest that graduates acquire skills that are valued by employers. This interpretation is consistent with the human capital model of education (Mincer, 1958; Schultz, 1961). Yet, one could argue that university degrees just function as a screening device for employers to select the most innately capable workers. The argument is that university education does not teach additional skills, but students who graduate send a signal to employers that they are smart and productive workers because they were able to apply for and complete university training. This interpretation is consistent with the screening theory of university education (Arrow, 1973; Spence, 1973; Stiglitz, 1975; Weiss, 1995).

studying is a mix of "skill and will" (Paris, Lipson, and Wixson, 1983, p. 309), where will refers to a disposition to exert effort, to persist, and to seek and transform information. Thomas and Rohwer Jr. (1986) thus describe academic learning as a combination of emotional factors and cognitive abilities. Because of the challenging nature of academic learning and the demands that are placed on students, exposure to university education may alter students' NCS maturation trends.

A second avenue through which university education may shape NCS is exposure to degree-specific content or teaching styles. Most of the degrees offered at university teach technical or subject-specific skills (e.g. law, engineering, medicine), but liberal arts degrees – degrees that combine humanities and science – teach a broader framework for thought. Liberal arts degrees are intended to produce well-rounded and well-read graduates, who are contemplative, collaborative, and creative.[3] Furthermore, many universities have embarked on a mission to incorporate into their curricula "generic or transferable skills", "...skills which all graduates should possess, and which would be applicable to a wide range of tasks and contexts beyond the university setting" (Gilbert et al., 2004, p. 376).[4] If university education shapes NCS through exposure to degree-specific contents or university-specific teaching strategies, we would expect to see heterogeneity in the skill-building effects of university education across universities and degrees.

A third avenue through which university education may impact on NCS development is exposure to new peer groups and extracurricular activities including sport, politics, and art. Because students from disadvantaged backgrounds are likely

[3]Some argue that the breadth of skills taught through Liberal arts degrees makes graduates more likely to succeed in leadership roles. For a discussion of these issues see Fisher (2016).

[4]Critics in the popular press contend that universities in fact do not but should teach their students NCS, because these prepare students for a fast-paced, globally connected labour market. Journalists Laura Pappano and Thomas L. Friedman suggested that universities should teach their students creativity (New York Times, 5 Feb 2014), humility, leadership, and the ability to learn on the fly (New York Times, 22 Feb 2014). Katie Allen even went so far to claim that British undergraduate university education does nothing more than leading students into high debt and under-employability (The Guardian, 19 Aug 2015).

to be more affected by a change in peer groups through day-to-day interaction with academically inclined peers and academic staff (e.g. Terenzini et al., 1996), we would expect to observe a greater effect of university education on students from disadvantaged backgrounds.

If university education truly builds NCS for students, then we should find that the more time a student has spent at university, the greater should be the impact of university education on NCS. We can test this because we observe the amount of time the students in our data spend at university. Furthermore, since we have information on socioeconomic background, university type and course, we can directly test for the second and third hypotheses, and thus may indirectly test for the validity of the first hypothesis.

2.3 Identification of the impact of university education on NCS development

2.3.1 Endogeneity in university education

Identifying the causal impact of university education on NCS development is challenging because of the inherent endogeneity in university education. Students self-select into tertiary training if the benefit from education is greater than its cost. A typical model of education choice considers the benefit of university education in terms of net lifetime earnings (Cunha, Heckman, and Navarro, 2005):

$$\begin{aligned} UNI_i &= 1 \text{ if } E(Y_{U,i} - Y_{NU,i} - C_i | I_{i0}) > 0 \\ &= 0 \text{ otherwise.} \end{aligned} \tag{2.1}$$

$Y_{U,i}$ measures life-time earnings of the individual when entering university, and $Y_{NU,i}$ measures life-time earnings when not entering university. An individual chooses to enter university ($UNI_i = 1$) if $E(Y_{U,i} - Y_{NU,i} - C_i|I_{i0}) > 0$, which means that the net lifetime return to university is greater than the costs associated with university (C_i). I_{i0} captures the information set available to the adolescent at the time he or she has to make the decision. Cunha, Heckman, and Navarro (2005) assume that at time $t = 0$, the choice for university is based on expectations about the returns to university education, observed factors that proxy the costs associated with university education, and factors that are observed to the adolescent but that are not observed to the researcher.[5]

Costs usually include tuition fees and living expenses, opportunity costs as well as psychic costs (Jacob, 2002). Similar to Jacob (2002), we assume that NCS determine the psychic costs of university education and thus help students to more easily "navigate college life" (p. 591).[6] Both Jacob (2002) and Lundberg (2013) find empirical evidence in favor of this hypothesis. For instance, Lundberg (2013) demonstrates that individuals high on sociability and low on emotional stability are less likely to have completed a college degree, while individuals high on conscientiousness and agreeableness are more likely to do so. Therefore, selection into university is likely to depend on baseline levels of NCS.

Modeling the impact of university education on NCS development is furthermore complicated by the fact that the transition process from adolescence into young adulthood coincides also with the uptake of social responsibilities, changes in relationships and work experiences, all of which are likely to determine both education choices and the NCS maturation process (e.g. Luedtke et al., 2011; Roberts, 1997). Similarly, those factors may also be related to the probability of completing

[5]Moreover, models of major choice include an expectation about performance in the specific subject. We abstract from modeling expectations about performance in major choice because we are not modeling this choice. One could also argue that students self-select into university if they expected gains in non-cognitive skill development. One would expect self-selection by gains in NCS development if universities explicitly trained NCS as outlined in their academic strategies and curriculum contents, which Australian universities do not do. To the best of our knowledge, this possibility has also not been considered in the theoretical and empirical literature.

[6]In contrast, Heckman, Lochner, and Todd (2006); Carneiro, Hansen, and Heckman (2003); and Cunha, Heckman, and Navarro (2005) assume that psychic costs are related to cognitive ability.

university education. Thus, an evaluation of the impact of university education on NCS development needs to account for the self-selection into university education by NCS and other important confounding factors, some of which may be unobservable. In the next section, we describe our identification strategy to isolate the effect of university education from these confounding factors.

2.3.2 Modeling and estimating the production function of NCS

In this section we lay out the empirical framework within which we test the effect of university education on youth NCS. We depart our analysis from the perspective that NCS in young adulthood are the result of a cumulative dynamic process that is similar to the acquisition of cognitive and non-cognitive skills of children (Todd and Wolpin, 2003; Cunha and Heckman, 2008; Fiorini and Keane, 2014; Del Bono et al., 2016). We follow the adolescent skill production functions defined in Elkins, Kassenboehmer, and Schurer (2017) which considers input choices made by the individual including educational choices, negative and positive experiences including the uptake of social responsibilities, formation of new relationships, health changes, and some fixed mental capacity. The skill production function of individual i at age a is:

$$NCS_{ia} = NCS_a[E_i(a), X_i(a), \theta_{i0}, \varepsilon_{ia}], \tag{2.2}$$

where $E_i(a)$ are past educational inputs up until age a, $X_i(a)$ are all other relevant inputs, θ_{i0} is the initial skill endowment and ε_{ia} is measurement error in skills or age-specific shocks, which are assumed to be independent of E, X, and θ. In this flexible specification, the impacts of all inputs are allowed to vary by age. However, estimating this specification is not feasible, because information on all relevant, historical inputs and initial endowments in skills is usually not available. To control for all historical inputs into the skill production function and initial skill endowment, we follow an empirical specification widely used in the literature, which controls for these unobservable inputs by conditioning the analysis on a past or baseline value of NCS (See Todd and Wolpin, 2003; Cunha and Heckman, 2008; Fiorini and Keane, 2014; Del Bono et al., 2016; Black and Kassenboehmer, 2017). One assumption of this approach is that the impact of each input is independent of the age at which the input occurs (See Fiorini and Keane, 2014; Todd and Wolpin, 2003, for a discussion).

We adjust this so-called value-added model to the context of the young adult skill production function similar to Elkins, Kassenboehmer, and Schurer (2017) and Cobb-Clark and Schurer (2012). Our outcome measure is personality trait j in time period $t+1$ ($PT^j_{i,t+1}$):[7]

$$PT^j_{i,t+1} = PT'_{i,t}\alpha^j_0 + \alpha^j_1 U_{i,t+1} + \alpha^j_2 FOPS_{it} + X'_{i,t}\beta^j + Z'_{i,t+1}\gamma^j + \varepsilon^j_{i,t+1}. \qquad (2.3)$$

We observe the same individual i in time periods $t+1$ (2013) and t (2005). University education is measured by an indicator variable $U_{i,t+1}$ which takes the value 1 if the individual "has completed university" or "has been studying at university at least in the second year" in $t+1$. The coefficient α^j_1 is the key parameter of interest, measuring the effect of university education on NCS. The reference group consists of adolescents who did not go to university in the time period considered, but pursued alternative vocational training or work experiences. In an extension to the baseline model, we interact the university indicator with the father's occupational prestige score (FOPS, see Section 2.4.2 for a definition) measured at baseline t ($U_{i,t+1} \times FOPS_{i,t}$) to allow for heterogeneous effects by parental SES. We include this interaction term because university education is likely to change

[7]Note, we allow in our specification the impact of baseline skills on contemporaneous skills to differ from 1 in contrast to Elkins, Kassenboehmer, and Schurer (2017) and Cobb-Clark and Schurer (2012), who assume that the impact is 1 and thus use the changes in NCS between two time periods as outcome variable.

the peer group or exposure to new ideas for adolescents from low socioeconomic backgrounds to a greater extent than for adolescents from high socioeconomic backgrounds (as suggested in the third mechanism discussed in Section 2.2.

$PT_{i,t}$ is a vector of five baseline personality traits to capture all unobservable past inputs and baseline endowments in skills (θ_{i0}) (see Section 2.3.1). The vector $X^j_{i,t}$ includes baseline control variables which are associated with the decision to go to university as well as the changes in NCS between time t and $t+1$. These control variables are measured strictly before our sample members potentially enter university and include age, sex, country of birth, region and state of residence, socioeconomic status and health. The vector $Z_{i,t+1}$ captures the uptake of social responsibilities, changes in relationships, work experiences (and other relevant life events) which are important for NCS formation. It includes indicator variables for each life event that occurs after period t up until period $t+1$, and accumulated work experience in $t+1$ since period t. All variables are described in Section 2.4.2 and their summary statistics are listed in Table 2.1.

The error term $\varepsilon^j_{i,t+1}$ is assumed to be the sum of remaining individual-specific heterogeneity (μ^j_i) and period-specific shocks ($\varphi^j_{i,t+1}$). Given that we condition not only on baseline personality trait j, but also all other personality traits, we hope to proxy with these controls most of the unobservable variation in μ^j_i. In contrast, $\varphi^j_{i,t+1}$ includes all accumulated shocks until $t+1$ that are not captured by $Z_{i,t+1}$. We assume that this value-added model controls for selection into university by conditioning on baseline NCS and a set of other baseline controls. Under the assumption of zero remaining covariance between both components in $\varepsilon^j_{i,t+1}$ and university education ($U_{i,t+1}$), estimating α^j_1 with Ordinary Least Squares (OLS) would yield an unbiased impact estimate of university education on personality trait j.

To complement the value-added specification, we also use coarsened exact matching (CEM) to more transparently narrow down the control group against which the treatment group is compared [Hoetal2007}. Although matching methods also rely on the conditional independence assumption, they require fewer functional-form assumptions due to their semi-parametric nature. CEM improves upon standard matching methods, because it is less dependent on a regression model as propensity-score matching and achieves an improvement of the balance of one covariate without making the balance of other covariates between treatment and control group worse, which is a common problem with nearest-neighbour matching (for an overview see Iacus, King, and Porro (2011)). For applications in estimating educational outcomes, see Jones, Rice, and Rosa Dias (2011) and in health care expenditures, see Schurer et al. (2016).[8]

The assumptions underlying both OLS and CEM estimation of the value-added model to identify a causal impact of university education are that (a) the lagged dependent variable is a sufficient statistic for unobserved input histories and initial endowment, (b) the impact of previous inputs is independent of the age at which they occur, and (c) there is no remaining unobserved heterogeneity which correlates with university education (See Fiorini and Keane, 2014, for a discussion of these assumptions). These assumptions may be violated, specifically the latter of no remaining unobserved heterogeneity. Instead of conditioning on the lagged dependent variable, we therefore also estimate a fixed effects model

[8]We are matching youth who have entered the university track by 2013 to a statistical twin on the basis of the following categories of pre-treatment variables: Sex (0,1); Age: Being above versus being at or below age 17 in 2005 (0,1); Father's occupation class: Being above or at versus below the sample average on the occupational prestige score (0, 1); Family household income: Being above or at or below the most common income band between A$60,000 and A$70,000 (0,1); Degree of urbanization: Major urban versus non-major urban (0,1); Being from an English-speaking background (0,1); Big Five personality traits in 2005: Being above versus below the median of 4 (on a 7-point scale) for each of the five personality domains, respectively ($5 \times 0,1$). Out of 192 individuals in the treatment group, we found a perfect match for 112 adolescents (58%). The means of all relevant pre-treatment covariates, which were not balanced before matching (Table A.1 in supplementary data), were well balanced between the treatment and the control groups after matching (Table A.2 in supplementary data).

as an alternative specification in which we regress changes in personality trait j between time periods t and $t+1$ ($\Delta PT^j_{i,t+1} = PT^j_{i,t+1} - PT^j_{i,t}$) on changes in all time-varying regressors between time periods t and $t+1$ (T=2):

$$\triangle PT^j_{i,t+1} = \alpha^j_1 \triangle U_{i,t+1} + \triangle Z'_{i,t+1}\gamma^j + \triangle \varphi^j_{i,t+1}. \tag{2.4}$$

First-differencing the data allows us to eliminate the influence of all baseline control variables and all remaining unobserved time-invariant factors (μ^j_i) that correlate with $U_{i,t+1}$ and thus confound the parameter estimate of α^j_1. Under the assumption of zero covariance between $\Delta U_{i,t+1} and \Delta\varphi^j_{i,t+1}, \alpha^j_1$ identifies the causal effect of university education on changes in NCS. As we control for changes between t and $t+1$ in the most common period-specific, observable shocks and other individual choice variables through $\Delta Z_{i,t+1}$ (life events, work experience, health), we reduce the possibility of a residual correlation between $\Delta U_{i,t+1}$ and $\Delta\varphi^j_{i,t+1}$.[9] Because of data limitations, we are not able to estimate a dynamic fixed effects model which would additionally allow us to control for a lagged dependent variable, and thus for previous NCS inputs.

2.4 Study Design

2.4.1 The Australian tertiary education sector

We evaluate the impact of university education on NCS development from the perspective of the Australian tertiary education sector. The Australian higher education system consists of independent, self-governing public (38) and private (3) universities and institutions that award higher education qualifications. The Australian Government regulates the tuition fees universities can charge and subsidises both teaching and research activities. Core teaching activities are

[9]Note, fixed effects estimation comes at a cost of inefficiencies, and thus standard errors are often too large to identify significant effects, especially in smaller sample sizes.

subsidised in the magnitude of 0.95% of GDP in Australia, which is a higher contribution than in Britain (0.8%), but lower than in the United States (2.15%) (See Table B2.4 OECD, 2013). Although students pay tuition fees for their university degree, they do not face the same credit constraints as students in Britain or the US because of the existence of the Higher Education Credit System (HECS). The financial incentive to invest into a university degree in Australia is comparable to the incentives in Britain, but lower than in the US.[10] Similar to the British and US American tertiary education sector, Australian universities select students mainly on the basis of standardised high-school entry exams.

Australian university life is characterised by a large degree of diversity because of a high proportion of international students and students in postgraduate training. Almost a quarter of all on-shore students are international students, which is large in comparison to both the US (7%) and Britain (17%). International enrollments in Australia capture 8% of global share, ranking fifth world wide after the US, Britain, France, and Germany. Although postgraduate training is among the most expensive in the world, a quarter of all students study for post-graduate degrees in Australia, the same number as in the US, while the proportion is only 17% in the UK. However, full-time students are less frequent in Australia (70%) and Britain (67%) than in the US (80%) (See Moodie, 2015, for a review).

In contrast to the US, liberal arts colleges and degrees do not exist in Australia (and Britain). The idea of liberal arts degrees is to provide students with the opportunity to focus on one major in combination with general education courses in other basic subjects (e.g. philosophy, music).[11] Yet, in the past decade, almost all of Australian universities have started to offer combined science and arts degrees as an attempt

[10] For instance, the internal rate of return (IRR) for a man who obtained a university degree in Australia, as compared with a man attaining upper secondary or post-secondary non-tertiary education is 9%, similar as for a man in Britain (8.2%), but lower than for a man in the US (12.3%) (See Table A7.3a OECD, 2013).

[11] Liberal arts degrees aim at helping students to think deeply about their major choice, and the programs encourage collaboration and mentorship. For an excellent review of the nature of liberal arts degrees see Vallone (2013).

to give students greater flexibility and a wide range of skills transferable across industries.

In contrast to British and US American universities, the majority of Australian students live in capital cities (almost 80%) and many live with their parents (35%), while only 5% live in colleges or residential halls. Over 60% of students rely on a wage to finance their studies.[12] Australian student life is less focused on campus than for instance in the United States or Britain, which implies that every-day life changes less for the average Australian student than for the average British or US student. We would therefore expect that the contribution of university education on NCS formation may be stronger in other countries than in Australia.

2.4.2 The Data

To conduct our analysis, we use data from the Household, Income and Labour Dynamics in Australia (HILDA) Survey. The first wave of the annual survey began in 2001 with 19,914 panel members from 7,682 households, with a top-up sample of 5,477 individuals from 2,153 households in the eleventh wave (Summerfield et al., 2013). It collects information on a wide range of household and individual characteristics, such as labour market dynamics, household income and formation, self-assessed well-being and other health-related outcomes, educational background of both the participants and their parents, lifestyle and values. Of particular interest to our analysis is a module on personality traits that was collected as part of the self-completion questionnaires in waves 5, 9 and 13. We restrict our analysis to nine waves of data collected between 2005 and 2013.

There are 758 adolescents in the HILDA observed in wave 5 (2005) and wave 13 (2013) who were between 15 and 19 years of age in wave 5 and gave full information on their personality traits and education. We lose 17% because they

[12]These statistics are derived from a survey conducted by the Australian Bureau of Statistics (2013).

did not complete information on their personality traits or education in wave 13 and another 6% either due to missings in the control variables or because of sample restrictions described below. This leaves us with an estimation sample of 575 adolescents. Summary statistics of all variables used in the analysis are reported in Table 2.1.

Non-cognitive skills (NCS): We use the the Big-Five Personality Inventory (Goldberg, 1992a, 1993b) collected in waves 5, 9, and 13 to proxy NCS. Of the 40-item Trait Descriptive Adjectives in Saucier (1994b), 30 are included in the version used in the HILDA Survey, with an additional six from different sources. Respondents were asked to self-assess on a seven-point scale the degree to which each adjective describes them, with 1 indicating "not at all" and 7 indicating "very well". Of the 36 items, only 28 are used in the derivation of the five personality scales (extraversion, agreeableness, conscientiousness, emotional stability, and openness to experience). Eight items are not used after testing for item reliability (e.g. an item was omitted if the highest factor loading was not on the expected factor). The distribution of most traits is left-skewed, which means that a larger proportion of the sample agrees with the statements about their personality underlying each trait (See Elkins, Kassenboehmer, and Schurer, 2017; Cobb-Clark and Schurer, 2012, for a detailed description). We standardise each measure to mean 0 and standard deviation of 1.

University education: Whether an individual has university education is derived from information provided about the highest level of education attained and whether the individual is currently studying in any year between 2005 and 2013. We distinguish between two types of university education status: (1) degree completion by 2013, where a degree includes bachelor/honors, a graduate diploma or certificate, or postgraduate training, or (2) a combination of degree completion by 2013 or still studying in 2013 with at least two years at university. We dropped

Table 2.1: *Summary statistics of all control variables*

	mean	SD	min	max
Extraversion 2013	4.55	1.02	1.7	7
Agreeableness 2013	5.34	0.91	1	7
Conscientiousness 2013	4.94	1.02	1.2	7
Emotional stability 2013	4.93	1.11	1.7	7
Openness to experience 2013	4.34	1.05	1	7
Extraversion 2005	4.68	1.00	1	7
Agreeableness 2005	5.10	0.93	1	7
Conscientiousness 2005	4.52	0.99	1.2	6.8
Emotional stability 2005	4.84	1.05	1.8	7
Openness to experience 2005	4.32	1.04	1	7
At university	0.33	0.47	0	1
FOPS	49.17	23.60	4.9	100
FOPS missing	0.09	0.28	0	1
$0 \leq$ FOPS ≤ 30	0.23	0.42	0	1
$30 <$ FOPS < 70	0.42	0.49	0	1
$70 \leq$ FOPS ≤ 100	0.26	0.44	0	1
Age	24.73	1.38	23	27
Female	0.55	0.50	0	1
Country of birth				
Australia	0.95	0.22	0	1
Other English speaking	0.02	0.12	0	1
Other non-English speaking	0.04	0.19	0	1
Geographic region				
Major urban	0.67	0.47	0	1
Non-major urban	0.33	0.47	0	1
Type of secondary school				
Government school	0.67	0.47	0	1
Catholic non-government school	0.19	0.39	0	1
Other non-government school	0.13	0.34	0	1
Other	0.01	0.10	0	1
Life events since 2005				
Got married	0.14	0.35	0	1
Separated from spouse	0.27	0.45	0	1
Pregnancy	0.22	0.42	0	1
Birth/adoption of new child	0.18	0.38	0	1
Serious personal injury/illness	0.25	0.43	0	1
Serious injury/illness to family member	0.44	0.50	0	1
Death of close relative/family member	0.46	0.50	0	1
Death of a close friend	0.32	0.47	0	1
Victim of a property crime	0.28	0.45	0	1
Fired or made redundant	0.25	0.44	0	1
Changed jobs	0.77	0.42	0	1
Promoted at work	0.38	0.49	0	1
Changed residence	0.78	0.42	0	1
Years in paid work between 2005 and 2013	6.30	0.94	0	7.8
Difference in physical functioning between 2005 and 2013	0.89	20.05	−95	100
University group				
Go8	0.32	0.47	0	1
ATN	0.14	0.35	0	1
IRU	0.20	0.40	0	1
RUN	0.07	0.25	0	1
Other	0.27	0.45	0	1
Field of study				
STEM	0.18	0.39	0	1
Medicine and health related	0.21	0.41	0	1
Education	0.13	0.34	0	1
Management, Commerce, Law	0.23	0.42	0	1
Others	0.25	0.43	0	1
Observations	575			

Note: Estimation sample is 575 teenagers who were aged between 15 and 19 in 2005.
Source: HILDA, waves 5 and 13.

six individuals who were at university for the first year in the year 2013 and 24 individuals who had been at university before 2005. According to this definition, 192 individuals (33%) are or will be university graduates.[13] We observe a large degree of heterogeneity in university participation, where only 25% of adolescents from a disadvantaged background will enter the university track, while 50% do so from advantaged backgrounds. The non-university group comprises individuals who obtained some post-secondary training (teaching college diploma (8%), vocational training certificate (29%)) or who completed Year 12 (42.5%), or completed year 11 or dropped out of high school (19.5%). Hence, we compare NCS between adolescents who entered university training with adolescents who had either vocational training after high school completion or no additional qualification.

Control variables: We control for baseline characteristics measured in 2005 before adolescents potentially enter university. Since our baseline personality traits are measured at different stages of the adolescent lifecycle, we control for age by including dummy variables for each age-group (relative to age 15). This is an important set of control variables, because previous research has shown that some personality traits rapidly change during adolescence (e.g. Elkins, Kassenboehmer, and Schurer, 2017).[14] We further control for sex, country of birth (Australia, English-speaking country, any other country), region of residence (major urban area versus rural), state of residence, socioeconomic status of the father, and health. Although, we have no information on parental input factors in the skill production function, we use parental socioeconomic status to proxy parental attitude

[13]This includes 29 individuals who had just started university in March 2005 when they were interviewed, assuming that the time period at university was too short to have affected baseline NCS. There are 60 individuals in the non-university group who reported to have been enrolled in some form of tertiary education between 2006 and 2012 but who were neither at university in nor did they complete a degree by 2013. Of these 60 - what we refer to as university drop-outs - 27 provided the name of the university and their field of study. As the majority of the university dropouts stayed at university for less than two years, we left them in the non-university group. However, in a robustness check we omitted the 27 individuals for whom we could corroborate that they really studied at a university from the analysis. We are able to show that our main conclusions are not sensitive to omitting these observations. These results are provided upon request.

[14]Strictly speaking, our baseline measures of personality traits are not measured just before our sample members potentially enter university, which is particularly true for our youngest sample members. We assume that a non-linear control for age difference at baseline assessment captures a non-linear age trend in NCS maturation, so that all remaining differences are measurement error only.

to education, past educational inputs and resource constraints. Socioeconomic status is measured both by the Australian Socioeconomic Index 2006 (AUSEI06) occupational status scale (McMillan, Beavis, and Jones, 2009) - referred to as Father Occupational Prestige Score (FOPS), and household income bands. The FOPS scale is derived from the first edition of the Australian and New Zealand Standard Classification of Occupations. The reference point for the classification is when the individual was aged 14, or, if the father was not in employment or deceased, the classification would be based on any previous employment. Our FOPS measure is bound between 0 (lowest status) and 100 (highest status), and we standardise it to mean 0 and standard deviation of 1. Values of 70 and above indicate high-skilled occupations (professional, managerial, and legislative), while values of under 30 indicate low-skilled occupations (elementary and manual). In a robustness check we explore alternative definitions of FOPS based on quartiles of the FOPS distribution or occupational skill levels.

To capture further differences in previous educational inputs, we control for the type of high school from which the individual graduated (public, private catholic, private independent) as a crude proxy for quality. Finally, we also control for physical health status using the SF-36 instrument (Ware Jr, 2000), because health problems are important input factors into the skill production function of adolescents and young adults (Fletcher and Schurer, 2017; Elkins, Kassenboehmer, and Schurer, 2017).

It is possible that the decision to go to or complete university is affected by family-, employment- and health-related shocks that may also affect the maturation of NCS. We therefore control for a battery of life events that occurred in any time period after baseline similar to Elkins, Kassenboehmer, and Schurer (2017) and Cobb-Clark and Schurer (2012). Some of these life events are positive (e.g. marriage, job promotion, birth of a child), while some are negative (e.g. death of a family member, unemployment). Some of these events are under individuals' control, however, others are not (e.g. death of a spouse, becoming a property-crime victim). A full list of these life events is presented in Table 2.1. Finally, we control for the accumulated work experience of each individual either during or after graduation to ensure that changes in NCS are not driven by post-education-specific work experiences.

University type and degree of study An important component of our analysis is the exploration of heterogeneity in the effect of university education by the type of university or by degree as a potential mechanism through which university affects NCS. In 2012 of the HILDA survey, participants were asked to provide information on the name of the university where they graduated from and their field of study. We follow Norton and Cherastidtham (2014) to group universities according to their self-identified group membership which is a crude measure of their research intensity and teaching focus. In our sample, 176 out of 192 youth who completed or entered the university track provided information on which university they study or studied at, and 188 provided information on their field of study.

The Group of Eight (Go8) universities market themselves as "Australia's Leading Universities", because they are ranked as the top eight performers in national research evaluations and international standing. The Go8 universities provide general training in all sciences including medical training and the Australian equivalent of liberal arts degrees.[15] The Australian Technology Network (ATN) is a coalition of five universities that share a common focus on the practical application of tertiary studies and research. The members of this network distinguished themselves as technical colleges before they became accredited as universities. The group of Innovative Research Universities (IRU) comprises seven young

[15] For instance, Monash University, University of Sydney, and University of Adelaide offer a Bachelor of Arts and Sciences, a double degree that takes four instead of the common three years of duration to complete.

universities that share a common mode of operation and a common background, all of which have been founded in the 1960s and 1970s as research universities. The Regional Universities Network (RUN), which comprises six universities, was formed in 2011 to provide tertiary training in remote areas, and to build research expertise in agriculture, fisheries, and environmental science. Finally, there are six universities that do not belong to any network ("Other universities"), some of which are highly-ranked in terms of their research quality.[16] In our sample there are no individuals who received their university education overseas. The majority of students obtained their degree from, or currently study at, a Go8 (31.8%), other (27.3%), or IRU institution (20%). Only 14.2% and 6.8%, respectively obtained their degree from an institution of the ATN and RUN.

The field of study classification is based on the Australian Standard Classification of Education (ASCED) (Australian Bureau of Statistics, 2001). University students or graduates are grouped into five broad groups: (1) Science, Technology, Engineering or Mathematics (STEM), which includes Architecture and Environment and Agriculture (18%); (2) Medicine, Nursing, and other health-related studies (21%); (3) Education (13%); (4) Management, Commerce, and Law (23%); (5) and Society and Culture, Creative Arts and Food and Hospitality which we refer to as "Other" (25%).

2.4.3 Descriptive analysis of NCS trends

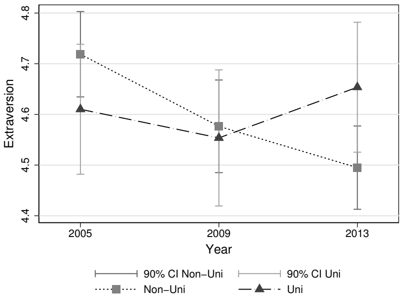
Figure 2.1 describes the trends in each of the five NCS over three measurement periods (2005, 2009, 2013) separately for the university entrants (triangle) and non-university entrants (square). Each symbol represents the mean level in personality trait j and capped lines illustrate their 90% confidence intervals. Individuals in the non-university group start out at a higher level of extraversion in 2005 than

[16] Deakin University, University of Tasmania, University of Wollongong, or Swinburne University of Technology, which are part of this network rank in the top 17 of Australia's university ranking according to the 2012 Excellence in Research for Australia (ERA) initiative.

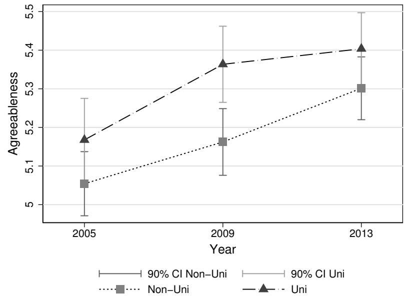
individuals in the university group, however a different trend emerges for the two groups over the next eight years (Figure 2.1(a)). Extraversion scores continuously decline for individuals in the non-university group, and the difference of 0.2 units between their 2005 and 2013 extraversion score is statistically significant. In contrast, the extraversion scores remain constant for individuals in the university group until 2009 and then they slightly increase - although not significantly - in 2013. These differential trends lead to a reversal in level rankings in extraversion over the eight-year window. The gaps in all other NCS between university and non-university entrants remain relatively constant over the eight-year time period (Figure 2.1(b)-2.1(e)), although university entrants score consistently higher on all four skills before they enter university.[17]

Figure 2.2 reveals important heterogeneities in levels and changes in NCS by socioeconomic status (below versus above the median value on FOPS) for agreeableness. Figure 2.2(b) shows that agreeableness scores increase for all adolescents between 2005 and 2013 independent of their SES and education choice. However, high SES adolescents who will enter university start out at the highest level of agreeableness (5.25), while their growth trajectory is the weakest. In contrast, low SES adolescents start out with the lowest levels of agreeableness in 2005 independent of whether they will enter university (5.10). Yet, low SES adolescents who will enter university experience the steepest growth in agreeableness by 0.5 units over the next eight years. In 2013, their agreeableness scores are 0.2 units higher than the scores of high SES adolescents who enter university.

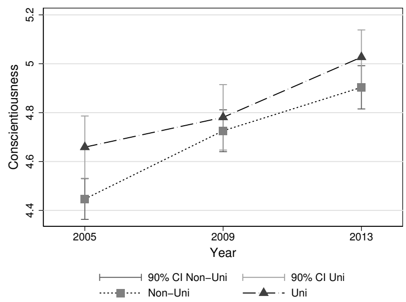
[17] We cannot say whether the changes between 2005 and 2009 represent the trends in the pre-university period, because most of the adolescents in the university group have entered university already by 2009. We have conducted an informal test on a sample of 32 individuals who will go to university but who have not done so yet by 2009. For this small sample, we can exploit the changes between 2005 and 2009 as pre-university trend and compare this pre-university trend against the trend of the non-university group. For all five personality traits, the changes over the four-year window are not statistically significant for any of the two groups - a result that is consistent with the findings in Cobb-Clark and Schurer (2012) - and they do not differ between the two groups. Although this is not a sufficient test to completely rule out differential growth trends between university and non-university group pre-university, it is tentative evidence against the differential pre-university trend hypothesis.



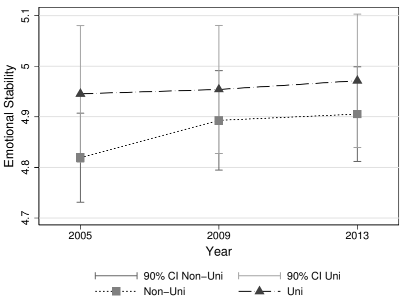
(a) Extraversion



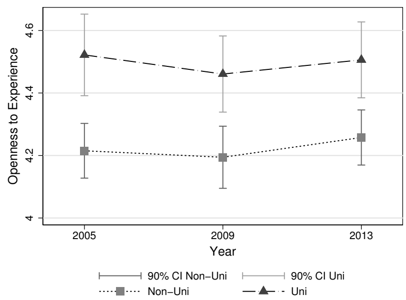
(b) Agreeableness



(c) Conscientiousness

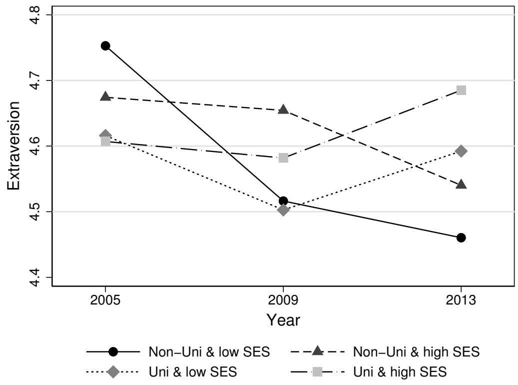


(d) Emotional stability

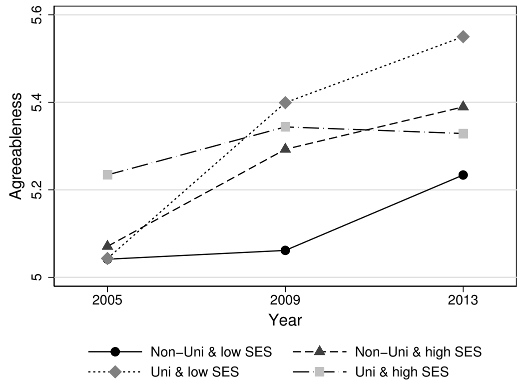


(e) Openness to experience

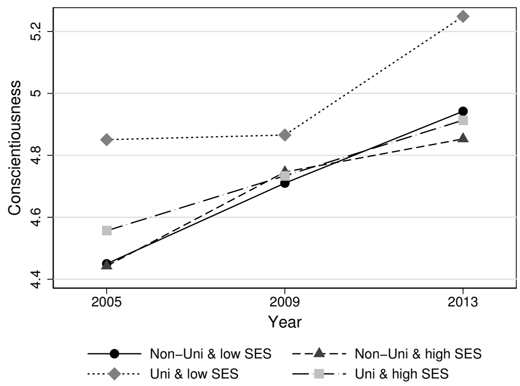
Figure 2.1: *Distribution of personality traits in 2005, 2009, and 2013 by university status*



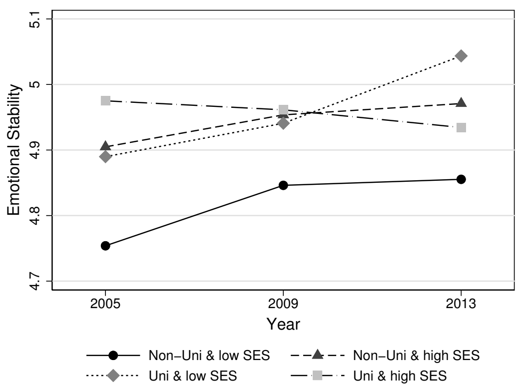
(a) Extraversion



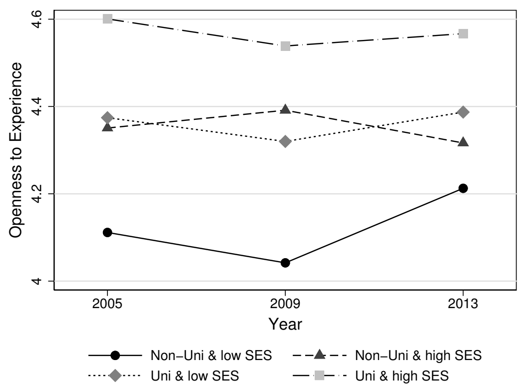
(b) Agreeableness



(c) Conscientiousness



(d) Emotional stability



(e) Openness to experience

Figure 2.2: *Distribution of personality traits in 2005, 2009, and 2013 by university and socioeconomic (SES) status*

Figure 2.2 further reveals the strong self-selection into university by conscientiousness for low SES adolescents who will enter university. At baseline, they score already highest in conscientiousness (4.85), exceeding the scores of all other adolescents by between 0.3 and 0.4 units. This is tentative evidence that conscientiousness may be a key determinant of upward mobility (Figure 2.2(c). Furthermore, low SES adolescents who will not enter university score lowest on emotional stability and openness to experiences at baseline (Figures 2.2(d) and 2.2(e)). Yet, growth trends in these NCS are similar for all groups, and thus gaps in NCS by 2013 are the same as in 2005.

In the next section, we test whether differences in the levels of the five personality traits between university and non-university groups in 2013 still remain when conditioning the analysis on starting levels of personality traits in 2005, before the adolescents potentially enter university, and all other relevant input factors into the adolescent skill production function.

2.5 Estimation results

2.5.1 Impact of university education on NCS development

To test whether university education shapes NCS skills, we first present the estimation results based on the value-added model described in Eq. 2.3. The estimated parameters of interest are reported in Table 2.2. Full estimation results are reported in Table A.3 in supplementary data. The estimation sample includes 575 adolescents who were aged between 15 and 19 in 2005. The dependent variable is personality trait j measured in 2013, standardised to mean 0 and standard deviation of 1. UNI measures whether the sample member completed a university degree or has entered the university track more than one year ago by 2013 (192 individuals). The reference group comprises individuals who went on to either some form of vocational training outside the university track or who did not complete any form of post-secondary training.

Panel A (Without interaction term) shows that university education is significantly associated with extraversion, whereby extraversion scores in the university group are almost one-third of a standard deviation (0.29 SD, significant at the 1% level) greater than the average score in the non-university group, ceteris paribus. Yet, university education is not associated significantly with any other NCS for the average adolescent. Almost identical estimates of the effect of university education on extraversion are obtained when using the CEM approach (0.26 SD, significant at the 10% level) based on 113 matched pairs for which all observable characteristics are well balanced between university and the non-university groups (Tables A.2 and A.4 in supplementary data).

Panel B (With interaction term) presents the estimation results of a model in which we interact UNI with a continuous measure of the father's occupational prestige score ($UNI \times FOPS$). As we standardise the paternal occupational class score to mean 0 and standard deviation 1, the coefficient on this interaction term is interpreted in terms of 1 SD increase away from the zero mean. A 2-SD up- and downward movement away from the mean implies that the father worked in a high-skilled (professional, managerial) and low-skilled occupation (manual, elementary), respectively.

Our conclusions do not change for extraversion, because the interaction effect is zero both in size and significance (0.02 SD). However, we find a statistically significant interaction effect for agreeableness in the magnitude of -0.23 SD (significant at the 1% level). Therefore, the marginal effect of university education for adolescents whose fathers rank 1 SD below the mean FOPS (lower levels of SES) score by 0.24 SD higher on agreeableness, while adolescents whose fathers rank 1 SD above the mean FOPS (higher levels of SES) score by 0.21 SD lower on agreeableness.

Table 2.2: *Estimated effects of university participation on Big-Five personality traits: Value-added and fixed effects models*

	Extrv	Agree	Consc	Emote	Openn
Value-added model (N = 575)					
Panel A: Without interaction term					
UNI	0.292***	−0.023	0.021	0.018	0.101
	(0.085)	(0.080)	(0.079)	(0.084)	(0.085)
R^2	0.40	0.26	0.33	0.27	0.31
Panel B: With interaction term					
UNI	0.289***	0.019	0.026	0.028	0.107
	(0.087)	(0.078)	(0.080)	(0.087)	(0.087)
UNI × FOPS	0.018	−0.225***	−0.025	−0.053	−0.033
	(0.078)	(0.085)	(0.080)	(0.083)	(0.082)
Marginal effect for adolescents from high and low SES					
High	0.306***	−0.206*	0.001	−0.024	0.074
Low	0.271**	0.244**	0.051	0.081	0.140
R^2	0.40	0.27	0.33	0.27	0.31
First-difference fixed effects model (N = 575, T = 2)					
Panel C: Without interaction term					
UNI=1	0.261***	−0.088	−0.113	−0.043	−0.105
	(0.093)	(0.095)	(0.096)	(0.110)	(0.105)
R^2	0.11	0.09	0.18	0.04	0.04
Panel D: With interaction term					
UNI	0.253***	−0.033	−0.118	0.009	−0.075
	(0.096)	(0.094)	(0.097)	(0.115)	(0.111)
UNI × FOPS	0.022	−0.153**	0.015	−0.145*	−0.082
	(0.077)	(0.072)	(0.074)	(0.077)	(0.078)
Marginal effect for adolescents from high and low SES					
High SES	0.274***	−0.186*	−0.103	−0.136	−0.157
Low SES	0.231*	0.120	−0.133	0.154	0.007
R^2	0.11	0.10	0.18	0.04	0.04

Note: Each model controls for the full set of control variables. Full estimation results are reported in Table refT:olschange-all. The value-added model includes lagged personality traits as additional control variables. Sample includes all respondents aged 15 to 19 in wave 5. FOPS stands for Father Occupational Prestige Score. A one-standard deviation increase in occupational prestige is 23.48 points on a scale from 0 to 100.

Source: HILDA, waves 5 and 13.

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Standard errors in parentheses.

Again, university education is not associated with conscientiousness, openness to experience, or emotional stability in the interaction model.

When controlling for individual-specific, time-invariant heterogeneity by exploiting changes in NCS between 2005 and 2013 and changes in all time-varying covariates, our conclusions remain unchanged (Panels C and D). The effect of university education on extraversion remains large and statistically significant in the first-difference model without (0.26 SD) and with interaction effects (0.25 SD), respectively. The heterogeneous effects of university education on agreeableness for students from low (0.12 SD) and high (-0.19 SD) SES backgrounds are similar to the ones obtained from the value-added model (Panel D), although the effect is halved for adolescents from low SES backgrounds.

As demonstrated in Figure 2.2(b), differences in agreeableness scores that emerge between university entrants from high and low SES stem from the observation that youth from disadvantaged backgrounds start from the lowest levels and experience the steepest growth. Adolescents from privileged backgrounds start at the highest levels of agreeableness but experience no growth during university education. One explanation, that is consistent with the observed data, is that for low SES adolescents university education changes the environment and peer groups more dramatically than for high SES students, and thus low SES students are more likely to adapt their behavioural styles governing interpersonal relationships in response to this change.

We are able to demonstrate in a series of robustness checks that our results for extraversion and agreeableness are not driven by the strong assumption underlying the linear and symmetric specification of the interaction effect between FOPS and university education. Using non-parametric estimation methods (Figure A.1 in supplementary data) or discrete categories for father's socioeconomic status (Table

A.5 in supplementary data), reveal the same patterns as described above. We further demonstrate no systematic differences in the impact of university education by sex (Table A.6 in supplementary data).

2.5.2 Potential mechanisms

So far, we have demonstrated a positive and robust link between university education and extraversion, and agreeableness for youth from disadvantaged backgrounds. In Section 2.2, we hypothesised that youth from disadvantaged backgrounds might be more affected by university education because they experience a bigger change in peer groups and in extracurricular activities available to them.

Another mechanism through which university education may shape NCS is through specific course contents or university-specific teaching programs or quality. To test for this mechanism, we re-estimated the value-added model by regressing personality trait j in 2013 on a set of dummy variables that each captures one of the five university groups or a set of dummy variables that represent the field of study if the individual attends or has completed university education. Table 2.3 reports the estimated coefficients for a model without interaction effects.

Overall, we find little systematic differences in the effect of university education by university grouping (Panel A). The effect of university education on extraversion – which stands out as the most important effect from our previous analysis – is equally strong across all universities. In magnitude, the effect is strongest for students who study at one of the OTHER universities (0.37 SD, significant at the 1% level), and weakest for students at one of the IRU universities (0.15 SD, not significant). We conducted an F-test of equality of differences in means across

Table 2.3: *Effect of university education by university type and field of study*

	Extrv	Agree	Consc	Emote	Openn
Panel A: Treatment effect by university type					
Go8	0.281**	−0.154	0.146	−0.157	0.281**
	(0.132)	(0.125)	(0.130)	(0.132)	(0.116)
ATN	0.282	−0.124	−0.067	0.080	0.078
	(0.193)	(0.157)	(0.190)	(0.180)	(0.190)
IRU	0.147	0.001	−0.095	0.092	−0.105
	(0.153)	(0.158)	(0.152)	(0.148)	(0.149)
RUN	0.358	0.040	0.123	−0.081	0.021
	(0.239)	(0.221)	(0.294)	(0.303)	(0.277)
Other	0.368***	0.050	0.022	0.104	0.009
	(0.124)	(0.130)	(0.112)	(0.127)	(0.148)
R^2	0.39	0.25	0.32	0.27	0.31
Observations	559	559	559	559	559
Equality of coef. (p-value)	.802	.682	.706	.470	.195
Panel B: Treatment effect by degree discipline					
STEM	0.422***	−0.220*	−0.042	−0.055	0.167
	(0.146)	(0.129)	(0.149)	(0.151)	(0.154)
Medicine and health related	0.287*	0.148	0.230	−0.155	−0.161
	(0.147)	(0.151)	(0.147)	(0.135)	(0.138)
Education	0.273	0.228	−0.020	0.303	0.043
	(0.186)	(0.151)	(0.178)	(0.199)	(0.183)
Management, Commerce, Law	0.158	−0.018	0.088	0.027	0.172
	(0.132)	(0.121)	(0.124)	(0.138)	(0.144)
Others	0.224	−0.199	−0.132	0.044	0.152
	(0.137)	(0.140)	(0.128)	(0.127)	(0.148)
R^2	0.40	0.26	0.33	0.27	0.30
Observations	571	571	571	571	571
Equality of coef. (p-value)	.657	.074	.330	.324	.315

Note: Respondents aged 15 to 19 in wave 5. In total, 176 individuals who completed or are completing their university degree provided information on their university of study and 188 provided information on their field of study. **Group of 8 (Go8)**: The University of Adelaide, The Australian National University, The University of Melbourne, Monash University, The University of New South Wales, The University of Queensland, The University of Sydney and The University of Western Australia; **The Australian Technology Network (ATN)**: Curtin University, University of South Australia, RMIT University, University of Technology Sydney and Queensland University of Technology; **Innovative Research Universities (IRU)**: Flinders University, Griffith University, La Trobe University, Murdoch University, University of Newcastle, James Cook University and Charles Darwin University; **The Regional Universities Network (RUN)**: Central Queensland University, Southern Cross University, University of Ballarat, University of New England, University of Southern Queensland and University of the Sunshine Coast; **Other**: Australian Catholic University, Australian Defence Force Academy, Bond University, Charles Sturt University, Deakin University, Edith Cowan University, Macquarie University, Swinburne University of Technology, University of Canberra, University of Notre Dame Australia, University of Tasmania, University of Western Sydney, University of Wollongong, Victoria University (Victoria University of Technology), Other (please specify). Each model controls for the full set of control variables including lagged personality measures.

Source: HILDA, wave 5 and 13.

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Standard errors in parentheses.

all university groups, and fail to reject the null hypothesis.[18] We also find little

[18]One exception to previous results is that the effect of university education on openness to experience is positive and significant for Go8 university students (0.28 SD, significant at the 5% level) and zero for all other university types.

evidence that the effect of university education on NCS differs by field of study (Panel B).[19]

We conduct another test to confirm that the effects we are measuring are attributable to university education and not to other unobserved factors. If university education truly builds extraversion for all students and agreeableness for students from disadvantaged backgrounds, then we should find that the more time a student has spent at university, the greater this impact should be. We re-estimated the value-added model including years of exposure to university education, a variable which varies between 0 and 8 years, as a dose indicator for university exposure. We find that every additional year spent at university is associated with a 0.07 SD (significant at the 1% level) increase in extraversion and a 0.10 SD (significant at the 10% level) increase in agreeableness for youth from low SES backgrounds.[20] Hence, the length of exposure to university life contributes to NCS formation.

We conclude that a likely channel through which university education contributes to changes in agreeableness for youth from disadvantaged backgrounds is exposure to new peer groups and/or extracurricular activities. Furthermore, our findings are consistent with the hypothesis that NCS are shaped through exposure to new learning environments that place higher demands on students, independent of where students study. There is no evidence in support of mechanisms related to degree-specific content, teaching styles, or quality.

2.6 Discussion and conclusion

Recently, a public debate has emerged on whether universities teach the right skill-sets that prepare students for a continuously changing and globally expanding

[19] The effect sizes vary slightly for extraversion across field of studies, but we fail to reject the null hypothesis that the means are the same across all groups.
[20] Full estimation results are provided upon request.

labour market. Employer surveys have shown that non-cognitive skills are valued highly by employers, but leading scholars have raised concerns that university education may fall short of teaching students such skills. Yet, no empirical evidence has existed on the matter.

We contribute to this discussion by providing a first empirical glance at the role that university education plays in the skill production function of adolescents. Following the education decisions of a sample of Australian youths from 2005 until 2013 and controlling for the self-selection by NCS into university education, and other important factors, as best as we can, we find robust evidence that Australian universities contribute to building sociability (extraversion) and the tendency to cooperate (agreeableness).

Youth who enter the university track or complete tertiary education have significantly higher levels of extraversion. Importantly, this effect does not differ by sex, university type or field of study. Despite the strong self-selection into university type and field of study, and the heterogeneity in teaching quality and curricula across universities, we suggest that the likely mechanisms through which university education shapes outward orientation is through exposure to university life and less so through quality and what is being taught in class. University education may foster these tendencies because it encourages participation in club activities, social functions, and communication with fellow students and academic staff on a continuous basis. This conclusion is strengthened by the finding that years spent at university is positively associated with extraversion. We propose therefore that university education may be associated with higher earnings because university education shapes sociability skills which also have high labour-market returns (Gensowski, 2018; Fletcher, 2013; Heineck and Anger, 2010).

In addition, university education is associated with higher levels of agreeableness for both male and female students from low socioeconomic backgrounds, who started from the lowest baseline scores in adolescence and experienced the steepest growth curve as they entered university. This implies that students from disadvantaged backgrounds catch up with their peers from more privileged backgrounds, thus reducing initial levels of inequality in agreeableness. This is likely due to exposure to new peer groups and/or extracurricular activities. This convergence in agreeableness may have important welfare effects, because agreeableness has been linked to reciprocity and altruism (Becker et al., 2012) and prosociality (Hilbig, Gloeckner, and Zettler, 2014), economic preference which are considered to be at the basis of socioeconomic development (e.g. Bigoni et al., 2016) and population wellbeing (Post, 2005). Therefore, the benefits of university education is not only that it increases individuals' employability and earnings, but that it may directly contribute to the formation of socially-beneficial preferences (See Arrow, 1997, for similar arguments).

We cannot say with certainty whether these effects are permanent or temporary, because our data does not provide long-term follow up personality data. Our findings support the conclusion that Australian universities are at least in the short run successful in shaping life skills which employers and society value. The skill-returns of university education and its psychic benefits are substantial for youth from disadvantaged backgrounds. The public discourse is therefore misguided on claiming that university education does not contribute to human capital formation.

Second, the current policy focus on early childhood education to boost non-cognitive skills (See Kautz et al., 2015, for a review) may be too narrow. Our findings suggest that non-cognitive skills can still be shaped at later stages as suggested in Schurer (2017) who summarises such evidence for the post-primary school sector. We conclude that interventions in the secondary or tertiary education sector may be a promising avenue to boost non-cognitive skills. Future research that identifies the exact channels through which university attendance impacts upon skill formation, for instance the role of peer effects, teaching innovations or curricula reforms, could provide useful information for universities that seek to strategically target their students' NCS development.

Chapter 3

Measuring non-cognitive skills: Beyond the factor analytic framework

3.1 Introduction

The Big Five personality traits, and non-cognitive skills more generally, have been extensively studied in labour, education and health economics. Non-cognitive skills both predict and cause later life outcomes (Heckman, Jagelka, and Kautz, 2021; Almlund et al., 2011b). Personality inventories for the Big Five can now be found in many longitudinal household surveys.[1] Despite the success of the Big Five personality traits, which can be considered a variable-centric approach to personality, some researchers have approached personality from a person-centric perspective, in which individuals are grouped into different personality types. In terms of replications studies, though not as extensive compared to those

[1]Including the German Social-Economic Panel; the Household, Income and Labour Dynamics in Australia Survey; the US Panel Study of Income Dynamics; and the UK household longitudinal study, Understanding Society.

for the Big Five, personality types have been replicated. More recently, using some of the largest personality data, Gerlach et al. (2018) found evidence for four personality types. This is a puzzling finding given the abundance of research that has replicated the Big Five. More importantly, the underlying statistical models – a linear factor analysis model for the Big Five personality traits and a Gaussian mixture model for personality types – make incompatible assumptions about the number of clusters in the data. In a linear factor model, the observed variables are assumed to be jointly normally distributed and therefore form just one cluster. In a Gaussian mixture model, however, it is implicitly assumed that there are at least two clusters within the data.

To begin to reconcile findings of personality types with studies that have replicated the Big Five, we revisit some of the early but foundational studies on the Big Five. Specifically, we show that many key statistics (e.g., eigen values in a scree plot, Cronbach's alpha, the Kaiser-Meyer-Olkin index, and the Tucker-Lewis index) that are routinely used to guide factor extraction, show internal consistency, and demonstrate factor replicability do not imply the absence of clusters in the data. Using two personality data sets, we compare the relative model fit of a multitude of factor analysis models and Gaussian mixture models and find that a mixture model provides a relatively better fit. Importantly, we show that even the best mixture model fails to capture all saliency of the personality data analysed. This suggests that alternative latent variable models may be needed.

In their seminal review on personality psychology and economics, Almlund et al. (2011a) developed a theoretical framework that integrates personality traits into economic models and suggested that personality psychology has the potential to deepen our understanding of conventional economic preferences such as risk aversion and time discounting. In order to pursue research that seeks to integrate personality traits and economic preferences, it is therefore important to reconcile the different approaches to measuring personality traits.

This chapter is organised as follows. Section 3.2 reviews the literature on personality traits and personality types, studies that compare the two approaches, and other approaches that are not based on a latent variable model. In section 3.3, we explain why studies that have replicated the Big Five do not necessarily imply the absence of personality types. We describe in section 3.4 three empirical checks and illustrate in section 3.5 how they perform on simulated data. We apply them to two datasets, described in section 3.6, and discuss the results in section 3.7. Section 3.8 concludes.

3.2 Literature review

In this section, we briefly review the history of the research on personality traits and personality types and highlight some recent developments. Personality traits and types reflect two different approaches to the study of personality: variable-centric and person-centric, respectively. There are replication studies for both approaches but few had directly compared the two. Other approaches that do not make use of latent variable models are also discussed.

3.2.1 Personality traits

The Big Five personality taxonomy is the result of decades of research, with its roots in the lexical hypothesis of Francis Galton (Goldberg, 1993a). The application of factor analysis to psychometric data has generated multiple taxonomies over the years before a consensus began to emerge. From the early lexical studies that began in the 1930s (Allport and Odbert, 1936) to the emerging consensus on the Big Five in the early 1990s, factor analysis has been fundamental as a statistical method in the study of personality. Much research effort had been devoted to determining the number of factors. Before the 1990s, Cattell's 16-personality factor model and Eysenck's 3-factor model were common in applied personality research; studies that used the Big Five are much fewer in number. However, by the early 1990s, a consensus began to emerge, with many studies successfully replicating the Big Five. The five traits are openness to experience, conscientiousness, extraversion, agreeableness and neuroticism (the opposite is also known as emotional stability). Since then, the Big Five have continued to gain popularity, surpassing previous models of personality traits (Cattell's and Eysenck's) in terms of the number of publications.

The Big Five personality traits have been replicated in both self-rated and peer-rated personality data (McCrae and Costa, 1987) and across different cultures and languages (McCrae and Terracciano, 2005; Saucier and Goldberg, 2001; Hofstee et al., 1997; McCrae and Costa Jr, 1997). There has been debate over the numbder of factors. Besides Cattell's and Eysenck's, other proposed taxonomies include, for example, the six-factor HEXACO structure[2] (Ashton and Lee, 2007; Ashton et al., 2004) and the seven-factor Multi-Language Seven (ML7) that is based on multiple languages and cultures (Saucier, 2003). A higher-order factor called the General Factor of Personality has also been postulated (Linden, Dunkel, and Petrides, 2016; Musek, 2007).

These various taxonomies, though differing in the number of factors, are nonetheless based on the use of factor analysis as applied to psychometric data. They are therefore subject to the same criticism of the Big Five personality traits that they are "atheoretical" and therefore to be regarded as descriptive only (Block, 2010, 1995).

[2]The HEXACO model is closely related to the Big Five but adds the sixth dimension of Honesty-Humlity.

3.2.2 Personality types

Despite the dominance of the variable-centric approach represented by the Big Five, others have emphasised the importance of a person-oriented perspective (Asendorpf, 2002). Outside of the field of personality psychology, the research on personality types appears to be relatively lesser known. For example, in the highly influential handbook chapter by Almlund et al. (2011b) on personality psychology and economics, there is no mention of personality types. In the more recently published Handbook of Personality: Theory and Research, its coverage is also limited (John, 2021). However, not unlike the course of research on personality traits, personality types also originated from a hypothesis and have been replicated, extended, or refined over the years.

The notion of personality types dates back to Block and Block (1980), whose hypothesis introduced the concepts of ego-control and ego-resiliency. However, early attempts to replicate it mostly failed. Two types of methods have been used: Q-sort, also known as inverse factor analysis, and cluster analysis, with the latter dominating later work. Robins et al. (1996) is among the first to empirically replicate three personality types: resilients, undercontrollers, and overcontrollers. Two other influential studies, by Asendorpf et al. (2001) and Caspi and Silva (1995), have also replicated the three types, which some have called the RUO types or the ARC types, named after the three respective lead authors.

While the ARC types have been replicated, the number and characteristics of personality types have continued to be debated. For example, Herzberg and Roth (2006) found evidence for five clusters based on a sample of 1,908 German adults. This person-centered approach has other drawbacks, as De Fruyt and Karevold (2021) noted. In terms of their predictive validity, personality types have not fared as well compared to the Big Five traits. Within the same personality type, there could also be substantial differences between individuals.

Recent studies have identified more than three personality types using large data sets. For example, Kerber, Roth, and Herzberg (2021) found a five-cluster solution in the German Socio-Economic Panel (SOEP) data. Gerlach et al. (2018) found four clusters with high density of individuals around each cluster in four large data sets. These include a data set of responses from over 145,000 individuals on a 300-item International Personality Item Pool (referred as the Johnson-300), and another from over 410,000 respondents on a shorter version with 120 items.

However, as Freudenstein et al. (2019) pointed out, the four personality types identified only account for about 42 percent of the Johnson-300 data set. In other words, most individuals could not be assigned to any distinct cluster. Katahira et al. (2020) also raised the possibility that the findings in Gerlach et al. (2018) could be affected by skewness in the distribution of responses. They did not evaluate to what extent skewness might play a role and concluded that it remains an "open question whether the distribution of personality should be characterized as categorical, dimensional, or their intermediate".

3.2.3 Studies comparing traits and types

Studies that directly compare the dimensional and typological approaches are relatively few. The main method of comparison in these studies is based on predictive validity. That is, how well personality traits and personality types can predict other outcomes. Costa Jr et al. (2002) compared the predictive power of traits and types on different measures psychosocial functioning and various personality disorder and found that personality traits were consistently more predictive than type memberships. Asendorpf (2003) reported similar findings in terms of cross-sectional predictions but, in a small sample of 99 participants, found that personality types at ages 4-6 were equally predictive of personality and cognitive ability at age 17 compared to personality traits. Similar results were found in a follow-up study (Asendorpf and Denissen, 2006). Roth and Collani

(2007) is another study in support of the predictive utility of personality types. They found that a five-cluster solution on Big Five data was more optimal than a three-cluster solution and could predict social attitudes as well as the Big Five traits.

3.2.4 Other approaches

The statistical models that are commonly used to derive personality traits and personality types – factor analysis and mixture models, respectively – are examples of latent variable models. There are other approaches to studying personality that do not make use of latent variable models.

The use of network analysis applied to psychometric data has gained popularity in recent years (Epskamp, Borsboom, and Fried, 2018; Golino and Epskamp, 2017; Cramer et al., 2012). A network or graph consists of nodes and edges. Nodes represent observed variables and edges represent statistical relationships, such as correlations. The use of a network allows researchers to directly study the interrelationships between the observed variables without assuming any hidden variables.

Another example that directly uses the observed variables is Mareckova and Pohlmeier (2017). Mareckova and Pohlmeier (2017) used machine learning techniques to evaluate how well non-cognitive skills could predict labour market outcomes. Typically, responses to personality inventories are either averaged to get the mean scores or factor analysed to obtain the factor scores. These scores are then used as explanatory variables together with other covariates to predict an outcome of interest. Instead, Mareckova and Pohlmeier (2017) used a group LASSO (least absolute shrinkage and selection operator) directly on the responses to non-cognitive skills questions to predict labour market earnings and employment status.

3.3 Limitations of statistics routinely used to support factor analysis

Many key statistics that are routinely used to guide factor extraction, show internal consistency, and demonstrate factor replicability depend only on the covariances of the observed data. In other words, given any two sets of data that have the same covariance matrices, these statistics would also have the same values. The implication is that in general these statistics alone cannot be used to argue in favour of or against any data generating process, as will be illustrated below. This section begins with examples from influential studies of the use of these statistics and ends with an illustration of how these statistics could be almost identical for two data sets with very different data generating processes.

3.3.1 Examples from influential studies

Among the most important findings regarding the five-factor model are:

1) the number of factors sufficient to account for the dimensions of personality traits is five,
2) the five factors show high internal consistency, and
3) the five factors can be replicated from different ratings of the same individuals, and for people from different cultures.

Consider some of the statistics commonly used to support the above findings.

To help determine how many factors to retain, commonly used techniques include the scree plot and the Kaiser rule, i.e., the eigenvalues-above-one rule. Both are heuristics that are based on the eigenvalues of the covariance matrix. As an example, when determining the appropriate factor structure, McCrae and Terracciano (2005) wrote, “[t]he first six eigenvalues were 6.67, 4.40, 3.51, 2.43, 1.46, and 0.84, unmistakably suggesting a five-factor solution”. In Digman and Inouye

(1986), as another example, the use of the Kaiser rule was also explicit; they wrote, "[a]s the five-factor model would predict, there were five eigenvalues in excess of unity", and also referenced the scree test.

To show factor reliability or internal consistency, Cronbach's alpha[3]is often used. For example, McCrae (1982) and McCrae and Costa (1987) referred to the relatively high alpha coefficients in support of the five-factor structure.

To demonstrate factor replicability – that two factor loadings matrices from two different studies are similar, researchers would either directly compare the factor loadings matrices from different studies or calculate the Kaiser coefficient[4] or the Tucker's factor congruence coefficient[5]. Some examples of the former include Fiske (1949), Tupes and Christal (1992), and Borgatta (1964). Examples of the latter include Norman (1963), Digman and Takemoto-Chock (1981), and McCrae and Terracciano (2005).

The statistics mentioned above can all be calculated from the covariance (or correlation) matrix of the data alone. This is not surprising. Some early replication studies (e.g. Digman and Takemoto-Chock, 1981), to re-analyse previous data sets, would do so using only the correlation matrices published in those previous

[3]Cronbach's alpha is defined as:

$$\alpha = \frac{N \cdot \bar{c}}{\bar{v} + (N-1) \cdot \bar{c}},$$

where N is the number of items, $\bar{c}$ is the average covariance, and $\bar{v}$ is the average variance of each item.

[4]The Kaiser coefficient is given by:

$$KMO = \frac{\sum\sum_{j \neq k} r_{jk}^2}{\sum\sum_{j \neq k} r_{jk}^2 + \sum\sum_{j \neq k} p_{jk}^2},$$

where r_{jk} and p_{jk} are the correlation and partial correlation coefficients respectively.

[5]The Tucker's factor congruence coefficient is defined as:

$$\phi = \frac{\sum P_i Q_i}{\sqrt{\sum P_i^2 \sum Q_i^2}},$$

where P_i and Q_i are the two factor loadings being compared.

studies. Also, certain software for factor analysis, e.g., the factor analysis module in Stata or the psych package in R, could take the correlation matrix as the only input. The most important implication is that generally these statistics cannot inform researchers if a factor analytic model is more appropriate for a given data set. The following example will illustrate this.

3.3.2 Illustration

Consider the following two simulated data sets, each with 2000 observations and 5 observed variables. One of them is generated from a one-factor model, and the other from a two-cluster Gaussian mixture model. The parameters of the two models are chosen such that they have identical population covariance matrices.[6] Then a sample of 2000 is drawn from each DGP to give the two data sets.

As expected, the two sample covariance matrices, as shown in Table 3.1, are highly similar. As the covariance matrices are highly similar, so too are their eigenvalues, and in turn the scree plots, as shown in Figure 3.1. For both data sets, the Kaiser rule and the alternative of Horn's parallel analysis (Horn, 1965) would suggest a one-factor model.

Estimating a one-factor model on both datasets, and inspecting the resulting factor loadings matrices, as reported in Table 3.2, show that they are also similar. Table 3.3 reports various statistics for the one-factor solution on both data sets. Judging by the Cronbach's alpha coefficients, the one-factor solution for both data sets has good internal consistency. The Kaiser-Meyer-Olkin (KMO) index that measures factor simplicity is also similar. The Tucker-Lewis index, a reliability coefficient, is high for both data sets. When comparing the similarity between the two factor

[6] Bartholomew, Knott, and Moustaki (2011) showed that a linear factor model having k factors and a mixture model with $k+1$ clusters can be constructed such that they have identical means and covariances.

Table 3.1: *Sample covariance matrices for the two simulated data sets from a one-factor model (left) and a two-cluster mixture model (right)*

	X1	X2	X3	X4	X5		X1	X2	X3	X4	X5
X1	1					X1	1				
X2	-0.65	1				X2	-0.66	1			
X3	-0.64	0.6	1			X3	-0.62	0.62	1		
X4	-0.15	0.14	0.11	1		X4	-0.15	0.16	0.17	1	
X5	0.41	-0.42	-0.39	-0.06	1	X5	0.43	-0.41	-0.4	-0.09	1

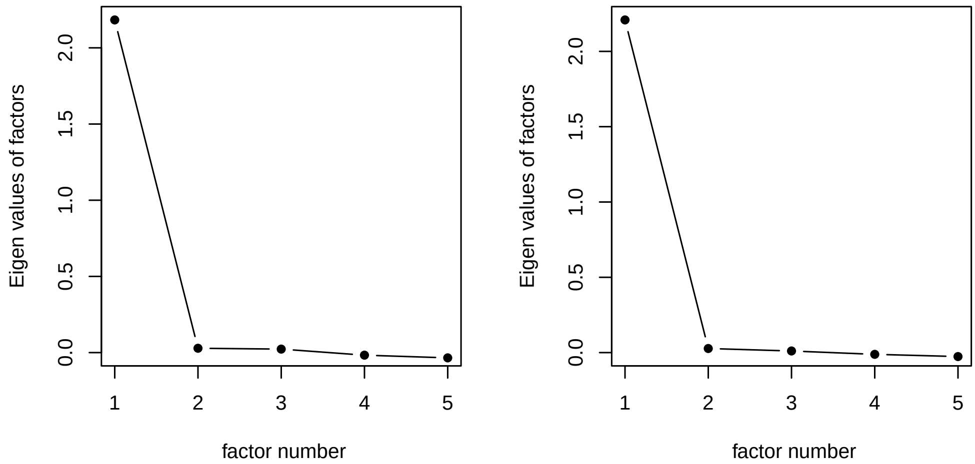


Figure 3.1: *Scree plots for the two simulated data sets from a one-factor model (left) and a two-cluster mixture model (right)*

loadings matrices, the Tucker's factor congruence coefficient (not reported in Table 3.3) is above 0.99.

In summary, just by comparing the scree plots and the various statistics reported, it is hard to tell which data set is generated from a factor analytic model. It may be hypothesised that some factor analytic model is more likely to be the DGP for a given data set, but these statistics alone can neither confirm nor disprove the hypothesis that the true DGP is a factor analytic model. In other words, if the task is to empirically show that a particular DGP is more suitable than others, then resorting to these statistics will not give any conclusive evidence.

Table 3.2: *Factor loadings matrices of a one-factor solution for the two simulated data sets from a one-factor model (left) and a two-cluster mixture model (right)*

	Factor 1		Factor 1
X1	-0.827	X1	-0.816
X2	0.789	X2	0.808
X3	0.769	X3	0.761
X4	0.165	X4	0.196
X5	-0.507	X5	-0.522

Table 3.3: *Measures of internal consistency, sampling adequacy, and factor reliability of a one-factor solution for the two simulated data sets*

	one-factor model	two-cluster mixture model
Cronbach's alpha	0.760	0.769
Kaiser-Meyer-Olkin index	0.792	0.798
Tucker-Lewis index	0.997	1.000

3.4 Three empirical checks

The most important implication of the last illustration is not that some mixture model could have been the data generating process for personality traits, but that any DGP that could produce data with a similar covariance matrix would give similar values for those statistics. After all, the DGP does not have to be a factor or mixture model. The illustration also demonstrates that even if those statistics appear to favour a particular factor analytic model, it is incorrect to infer that the DGP is or close to being one. In what follows, we will look at empirical checks that could differentiate different DGPs.

The central question is, how do we know if factor analysis is appropriate for a given data set? Putting it differently, given any two data sets of which only one is known to be generated from latent factors, how could we empirically tell them apart? The primary concern is therefore not how many latent factors there are or if they should be orthogonal to each other, but whether factor analysis, as an entire class of latent variable models, is the right tool in the first place.

To help answer this question, suggested are three empirical tests, motivated by the same idea: derive characteristics that would be true given a factor analytic data generating process (DGP), then test for those characteristics in a given data set. We defer the derivations to the next section, and state the three characteristics first:

1. A factor analytic DGP with linear Gaussian factors can only generate jointly normally distributed variables.
2. The true DGP would fit the data better than any other DGP.
3. If we were to generate another data set from the true DGP, it would have the same statistical properties as the given one.

Each of these can be checked empirically, and together we have the three tests.

It follows that, to determine if a data set has a factor analytic DGP, we could:

1. Use a multivariate normality test - to check that the observed variables are jointly normally distributed.
2. Compare the model fit - to check that the best model obtained, among those considered, is factor analytic.
3. Conduct a two-sample test - to check that the samples generated by the best factor analysis model are statistically no different from the given sample.

It is true that these checks could not help us verify if a particular factorial structure or some other model is the true DGP, but they could tell us whether the data set is at least consistent with a factor analytic DGP. If we found that the variables in some given data were not jointly normally distributed, that the best model was not factor analytic, and that the best factor model could not generate statistically similar data, then we would conclude that perhaps alternative models other than factor analysis are needed.

3.4.1 Are the observed variables normally distributed?

In any (linear Gaussian) factor analysis model, the manifest variables are jointly normally distributed. This suggests that, in determining if factor analysis is suitable, one could, among other considerations, assess the normality of the variables in a given data set. For example, if the data are far from normally distributed, one may not want to use factor analysis and instead consider other alternatives. In what follows, we first describe the factor analysis model and show that the manifest variables follow a joint normal distribution, and then outline how we plan to assess normality in a given data set.

Factor analysis model In a factor analysis model, the p-dimensional observed variables (x) are modelled as linear combinations of K latent factors ($\mathbf{F}$) with additive Gaussian noise (ϵ)

$$\mathrm{x} = \mu + \Lambda \mathbf{F} + \epsilon \,,$$

where $\mu = \mathbb{E}[\mathrm{x}]$, Λ is the ($p \times K$) matrix of factor loadings, and $\epsilon \sim \mathcal{N}(0, \Psi)$ with Ψ being a diagonal covariance matrix. The latent factors are, without loss of generality, assumed to have a standard normal distribution, i.e. $\mathbf{F} \sim \mathcal{N}(0, \mathbf{I})$. It can then be shown that the observed variables are jointly normal, with the marginal distribution given by

$$\mathrm{x} \sim \mathcal{N}(\mu, \Lambda\Lambda^\intercal + \Psi) \,.$$

That the observed variables are jointly normally distributed means that there is just one cluster, with μ as the center. It is worth noting that the implied normality holds for any (linear Gaussian) factor analysis models. That is, the observed variables are jointly normally distributed regardless of the number of latent factors, whether the factor loadings matrix is completely free or restricted in some ways (i.e. exploratory or confirmatory factor analysis respectively), whether the latent factors are orthogonal to each other or not, and whether frequentist or Bayesian factor analysis is used. In other words, the more the variables in a data set deviate from a normal distribution, the less well they can be approximated by any factor analysis model. From a generative perspective, no (linear Gaussian) factor analysis model could generate data that are not normally distributed. Therefore, checking normality of the observed variables could be a useful first step.

Doornik-Hansen multivariate normality test To check for normality, one could conduct a multivariate normality test. Of the wide variety of such tests, we suggest the Doornik-Hansen test, due to its desirable size and power properties as compared to other popular multivariate normality tests (Doornik and Hansen, 2008). In the Doornik-Hansen test, the data are first decorrelated with a Mahalanobis transformation step. Then the univariate skewness and kurtosis statistics for each column of the transformed data are computed and further transformed into normally distributed variables (denoted respectively as $z_{11}, z_{12}, \ldots, z_{1p}$ and $z_{21}, z_{22}, \ldots, z_{2p}$). The Doornik-Hansen test statistic is defined as

$$E_p = Z_1^\intercal Z_1 + Z_2^\intercal Z_2 \,,$$

where $Z_1^\intercal = (z_{11}, z_{12}, \ldots, z_{1p})$ and $Z_2^\intercal = (z_{21}, z_{22}, \ldots, z_{2p})$. The statistic is approximately χ^2 distributed with $2p$ degrees of freedom, i.e. $E_p \overset{\text{app}}{\sim} \chi^2_{2p}$.

To assess normality, it may not be enough to rely on data visualisation techniques. It would appear that a scatterplot matrix could be used for the task of checking normality, but it quickly becomes less practical when there are many observed variables, as the number of bivariate distributions increases quadratically (or $\frac{1}{2}(p^2 - p)$ to be exact) with the number of variables. Even if the number of variables is small, it can still be difficult to discern any departure from normality.

3.4.2 Does the proposed model has the best model fit?

In a factor analysis model, the latent variables are assumed to follow a standard normal distribution. This is but one assumption that can be made regarding the latent variables, and different assumptions give rise to different latent variable models. Given the many latent variable models, how would one show that a particular latent variable model is better or worse than any other? One way is to evaluate the relative fit of the model. It seems natural to expect the preferred latent variable model to have a better relative model fit. In the context of modelling non-cognitive skills, the key question is then, can it be shown that some particular factor analysis model, being the current model of choice, has a better model fit, not just within the class of factor analysis models but compared to other classes of latent variable models? In this section, we first present more formally the challenge of selecting a latent variable model, and then discuss the use of model selection criteria.

Model selection Denote the observed variables in a given data set by x, and the set of latent variables by η. The marginal distribution of the observed variables can then be expressed as

$$p(\mathrm{x}) = \int p(\mathrm{x}|\eta)p(\eta)\,\mathrm{d}\eta\,, \tag{3.1}$$

if η is continuous.[7] Since neither $p(\mathrm{x}|\eta)$ nor $p(\eta)$ is observed in a data set, the number of latent variables, how the latent variables relate to one another, and how they relate to the manifest variables, are all unknown. In fact, there are numerous combinations of $p(\mathrm{x}|\eta)$ and $p(\eta)$ that could give rise to the same $p(\mathrm{x})$ (Bartholomew, 2007, p. 18).

[7] If η is discrete, the marginal distribution is obtained by summing over η, i.e. $p(\mathrm{x}) = \sum p(\mathrm{x}|\eta)p(\eta)$.

Essentially, different latent variable models place different assumptions on $p(\mathrm{x}|\eta)$ and $p(\eta)$. In a factor analysis model, the assumption is that the observed variables are linear combinations of normally distributed latent variables, such that $p(\mathrm{x}|\eta) = \mathcal{N}(\mathrm{x}|\mu + \Lambda\eta, \Psi)$ and $p(\eta) = \mathcal{N}(\eta|0, \mathbf{I})$. Hence, from the general equation (3.1), the marginal distribution of the observed variables is reduced to the specific equation

$$\begin{aligned} p(\mathrm{x}) &= \int \mathcal{N}(\mathrm{x}|\mu + \Lambda\eta, \Psi)\, \mathcal{N}(z|0, \mathbf{I})\, \mathrm{d}\eta \\ &= \mathcal{N}(\mathrm{x}|\mu, \Lambda\Lambda^{\intercal} + \Psi)\,. \end{aligned} \tag{3.2}$$

How do we justify this reduction from equation (3.1) to equation (3.2)? For example, we could have made different assumptions such that the latent factors are not restricted to be Gaussian, leading to e.g. independent component analysis, or that $p(\eta)$ be a mixture, leading to a mixture model. Another option for even more flexibility is to allow different factorial structures in different subgroups, as in mixtures of factor analyzers (Ghahramani and Hinton, 1996). Yet more flexible latent variable models are available. Given so many choices, how do we justify any preference towards any particular class of latent variable model? It would seem the use of a particular model would be justifiable if it is shown to have a better model fit as compared to its alternatives. In the context of measuring non-cognitive skills then, in which factor analysis is almost exclusively used, the question is therefore, can we show that some factor analysis model provides a better fit relative to other classes of latent variable models?

Model selection criterion To compare the relative goodness of fit of the different models, we use the Bayesian Information Criterion (BIC):

$$\mathrm{BIC} = -2L + d\ln N,$$

where L is the log likelihood of the model and d is the number of parameters.

For a factor analysis model with k factors, the log likelihood is given by:

$$L = -\frac{Np}{2} ln(2\pi) - \frac{N}{2} ln|C| - \frac{N}{2} tr(C^{-1}S),$$

where $C = \Lambda\Lambda^{\intercal} + \Psi$ is the covariance matrix under the factor analysis model and S is the sample covariance matrix. The number of parameters is $p(k+1) - (k(k-1)/2)]$.

For a Gaussian mixture model, the log likelihood is given by:

$$L = \sum_{n=1}^{N} \ln\Big\{\sum_{i=1}^{k} \mathcal{N}(\mathrm{x}_n|\mu_i, \Sigma_i)\Big\}.$$

The number of parameters depends not just on the number of components but also on the type of covariance matrices of the components. That is, whether they are restricted in any way, as they typically are to reduce the number of parameters that needs to be estimated. For the unrestricted case in which each component has a different covariance matrix, the number of parameters is $kp(p+1)/2$.

3.4.3 Does the selected model describe the data?

Regardless of whether AIC or BIC is used, what we get from any relative model fit criteria is essentially a relative rank of the candidate models. The more important question is perhaps whether the best model among the candidates is good enough. To answer that question, what is needed is some criterion that could be used to evaluate a given model on its own. This is particularly important as there are potentially far more models to choose from than the ones already in the candidate set.

Assume that the true DGP is available for the sample at hand, say with a sample size of 1,000. If we were to generate another 1,000 new observations from the same

DGP, we would expect the two samples to be statistically indistinguishable. In reality, we approximate the true DGP with some chosen model, but the quality of the chosen model can be assessed based on how similar the new observations are to the original sample. More specifically, any model estimated from the data is falsifiable to the extent that new observations generated from it deviate from the original data. Any multivariate two-sample tests can be used for this purpose; here we will use the generalised edge-count test (Chen and Friedman, 2017).

The generalised edge-count test is a graph-based two-sample test for multivariate data. Given two samples, we first construct a three-nearest neighbour graph on all observed values (with each observed value represented as a node) such that every node is connected by an edge with its three closest nodes. If two samples of data have different distributions, data points from the same sample are more likely to be connected with themselves. The edge-count test works by comparing the number of within-sample edges to the expected number under the null hypothesis of equal distribution. The test statistic is a function of the number of edges between sample 1, R_1, and the number of edges between sample 2, R_2:

$$S = A\Sigma^{-1}A^{\intercal},$$

where $A = (R_1 - E(R_1), R_2 - E(R_2))$. The statistic is asymptotically χ^2_2-distributed (see Theorem 5 in Chen and Friedman (2017)).

3.5 Simulations

In this section, we start by applying the three tests to one data set simulated from a five-factor model. We then conduct two simulations. In the first simulation, all DGPs are factor-analytic, and in the second, all are mixture models.

3.5.1 An example

As a first example, consider a five-factor model with 30 observed variables. The 30x5 loading matrix has weights sampled uniformly in the range [0.7, 1.0]. Each observed variable loads onto exactly one factor, chosen with equal probability. The diagonal matrix is also randomly picked from a uniform distribution between 0 and 1. A sample of 4,000 is drawn.

First, we run the DH test. As explained earlier, data generated from a factor analytic DGP should follow a normal distribution. The null hypothesis of normality in the DH test, given our sample is generated from a five-factor model, is true. We would expect to fail to reject the null. This is indeed the case, with a test statistic of 63.3 and p-value of 0.36.

Next, we check the relative goodness of fit. Here we fit a range of models to the sample, namely factor analysis with $1, 2, \ldots, 9$ factors and Gaussian mixture models with $1, 2, \ldots, 9$ components. For Gaussian mixture models, certain restrictions can be placed on the covariance matrices, producing different variants. Four are considered:

- full - no restrictions at all; each of the p by p covariance matrix can take any values
- diagonal - off diagonal entries are restricted to zero
- tied - the covariance matrices are identical
- spherical - identical diagonal entries within each covariance matrix.

This gives a total of 45 models, of which nine are factor models and 36 are mixture models.

A brief explanation of the inclusion of the one-mixture model is in order. With a factor DGP we would expect a one-mixture model to fit better than any mixture

Table 3.4: *BIC and two-sample test results of top models fitted to simulated data*

		Two-sample test	
Fitted model	BIC	statistic	p-value
Factor analysis			
fac 5	**357436**	2.96	0.228
fac 6	357600	1.62	0.445
fac 7	357756	5.07	0.0792
Mixture model			
mix 1 full	359762	1.49	0.476
mix 1 tied	359762	1.49	0.476
mix 2 tied	360028	0.52	0.772

models with more components. The inclusion is mostly for the purpose of simulation only, to see that this is indeed the case. Excluding them would typically result in having a two-component mixture model with the best fit among the mixture models, which is not as intuitive.

Of all 45 models fitted to the sample, the five-factor model has the best fit overall, according to the BIC in Figure 3.2. Table 3.4 reports the BIC of the best three factor models and that of the best three mixture models. As expected, the best-fitting mixture model is a one-mixture model. The fact that the top two mixture models give identical results is also to be expected, as the two variants (full and tied) in the case of a one-component model are the same.

We illustrate now the last test. We know that our step 2 has picked the correct model but are the parameters estimated accurately enough to simulate observations that resemble the actual data? According to the two-sample test, not only is the five-factor model the best model relatively, it is also capable of producing samples that pass the two-sample test, suggesting the model describes the data adequately well. However, an important observation here is that the converse is not true: models that can pass the two-sample test are not necessarily the true DGP.

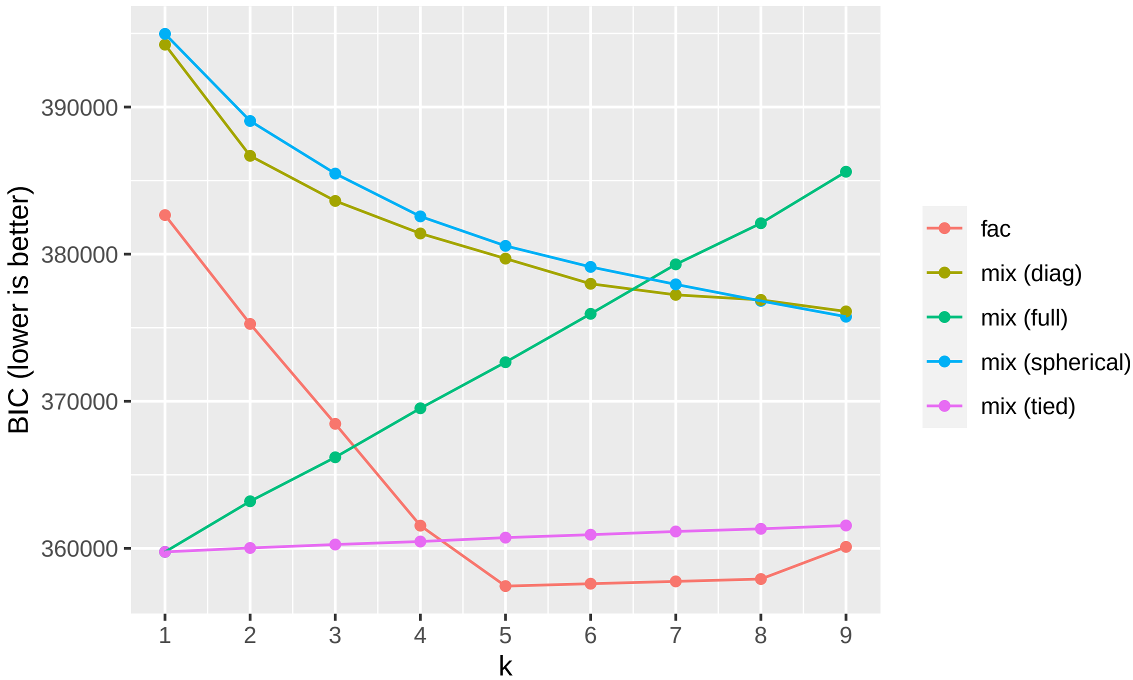


Figure 3.2: *BIC of models fitted to simulated data*

3.5.2 Factor analysis

To better understand the behaviour of the three tests, we consider a small number of different setups. Holding the number of factors at five, we vary the number of observed variables and sample sizes:

- observed variables, p: 15, 30
- sample size, n: 1000, 2000, 4000

In other words, there are two settings, one with 15 observed variables and the other with 30, both with five latent factors. For each DGP, we generate 500 loading matrices as described before, i.e. sampled uniformly in the range [0.7, 1.0] where each variable loads onto exactly one factor. For a given set of parameters, three separate samples of sizes 1000, 2000, and 4000 are drawn. For each sample, the three tests are performed.

For the Doornik-Hansen test of normality, of particular interest is the empirical Type I error rate. It is expected that, with large enough sample sizes, it would be close to the chosen significance level. Here it is chosen to be 5 percent. In the simulations, this is indeed the case - the empirical Type I error rates are all reasonably close to 5% for both DGPs across all sample sizes, ranging from 4.23 to 7.26 per cent.

Next consider the relative goodness of fit of the models. Here, of particular interest is how often the true DGP is ranked first. As the BIC is only comparable given the same data set, the BIC estimates are not reported but only used to report the proportion of time a fitted model is selected, i.e. with the smallest BIC among all 45 models. This is reported for models ever selected in Table 3.5. We see that only factor models are ever chosen, with k often close to the true value of 5. When p is 15, the true model ("fac 5" in the table) is selected in about 31 percent of the time, with BIC choosing a 4-factor model in about 54 percent of cases. However, when p is 30, we see that BIC only chooses either a 4- or 5-factor model and picks the correct model 95 percent of the time.

Lastly, we look at how often the chosen model can describe the actual data accurately enough to pass the two-sample test. We see that across different sample sizes the chosen model typically describes the data well enough, with the two-sample test rejected less than 5 percent of the time. In fact, with 15 observed variables and even 4000 observations the two-sample test never reject a 2-factor model in the four times the model is chosen. Similarly, when the 4-factor model is selected, the two-sample test rarely rejects the null.

3.5.3 Mixture model

In the next set of simulations, the DGPs are Gaussian mixture models, all with 5 components. We again consider two settings, with either 15 or 30 observed variables. For each DGP, we control how much the clusters can overlap. Following Melnykov, Chen, and Maitra (2012), we define the overlap between two clusters as

Table 3.5: *Summary of factor analysis simulations at different sample sizes (DGP is a 5-factor model)*

	N = 1000		N = 2000		N = 4000	
Fitted model	selected	rejected	selected	rejected	selected	rejected
15 observed variables						
fac 2	1%	0	1%	0.25	1%	0
fac 3	13%	0.031	13%	0.061	13%	0.045
fac 4	**55%**	0.018	**54%**	0.019	**54%**	0.022
fac 5	31%	0.019	31%	0.045	32%	0.019
30 observed variables						
fac 4	5%	0	5%	0.083	5%	0
fac 5	**95%**	0.0021	**95%**	0.0085	**95%**	0.021

the sum of two misclassification probabilities: the probability that an observation from the i-th cluster is classified as the j-th cluster, and the probability that an observation from the j-th cluster is classified as the i-th cluster. Here, we specify the pairwise overlap to be 0.03 on average. The covariance matrices are not otherwise restricted in any ways. For each of the 500 DGPs, four separate samples of sizes 1000, 2000, 4000 and 8000 are drawn.

For Gaussian mixture models, the null hypothesis of normality is not true. Therefore, we would expect the DH test to reject the null. We find that as the sample size increases, the rejection rate also increases. With 15 observed variables, the probability of rejecting the null increases from 82 percent when the sample size is 1000 and reaches 99 percent when the sample size is above 2000. With 30 observed variables, we find a similar pattern: the null hypothesis is rejected 39, 75 and 97 percent of the time as the sample size increases.

The relative goodness of fit of the various models is summarised in Table 3.6. We see that, with one exception, only mixture models are chosen. For the majority of the time, the BIC favours models with 5 clusters. When p is 15, a 5-cluster model is selected for 92.5 percent of the time when the sample size is 1000. We see that as the sample size increases, the best model switches from spherical to diagonal

Table 3.6: *Summary of mixture model simulations at different sample sizes (DGP is a 5-component mixture model)*

	N = 1000		N = 2000		N = 4000		N = 8000	
Fitted model	selected	rejected	selected	rejected	selected	rejected	selected	rejected
15 observed variables								
mix 4 diag			0%	0	0%	1		
mix 4 spherical	1%	0.5						
mix 5 diag			35%	0.011	**97%**	0.019	**90%**	0.029
mix 5 spherical	**92%**	0.13	**60%**	0.29	0%	0		
mix 6 diag			0%	0	1%	0.2	2%	0
mix 6 spherical	7%	0.15	4%	0.28	1%	0.33		
mix 7 spherical			0%	0				
30 observed variables								
fac 3	0%	0						
mix 4 spherical	15%	0.11	1%	0	0%	1		
mix 5 diag					4%	0.048	**90%**	0.024
mix 5 spherical	**84%**	0.036	**97%**	0.072	**95%**	0.22	1%	0.5
mix 6 diag							0%	0
mix 6 spherical	1%	0	1%	0	0%	0	0%	0
mix 8 spherical	0%	0						

covariances. When p is 30, a similar pattern is observed, with the BIC choosing the a 5-cluster spherical model most often for sample sizes of 1000, 2000 and 4000 but switching to a 5-cluster diagonal model when the sample size reaches 8000.

While the BIC selects the optimal number of clusters with high chances, the spherical models are generally more likely to be rejected in a subsequent two-sample test than the diagonal models. In our DGPs the covariance matrices are allowed to have non-zero values in the off-diagonal. In other words, mixture models with diagonal covariance matrices are not the true model. In our simulations, the two-sample test has a low Type I error rate but does not have high statistical power.

3.5.4 Discretisation

One major difference between the simulations thus far and the actual analysis is that responses to personality questionnaires are typically on a 5-point or 7-point scale. Actual personality data sets are therefore highly discrete. To find out how

Table 3.7: *Summary of factor analysis simulations at different sample sizes (DGP is a 5-factor model, discretised)*

	N = 1000		N = 2000		N = 4000	
Fitted model	selected	rejected	selected	rejected	selected	rejected
15 observed variables						
fac 2	1%	0	1%	0	1%	0
fac 3	15%	0.042	15%	0.014	15%	0.014
fac 4	**49%**	0.021	**50%**	0.028	**50%**	0.016
fac 5	35%	0.017	34%	0.006	34%	0.012
mix 9	1%	1				
30 observed variables						
fac 3	0%	0.5				
fac 4	7%	0.03	5%	0	5%	0
fac 5	**93%**	0.022	**95%**	0.019	**95%**	0.0064

discrete data could affect the performance of the three tests, we run the simulations again but this time all simulated data are rounded to the nearest integer.

First, we consider the case where the DGP is a discretised 5-factor model. The DH test appears to perform just as well, with the empirical Type I error ranging between 0.032 and 0.058, which is close to the chosen 5% significance level. As Table 3.7 shows, the other two tests also give broadly similar results to those for continuous data. For example, when there are 15 observed variables, the BIC also favours the 4-factor model over the 5-factor model. However, with 30 observed variables, it again chooses the correct 5-factor model over 90 percent of the time. The two-sample test is likewise rejected in less than 5 percent of cases whenever the correct model is chosen.

For the discretised 5-component mixture model, the performance of the DH test is the same as before when there are 15 observed variables, rejecting more than 99 percent of cases for all sample sizes. When there are 30 observed variables, the DH test only rejects the null of normality 20, 41, 76.6 percent of the time as the sample size increases. These are all lower than before when the simulated data are not discretised.

Discretisation also affects model selection adversely. As shown in Table 3.8, with discretised data, the BIC now chooses a 9-component mixture model most often in almost all scenarios. However, the BIC tends to choose mixture models with fewer components (still more than 5) as the number of observations and variables increases. Given that the BIC chooses the incorrect model, the performance of the two-sample tests suffers as a result, with the null hypothesis almost always rejected. Overall, we see that mixture models appear to be more sensitive to discretisation.

Table 3.8: *Summary of mixture model simulations at different sample sizes (DGP is a 5-component mixture model, discretised)*

	N = 1000		N = 2000		N = 4000	
Fitted model	selected	rejected	selected	rejected	selected	rejected
15 observed variables						
mix 5 diag			0%	1		
mix 6 diag	0%	1	0%	1	1%	1
mix 6 full					0%	1
mix 7 diag	6%	1	6%	1	8%	1
mix 7 full					0%	1
mix 8 diag	21%	1	27%	1	22%	1
mix 8 full					1%	1
mix 9 diag	**73%**	1	**66%**	1	**66%**	1
mix 9 full					3%	1
30 observed variables						
fac 1	0%	1				
fac 2	1%	1				
fac 3	1%	1				
mix 3 spherical	0%	1				
mix 4 spherical	14%	0.99	1%	1		
mix 5 spherical	30%	0.94	**79%**	1	6%	1
mix 6 spherical	1%	0.86	8%	1	8%	1
mix 6 diag	2%	1	0%	1		
mix 7 spherical			1%	1	20%	1
mix 7 diag	5%	1	1%	1		
mix 8 spherical					26%	1
mix 8 diag	12%	1	2%	1		
mix 9 spherical					**38%**	1
mix 9 diag	**33%**	1	8%	1	0%	1

3.6 Data

This section describes the two data sets used in the analysis. The first data set is from the Household, Income and Labour Dynamics in Australia (HILDA) survey, and the second from the Open Source Psychometrics Project (OSPP).

3.6.1 HILDA

The HILDA survey is a nationally representative data set. In the analysis, data from wave 13, collected in 2013, are used. Like other national household surveys, the HILDA survey has a personality inventory designed to measure the Big Five. This inventory is derived from the Trait Descriptive Adjectives (Saucier, 1994a), which in turn are based on Goldberg's Big-Five Factor Markers (Goldberg, 1992b). These questions are included in the Self Completion Questionnaire. In total there are 36 trait descriptive adjectives. For each, the respondents self-assessed the degree to which the adjective described them on a 7-point scale, from 1 "Does not describe me at all" to 7 "Describes me very well". Out of 15,361 individuals who returned their Self Completion Questionnaire, 14,421 (94%) responded to all 36 questions.

3.6.2 OSPP

The Open Source Psychometrics Project began in 2011. It offers various personality tests and other psychological assessments on their website.[8] Participants in these tests were informed and asked to confirm, both before and after the test, that their responses could be used for research purposes. The survey responses used in the analysis were collected in 2012. The Big Five personality test from the OSPP, like the one in the HILDA survey, uses the Big-Five Factor Markers by Goldberg (1992b). The questionnaire is based on the 50-item version from the International Personality Item Pool. For each of the 50 statements - 10 for each of the Big Five - the participants self-rated on a 5-point scale, from 1 "Disagree" to 5 "Agree". Out of 19,719 records, 19,718 have valid responses to all 50 questions.

3.7 Results

In this section, we apply the empirical tests to the two data sets from the HILDA survey and the OSPP. Given the consensus on the Big Five, the expectation would be that a factor model with 5 factors should fare better than any other models, and also that it should describe the data well. In terms of the epmirical tests then,

[8] https://openpsychometrics.org

it should have a lower BIC than other models, and should capture the salient features of the data such that a generated sample from the factor model should resemble the real data set. For the two data sets considered, we show that this is not the case.

First, we look at the results for the HILDA data. Like in the simulations, a total of 45 models are fitted to the data: 9 factor models varying the number of factors from 1 to 9, and 36 mixture models varying the number of clusters also from 1 to 9, and each with 4 different covariance structures. The relative goodness of fit of the models, as measured by the BIC, is shown in Figure 3.3. Overall the best model with the lowest BIC is a 3-cluster mixture model, with full covariance matrices.[9] The best factor model is one with 9 factors, although the BIC is relatively similar for factor models with 5 or more factors. However, compared to the best mixture model, none of the factor model has a lower BIC.

Next we check whether any of the fitted models could adequately describe the sample. This model diagnostic is conducted on the top 3 mixture models and the top 3 factor models, in which we perform a two-sample equality test between the simulated data from a given fitted model and the HILDA sample. If a particular fitted model can generate samples that pass the two-sample test, then it means that the fitted model captures the salient features of sample very well. The results of these two-sample tests are reported in Table 3.9. We see that none of those models could pass the two-sample test, where the null hypothesis of equality rejected with a p-value close to zero. In other words, these models fail to capture certain aspects of the actual data.

Turning to the results for the OSPP data, we see broadly similar results. In terms of the relative goodness of fit, as shown in Figure 3.4, the model with the smallest BIC is from an 8-cluster mixture model, with the same (i.e. tied) covariance matrices.

[9] The BIC is not shown for the mixture models with 6 or more clusters (with full covariance matrices) because for these models one or more of the estimated covariance matrices is singular.

Table 3.9: *BIC and two-sample test results of top models fitted to HILDA Big Five*

		Two-sample test	
Fitted model	BIC	statistic	p-value
Factor analysis			
fac 9	1708359	284.87	0
fac 7	1708732	284.29	0
fac 8	1709540	285.33	0
Mixture model			
mix 3 full	**1593191**	53.75	0
mix 4 full	1595015	54.34	0
mix 5 full	1597144	50.00	0

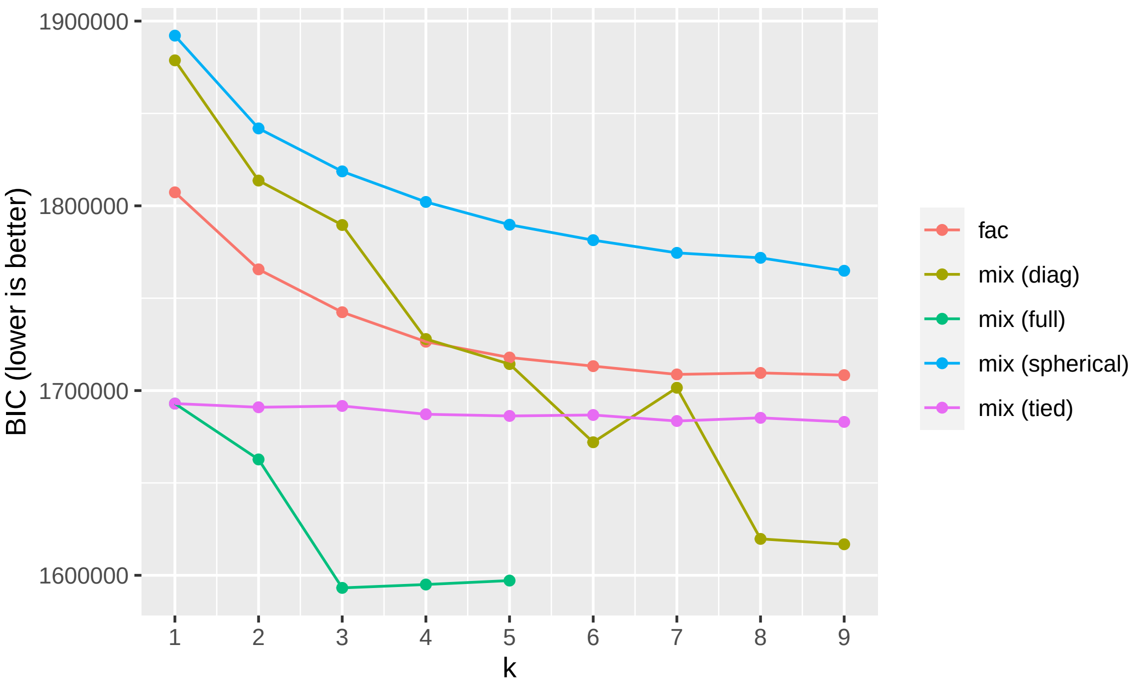


Figure 3.3: *BIC of models fitted to HILDA Big Five*

Table 3.10: *BIC and two-sample test results of top models fitted to OSPP Big Five*

		Two-sample test	
Fitted model	BIC	statistic	p-value
Factor analysis			
fac 8	2792699	65.57	0
fac 9	2793279	64.47	0
fac 7	2793925	58.66	0
Mixture model			
mix 8 tied	**2743725**	50.20	0
mix 9 tied	2744831	50.56	0
mix 7 tied	2746299	50.66	0

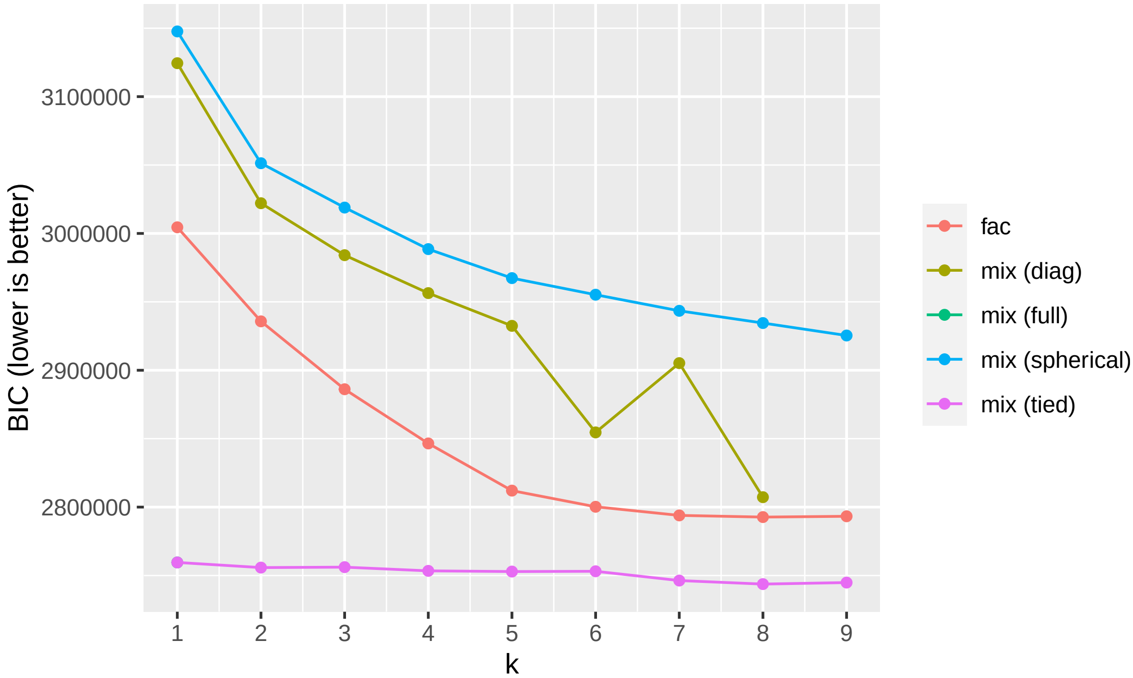


Figure 3.4: *BIC of models fitted to OSPP Big Five*

The best factor model is one with 8 factors, although again the factor models with 5 or more factors have relatively similar BIC. However, none of the best-fitting models can pass the two-sample test. As reported in Table 3.10, the null hypothesis of equality is rejected for the top 3 mixture models and the top 3 factor models.

To summarise the findings, in each of the data set we observe that:

1. the hypothesis of joint normality is rejected,

2. no factor model has an estimated BIC lower than the best mixture model, and
3. data simulated from the best-fitting factor model do not pass the multivariate two-sample test.

3.8 Conclusion

Despite the success and ubiquitous use of factor analysis to study personality, researchers have continued to find clusters in personality data. The two sub-literatures have existed alongside each other almost from the beginning, even though as we have shown the two make incompatible assumptions about the number of clusters in the data. We explained this phenomenon – replicating the Big Five on the one hand and finding clusters on the other – by pointing out that statistics used to support the replicability of factor analysis are ill-suited to compare relative model fit. When we compare the relative model fit between factor analytic and mixture models in the two data sets considered, we find that mixture models are relatively better according to the BIC. The best mixture model, however, still should not be treated as the true DGP, given the results from our two sample tests. In other words, the data sets exhibit patterns that are not captured by any of the factor analytic or mixture models.

Chapter 4

Interaction effects of non-cognitive skills on labour market earnings

4.1 Introduction

Both cognitive and non-cognitive skills have been shown to predict many socioeconomic outcomes (Heckman, Jagelka, and Kautz, 2021; Almlund et al., 2011b), including schooling decisions, occupational choices, and wages (Heckman, Stixrud, and Urzua, 2006b). The wage returns to these skills, especially their main effects, are relatively well understood. However, findings on their interaction effects on wages are less consistent. For example, Nyhus and Pons (2005) found evidence of interaction between personality and tenure, whereas Heineck and Anger (2010) did not. Green and Riddell (2003) found that cognitive skills did not interact with any unobserved skills. Deming (2017) and Weinberger (2014), on the other hand, found that cognitive and social skills had an interaction effect on wages.

One challenge facing researchers is that in a typical wage regression the number of two-way interactions that can be tested for is practically limited. It is common practice to restrict the inclusion of interaction terms to those that are driven by economic theory or have been considered by previous research. This is a problem because including too few interaction terms means false negatives but including too many runs into multiple testing problem. Another common practice is to restrict any interactions between two variables to their products. In this chapter, we address these limitations by combining two techniques: 1) multivariate adaptive regression splines (MARS) and 2) interaction test statistics. The use of MARS has the advantage that it can flexibly approximate two-way interactions without the need to explicitly specify them in the regression. It can also approximate non-linear main effects non-parametrically. The interaction test statistics could then be used to find which interaction terms, if any, are statistically significant. Understanding the interaction effects of various skills on wages can shed light on what combinations of skills are most valued in the labour market but also whether the returns to these skills differ by other labour attributes.

Using data from the Household, Income and Labour Dynamics in Australia (HILDA) survey, we explore the interaction effects on wages in the prime working age population. As some previous research (e.g. Nyhus and Pons, 2005) has shown gender differences in the effect of non-cognitive skills on wages, we estimate the wage regression separately for females and males. For females, we find no interaction effects on wages. For males, we find seven two-way interaction terms among five variables, namely Big 5 conscientiousness, occupation, industry, whether a casual or permanent worker, and full-time or part-time work. These are different interaction effects from what others in the literature have found, namely between cognitive skills and social skills (Deming, 2017; Weinberger, 2014) and between cognitive skills and emotional stability (Palczyńska, 2021).

This chapter is organised as follows. The next section summarises the key findings from previous research on the interaction effects of cognitive and non-cognitive skills on wages. Section 4.3 describes the key components of our approach, namely MARS and the H-statistics. In section 4.4 we illustrate how our approach can be used to discover interactions in three simulated data sets. In section 4.5, we describe the HILDA data used in the analysis, and in section 4.6 present our findings separately for females and males.

4.2 Recent studies on the interaction effects on earnings

Interaction effects on earnings have been found between social and cognitive skills in the US labour market (Deming, 2017; Weinberger, 2014). The complementarity between emotional stability and cognitive skills have also been found in the Polish labour market (Palczyńska, 2021). In this section we review these three studies in greater detail and summarise earlier research on other interactions briefly.

In the US labour market, workers with high social and cognitive skills were paid more, and this earnings premium has increased over time. Using longitudinal data on two cohorts of high school seniors in 1972 and 1992, Weinberger (2014) compared the returns to social and cognitive skills on earnings measured seven years after high school. Social skills were measured by participation in sports or leadership activities, whereas cognitive skills were measured by math scores in high school. Weinberger (2014) then showed that, in various pooled regressions of the two cohorts, the coefficient of the three-way interaction term between participation in sports or leadership roles, above-median math scores, and time trend was consistently positive and statistically significant. In other words, workers having both great social and cognitive skills were valued even more in 1999 than they were in 1979. This finding of growing complementarity between social and cognitive skills from 1979 to 1999 was also robust to the inclusion of additional controls, including family structure, parent education, and psychological measures such as self-esteem and locus of control. Weinberger (2014) estimated that the earnings premium for multiskilled workers on average increased by 4.8 to 6.8 percent every 10 years in that period.

To understand the increasing share of jobs requiring a high level of social skills and their increasing returns in the US labour market, Deming (2017) adapted a Ricardian model of trade to model team productivity and tested its predictions empirically using National Longitudinal Survey of Youth (NLSY) data. Similar to what Weinberger (2014) has found, Deming (2017) found evidence for increasing complementarity between cognitive skills and social skills. In his stylised model of team productivity, workers are more productive the higher their cognitive abilities are. Social skills, or the ability to work with others, do not enhance a worker's productivity directly. However, a worker with better social skills is assumed to have a lower cost of coordinating with other workers. Therefore, a team of workers endowed with more social skills would be better at trading tasks with one another, allowing each to specialise in tasks in which they have a comparative advantage. Though highly stylised, this model of team productivity not only explains the growth in social jobs over time, but it also derives the complementarity between cognitive and social skills from economic principles.

Using a representative sample of the Polish working-age population, Palczyńska (2021) analysed the complementarities between cognitive skills and non-cognitive skills, including grit, the Big Five personality traits, and social skills (as measured by extraversion). She did not find any interaction between cognitive and social skills but found an interaction between cognitive skills and emotional stability. That wage returns to cognitive skills differ by emotional stability, Palczyńska (2021) explained, could be because neurotic individuals tend to have lower self-esteem and therefore underestimate their productivity. Consistent with previous research, Palczyńska (2021) found that wages increase with conscientiousness but decrease with agreeableness and neuroticism.

Using data from the 1994 Canadian International Adult Literacy Survey, Green and Riddell (2003) selected a sample of 646 full-time male workers and found that cognitive skills did not interact with non-cognitive skills or any unobserved skills.[1] However, they found that cognitive skills and skills acquired through experience had a negative interaction effect on earnings, suggesting that cognitive skills could be more important at the beginning of one's career but experience became more valued later on.

Previous studies have examined the interaction effects between personality and tenure but findings are mixed. Using Dutch data from the DNB Household Survey, Nyhus and Pons (2005) found interaction effects between tenure and conscientiousness and between tenure and openness to experience for men only. On the other hand, using data from the German Socio-Economic Panel Study, Heineck and Anger (2010) only found interaction effects between tenure and openness to experience for women. Using data from the British Household Panel Study, Heineck (2011) found no interactions between tenure and personality at all.

The wage returns to personality traits may also differ by the level of education. Nyhus and Pons (2005) found that education interacted with extraversion and openness to experience for men but with emotional stability for women.[2] Among men with high IQ, Gensowski (2018) found that those with a graduate degree, compared to those with a bachelor's or less, had higher returns to extraversion and conscientiousness.

[1] With no measures on non-cognitive skills, Green and Riddell (2003) used quantile regressions to examine if the returns to cognitive skills varied across quantiles of the earnings distributions.

[2] In their analysis, education was used as a proxy for occupation and their findings could also mean that the wage returns to those traits could differ by types of jobs.

4.3 Method

In this section, we describe the three components of our approach, namely Multivariate Adaptive Regression Splines (MARS), a variable importance measure, and the H-statistics.

Our approach is not all that different from the conventional approach to identifying interactions. In the conventional approach, one would fit an ordinary least squares (OLS) model with all hypothesised interaction terms explicitly specified, and then perform a t-test on the coefficient of each term. In our approach, instead of an OLS, we use a MARS model to alleviate the need to manually specify every interaction term. For hypothesis testing, instead of t-tests, we use the H-statistics. To help reduce the number of tests required, it helps to first remove any variables that do not predict wages. To identify these non-predictors, we use a variable importance measure.

At a high level, our approach proceeds as follows:

1. fit a MARS model to capture the relationship between the explanatory variables and wages,
2. rank the variables by their relative importance and remove the weakest predictors, and
3. calculate the H-statistics and identify any interactions.

For clarity, some steps have been omitted. The exact procedure will be laid out step by step and illustrated with simulated data, but first we explain each of the three components.

4.3.1 Multivariate Adaptive Regression Splines

MARS was introduced by Friedman (1991). The following description, including the notations used, is based on section 9.4 in Hastie et al. (2009) (pp. 321-329).

MARS is a nonparametric regression method that uses piecewise linear basis functions (hence regression splines) to model the relationship between the outcome and the explanatory variables. MARS is adaptive in that it algorithmically selects the basis functions that best capture the main effects and interaction effects between the variables. These properties make it a better alternative to an OLS model as it removes the need to specify the functional form of the regression.

The regression function $f(\mathbf{X})$ in MARS, where $\mathbf{X}$ denotes the p explanatory variables $(X_1, X_2, \ldots, X_p)$, has the following form:

$$f(\mathbf{X}) = \beta_0 + \sum_{m=1}^{M} \beta_m h_m(\mathbf{X})$$

where β_0 is the constant term and β_m are the coefficients. What $h_m(\mathbf{X})$ is depends on what is called the degree of the MARS model. In a degree 1 model, $h_m(\mathbf{X})$ is a basis function of $\mathbf{X}$. In a degree 2 model, each $h_m(\mathbf{X})$ can be either a basis function of $\mathbf{X}$ or a product of two such functions. These products are what enable a MARS model to capture interaction effects.

A MARS model is built in two stages: forward selection and backward deletion.

In the forward selection stage, the algorithm begins with an initial model containing only the constant function and iteratively selects basis functions to add to the model until some stopping criterion is met. Specifically, the collection of basis functions the algorithm can select from is:

$$\mathcal{C} = \{max(0, X_j - t), max(0, t - X_j)\}_{t \in \{x_{1j}, x_{2j}, \ldots, x_{Nj}\}, j=1,2,\ldots,p},$$

where t is a knot location and x_{ij} is the i-th value of variable X_j. In other words, for each variable X_j and each observed value x_{ij} the algorithm creates two basis functions. They are called a reflected pair and form the building blocks of the terms $h_m(\mathbf{X})$ in the model. Starting from a constant function $h_0(\mathbf{X}) = 1$, in each iteration m the algorithm adds two terms to the model:

$$h_m(\mathbf{X}) \cdot max(0, X_j - t) \text{ and } h_m(\mathbf{X}) \cdot max(0, t - X_j),$$

where $h_m(\mathbf{X})$ can be any terms already in the model and the reflected pair chosen is such they lead to the largest decrease in the residual sum of squares. The forward selection stage stops when the number of terms reaches some predefined value or when the R-squared increases by less than some threshold.

In the backward deletion stage, to avoid overfitting, the algorithm removes terms from the model. The number of terms d to retain is chosen such that the generalised cross-validation is minimised:

$$\text{GCV}(d) = \frac{\sum_{i=1}^{n} \left(y_i - \hat{f}_d(x_i) \right)^2}{\left(1 - \frac{(r+3K)}{n} \right)^2},$$

where r is the number of linearly independent functions, K is the number of knots (Hastie et al., 2009, p. 325). That is, the GCV is the sum of squared residuals, penalised for the number of free parameters.

4.3.2 Variable importance

While MARS is more flexible than an OLS, it comes at a cost. In an OLS, the standard errors of the estimated marginal effects are readily available. It is therefore straightforward to test, for example, whether the estimated coefficient of any explanatory variable is statistically different from zero. By contrast, in MARS these standard errors are difficult to obtain. Also, each variable in a MARS model may be used in multiple basis functions, making it difficult to discern which variables are more important for prediction. As an alternative, we measure the importance of a variable with what is known as permutation feature importance (Fisher, Rudin, and Dominici, 2019). This allows us to rank the variable by the strength of its predictiveness, thereby answering the question which variables are more important. Identifying and removing any variables that do not or only weakly predict wages also helps reduce the number of tests needed when testing for interactions.

Permutation feature importance measures the importance of a variable by how much the model's prediction error increases when the values of the variable are randomly permuted. Permutation will result in a larger increase in prediction error the more a variable contributes to the predictiveness of the model. For a given variable X_j, it is calculated as the ratio between the mean squared error (MSE) of the model with permutation and the MSE of the original model:

$$\mathrm{PFI}_j = \frac{\frac{1}{P}\sum_{p=1}^{P}\mathrm{MSE}_p}{\mathrm{MSE}},$$

where MSE_p is the MSE of the p-th permutation and MSE is the MSE of the original model. We run 50 permutations. That is, P is 50.

Fisher, Rudin, and Dominici (2019) suggested a procedure that turns the PFI into a hypothesis test. While we calculate the upper and lower bounds of the PFI and use the lower bound to remove weak predictors, our goal is not to perform hypothesis testing. Ultimately there is a tradeoff between type I and type II errors. In fact, we err on the side of including than excluding variables.

4.3.3 H-statistics

The interaction test statistics were introduced by Friedman and Popescu (2008). There are two reasons why we need them even though it is straightforward to tell from any given MARS model what the interactions are, if there are any. Firstly, an interaction effect could be approximated by multiple pairs of basis functions in a MARS model. Therefore, direct comparison of the strength of two interaction terms can be difficult. Note also that the variable importance measure is not suitable because it captures both main and interaction effects of a given predictor on the outcome variable, and not just the interaction effects. Secondly, the interaction terms selected by the MARS model are driven by predictiveness. While the backward deletion stage helps mitigate overfitting, the resulting terms could still be spurious and, more importantly, dependent on the chosen hyperparameters. The interaction test statistics offer a more principled method of testing whether an interaction effect is statistically significant.

The interaction test statistics can measure the strength of low-order interactions in a given regression function. Below we describe two such statistics, H_j for overal interactions and H_{jk} for two-way interactions. In their paper, Friedman and Popescu (2008) also discussed a test statistic for a three-variable interaction. This is omitted below as we only consider MARS models of degree 2 and therefore any interactions will be of second order.

Both the H_j and H_{jk} statistics are formulated using partial dependence (PD) functions. To better understand these statistics, we will first introduce what PD functions are and their properties that these statstics rely on.

Let $\mathbf{X}_{\setminus j}$ denote all explanatory variables other than X_j. The PD function of X_j is defined as:

$$PD_j(X_j) = \mathbb{E}_{\mathbf{X}_{\setminus j}}\left[f(X_j, \mathbf{X}_{\setminus j})\right]$$

and is estimated by:

$$\frac{1}{n}\sum_{i=1}^{N} f(X_j, \mathbf{X}_{i\setminus j}).$$

To see how PD functions can be useful for detecting interactions in a regression function, note that if X_j does not interact with any other variables, then the regression function is equal to the sum of two PD functions:

$$f(\mathbf{X}) = PD_j(X_j) + PD_{\setminus j}(\mathbf{X}_{\setminus j}),$$

where the first depends only on X_j and the second only on variables other than X_j. This property of the PD functions underlies the H_j^2 statistics proposed by Friedman and Popescu (2008), which is defined as:

$$H_j^2 = \sum_{i=1}^{n} \left[f(\mathbf{X}_i) - PD_j(X_{ij}) - PD_{\setminus j}(\mathbf{X}_{i\setminus j}) \right]^2 / \sum_{i=1}^{n} f^2(\mathbf{X}_i).$$

If X_j does not interact with any other variables, then the numerator and therefore the H_j^2 statistic will be zero. It increases the stronger X_j interacts with other variables.

The H_{jk}^2 statistic is defined similarly. If two variables X_j and X_k do not interact, the PD function over the two variables can be expressed as the sum of two PD functions:

$$PD_{jk}(X_j, X_k) = PD_j(X_j) + PD_k(X_k).$$

The H_{jk}^2 is calculated as:

$$H_{jk}^2 = \sum_{i=1}^{n} \left[PD_{jk}(X_{ij}, X_{ik}) - PD_j(X_{ij}) - PD_k(X_{ik}) \right]^2 / \sum_{i=1}^{n} PD_{jk}^2(X_{ij}, X_{ik}).$$

It is equal to zero if X_j and X_k do not interact and increases with the strength of the interaction effect.

After calculating the interaction test statistics, to perform inference we also need to derive the reference distribution under the null hypothesis that no interaction exists. To derive this null distribution, Friedman and Popescu (2008) Friedman and Popescu (pp. 937-938) proposed a variant of the parametric bootstrap. In each bootstrap, the first step is to generate a data set such that the null hypothesis is true. For the overall H_j statistic, this requires generating the artificial outcome variable $\tilde{y}_i$ from the actual explanatory variables $\mathbf{X_i}$ such that there is no interaction involving X_j. For the two-way H_{jk} statistic, the artificial $\tilde{y}_i$ should be generated such that there is no pairwise interactions between X_j and X_k, while any other interactions are permitted. The following procedure is used to generate $\tilde{y}_i$:

$$\tilde{y}_i = f_0(\mathbf{X}_i) + (y_{p(i)} - f_0(\mathbf{X}_{\mathbf{p(i)}})),$$

where f_0 is a regression function such that the null hypothesis is true, and $p(i)$ is a random integer (between 1 and N) for the i-th row. Here, for f_0 we use a restricted MARS model such that the null hypothesis is true. In the second step of the bootstrap, on each artificial data set we apply the same modelling procedure used on the actual data, i.e. an unrestricted degree-2 MARS model. From the fitted MARS model, we then calculate the respective interaction test statistics. The bootstrap is then repeated 500 times, with the resulting 500 test statistics used as the reference distribution.

4.4 Illustrations

In this section, we demonstrate how the three-step procedure can help identify interaction effects. Considered are three scenarios. In each scenario, the simulated data set has 3000 observations, generated according to the following data generating process:

$$Y = -X_1^2 + X_2 + X_3 + X_4 + \beta X_2 \cdot X_3 + X_3 \cdot X_4 + \epsilon$$

The key difference between the three scenarios is in the value of the β coefficient. It is set to 1 in the first scenario, 0.3 in the second, and 0 in the third. In other words, the three scenarios have respectively strong, weak, and no interaction effects between X_2 and X_3. The interactive term between X_3 and X_4 is always present. Of the four main effects, X_1 enters as a squared term with a -1 coefficient (to mimic the stylised relationship between wage and age); the other three main effects, X_2, X_3 and X_4, enter linearly with a coefficient of 1. Also present in the simulated data set are two noise variables, X_5 and X_6, that are independent of the outcome. All four explanatory variables, X_1 to X_4, and the two noise variables, X_5 and X_6, are drawn from a standard normal distribution. The error term, ϵ, is also normally distributed with a zero mean. Its variance in each scenario is such that the signal-to-noise ratio is approximately equal to 1.

In all three scenarios, the success of the three-step procedure is measured by:

1. how well the fitted MARS model can approximate the relationship between each variable and the outcome;
2. whether the variable importance tests correctly identify the predictors, X_1 to X_4, from the non-predictors, X_5 and X_6, and;
3. whether the overall and two-way H-statistics can correctly identify the interactive terms.

4.4.1 Scenario 1: Strong interaction

In the first scenario, there are two strong interactive terms, one between X_2 and X_3 and another between X_3 and X_4. The six generated variables, X_1 to X_6, and the outcome, Y, are the only inputs to the MARS model. Importantly, no interactive terms are provided to or in any way specified in the model.

The fitted MARS model shows that it captures the true data generating process fairly well. The model has four main effects, X_1 to X_4, and two interaction effects, $X_2 \cdot X_3$ and $X_3 \cdot X_4$. The functional form is also well approximated, as can be seen in the partial dependence plots in Figure 4.1. We see that the outcome variable first increases with X_1 when X_1 is negative then decreases when it is positive. This correctly reflects the inverted-U relationship between Y and X_1 in the DGP. For the other main effects, we see that the outcome increases linearly with X_2, X_3 and X_4. This is expected. Given that all three variables have a coefficient (and marginal effect) of 1 in the DGP, the true partial dependence function is a straight line with

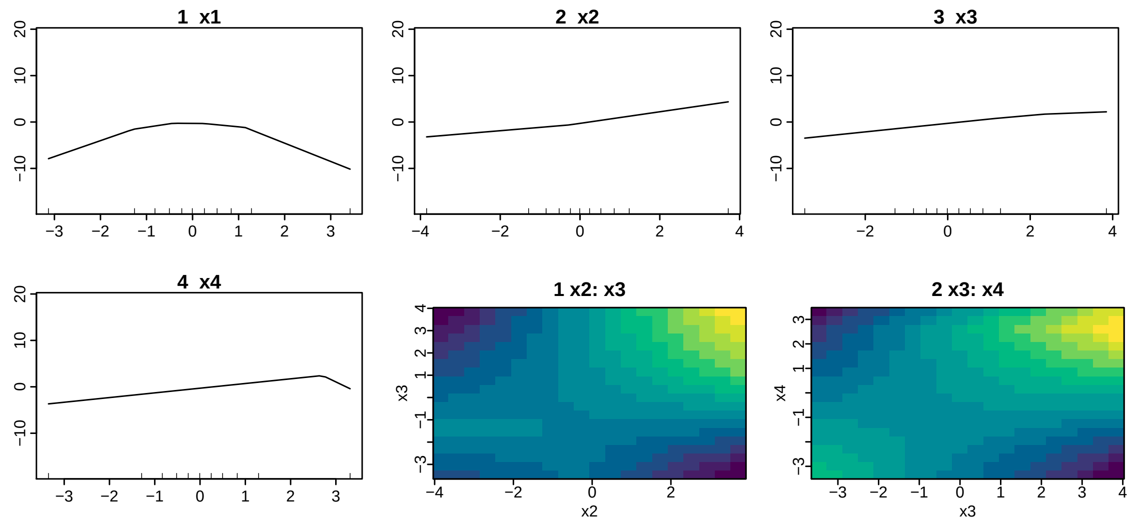


Figure 4.1: *Partial dependence plots of the final model- Scenario 1*

a slope of 1. The estimated partial dependence functions in Figure 4.1 are good approximations. For the two interactive terms, $X_2 \cdot X_3$ and $X_3 \cdot X_4$, the model also captures the salient feature of a product-form interaction, namely that the predicted outcome has higher values when the two variables being interacted are both positive or both negative, and lower when they take opposite signs.

The variable importance tests correctly identify X_1 to X_4 as the only predictors. We see in Figure 4.2 that permuting the values of X_5 or X_6 has no impact on the predictive performance of the model. Because we measure changes in predictive performance as a ratio, the value of 1 for both X_5 and X_6 means that the mean squared error (MSE) of the model is unchanged after permutation. This is expected given that the fitted model has no terms involving X_5 or X_6. Of the four predictors, X_3 is the strongest. The average value of 1.9 indicates that the MSE of the fitted model would increase by 90% if the values of X_3 are randomly permuted. Variables X_2 and X_4 are less strong compared to X_3, with variable X_1 being the weakest predictor. This is again consistent with the DGP. All four predictors enter as main effects but X_3 is present in both interactive terms, with X_2 and X_4 each present in just one interaction, and X_1 in neither.

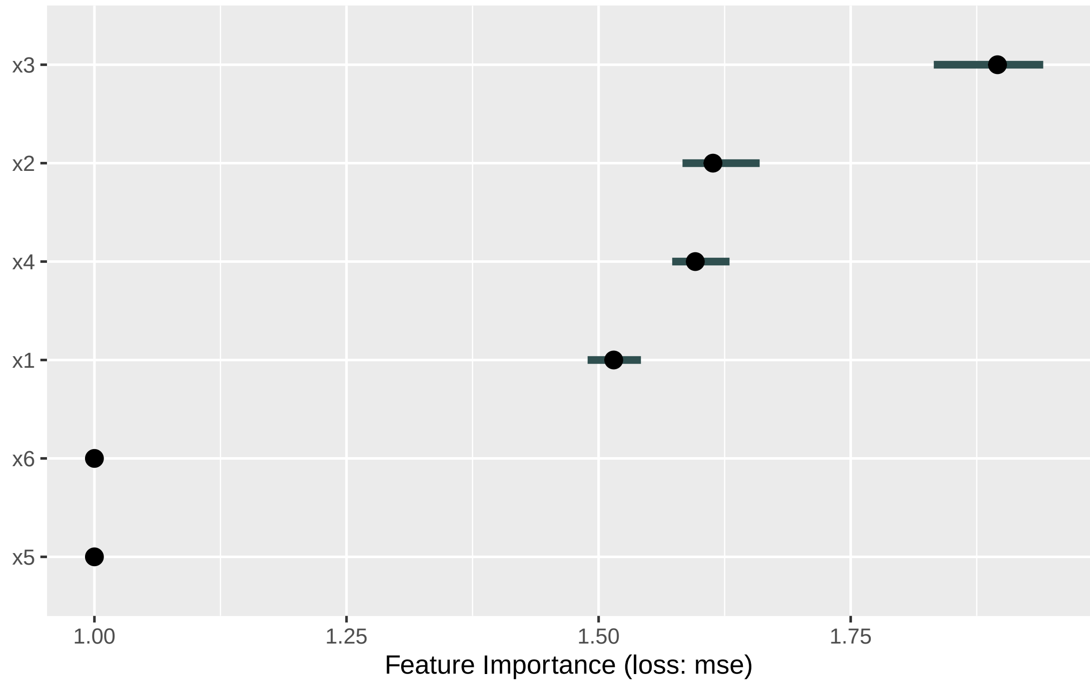


Figure 4.2: *Variable importance plot of the initial model - Scenario 1*

While the fitted model has $X_2 \cdot X_3$ and $X_3 \cdot X_4$ as interactive terms, we do not conclude they are present in the DGP yet. To check whether these interactive terms are spurious or not, we carry out the Friedman-Popescu tests. First, for each predictor in the model we calculate the H_j statistic to measure its overall interaction strength. In Figure 4.3, the H_i statistics are represented as dots and the bootstrapped 95th percentiles are represented as crosses. We see that only for X_2, X_3 and X_4 the overall interaction strength is above the respective 95th percentile. Therefore, at the 5% confidence level, we conclude that only X_2, X_3 and X_4 are involved in interactions. This gives us three possible combinations, namely $X_2 \cdot X_3$, $X_2 \cdot X_4$ and $X_3 \cdot X_4$. For the three combinations, we then calculate the two-way H_{jk} statistics and the bootstrapped 95th percentiles. As shown in Figure 4.4, we only find significant two-way interaction between X_2 and X_3, and between X_3 and X_4. In the DGP, this is indeed the only two interactions.

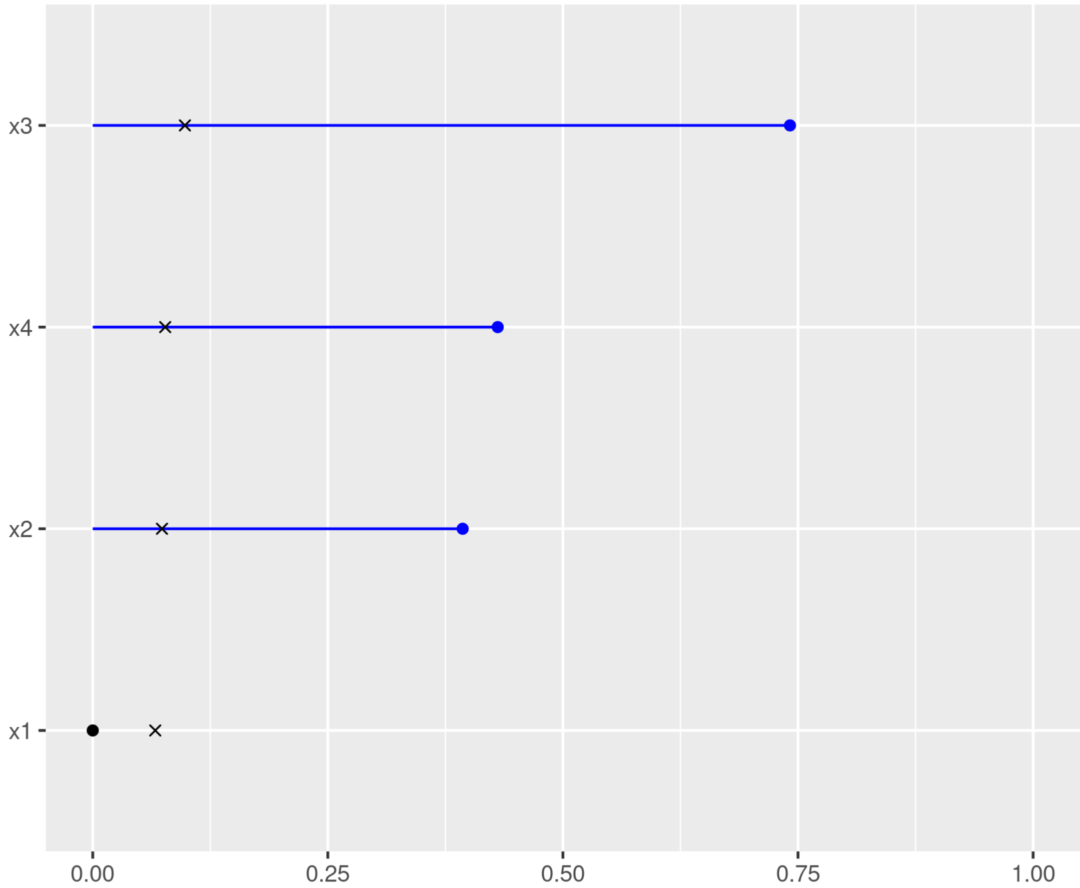


Figure 4.3: *Overall interaction strength of all variables - Scenario 1*

4.4.2 Scenario 2: Weak interaction

In this second scenario, the $X_2 \cdot X_3$ interaction is weaker, with a β coefficient of 0.3 instead of 1 in the first scenario. Again, we begin by fitting a MARS model. The partial dependence plots of the model are shown in Figure 4.5. The main effects are by and large the same as before in scenario 1. However, the fitted model now has an extra interactive term between X_1 and X_4, which is not in the DGP. For the $X_2 \cdot X_3$ term, the response surface is also more flat than in scenario 1. This is expected given that β is now smaller.

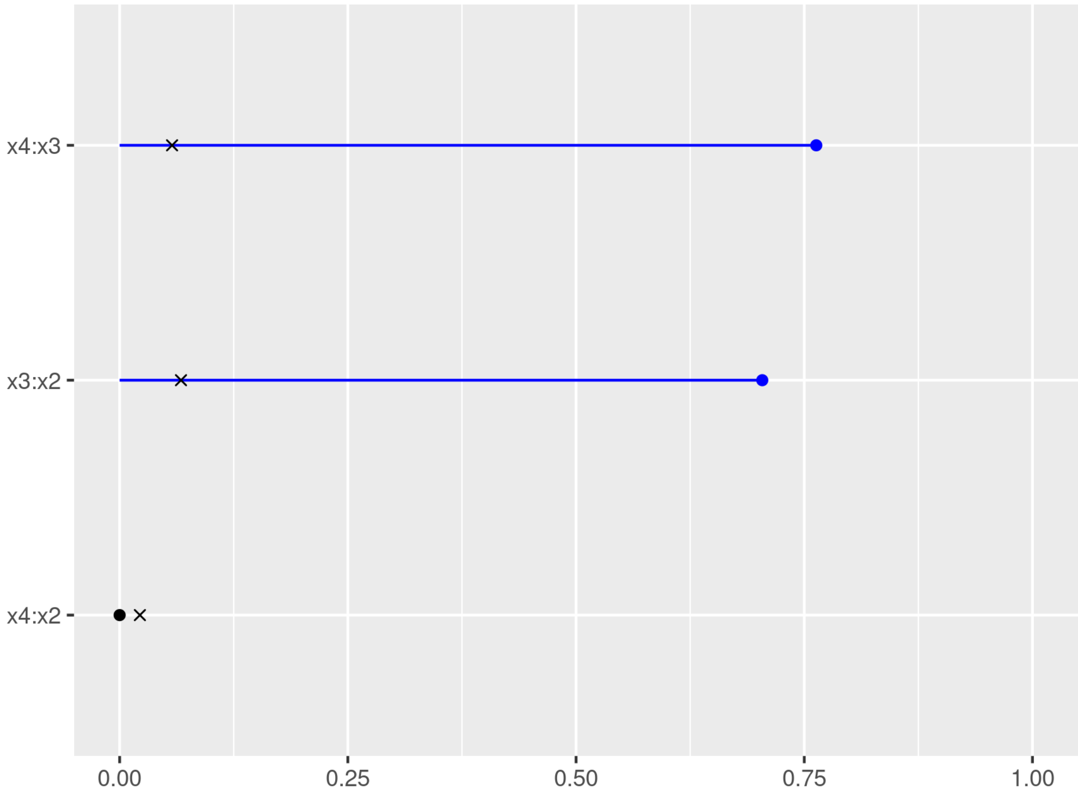


Figure 4.4: *Two-way interaction strength - Scenario 1*

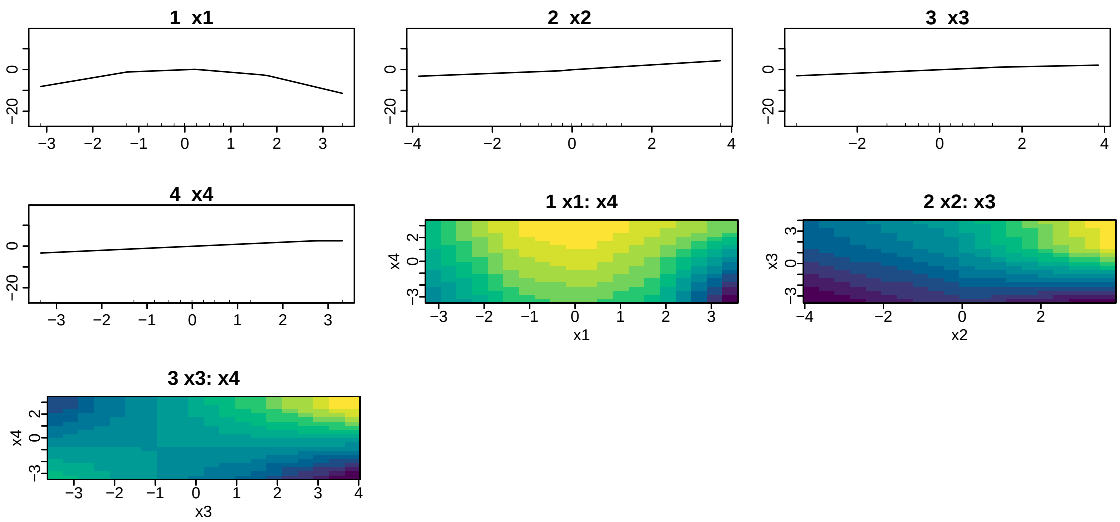


Figure 4.5: *Partial dependence plots of the final model - Scenario 2*

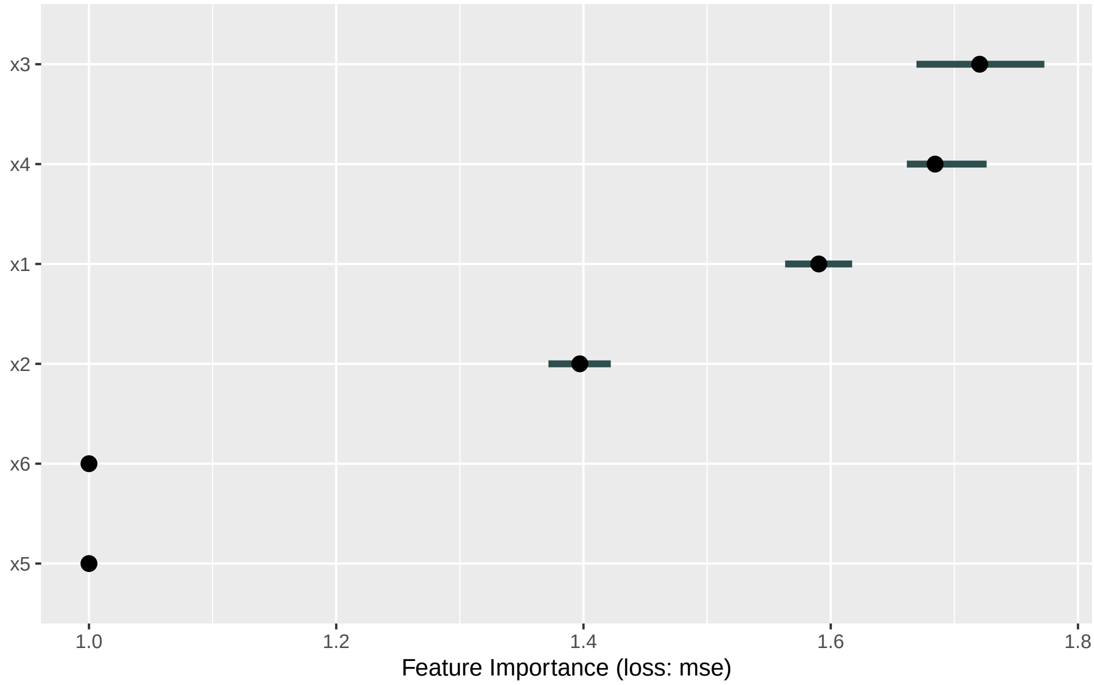


Figure 4.6: *Variable importance plot of the initial model - Scenario 2*

The variable importance plot, shown in Figure 4.6, correctly identifies the only four predictors. It shows that variables X_3 and X_4 are the strongest predictors, followed by X_1 then by X_2.

As in scenario 1, we calculate for each predictor the overall interaction strength. The results are in Figure 4.7. The H_j statistics show that X_2, X_3 and X_4 are the only variables that interact. That X_1 has an interaction strength of nearly zero (and is also smaller than the critical value) allows us to conclude that the $X_1 \cdot X_4$ term in the model is spurious. Finally, we calculate two-way interaction statistics for all three pairwise combinations of X_2, X_3 and X_4. As shown in Figure 4.8, we conclude that the only true interaction terms are $X_2 \cdot X_3$ and $X_3 \cdot X_4$.

4.4.3 Scenario 3: No interaction

In the last scenario, with β set to zero there is just one interaction term in the DGP, namely $X_3 \cdot X_4$. The initial MARS model fitted to all six variables, shown in Figure 4.9, shows similar main effects as in the previous two scenarios, but

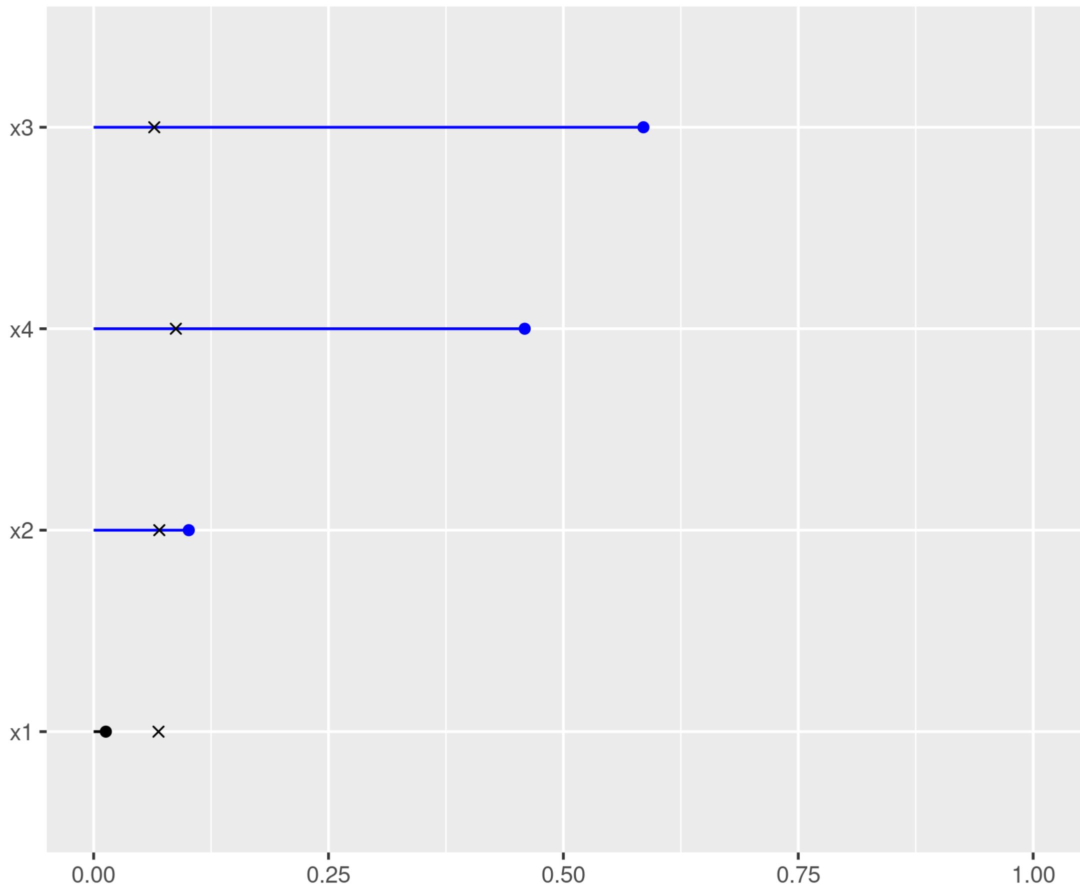


Figure 4.7: *Overall interaction strength of all variables - Scenario 2*

has four interaction terms: $X_1 \cdot X_2$, $X_1 \cdot X_4$, $X_2 \cdot X_3$, and $X_3 \cdot X_4$. Of the four, only $X_3 \cdot X_4$ is in the DGP.

We proceed to checking the variable importance. As Figure 4.10 shows, again the only predictors are X_1 to X_4. Refitting the model to only the four predictors produces a more parsimonious model. As shown in Figure 4.11, now the fitted model has two instead of four interaction terms, namely $X_1 \cdot X_4$ and $X_3 \cdot X_4$.

To identify the true interaction term, we calculate the overall H_j statistics. The results, Figure 4.12, shows that only X_3 and X_4 are involved in an interaction. We can therefore conclude that the $X_1 \cdot X_4$ pair in the fitted model is spurious. We can also deduce the only possible interaction has to be between X_3 and X_4. Indeed, the two-way H_{jk} statistics, in Figure 4.13, show that this is the case.

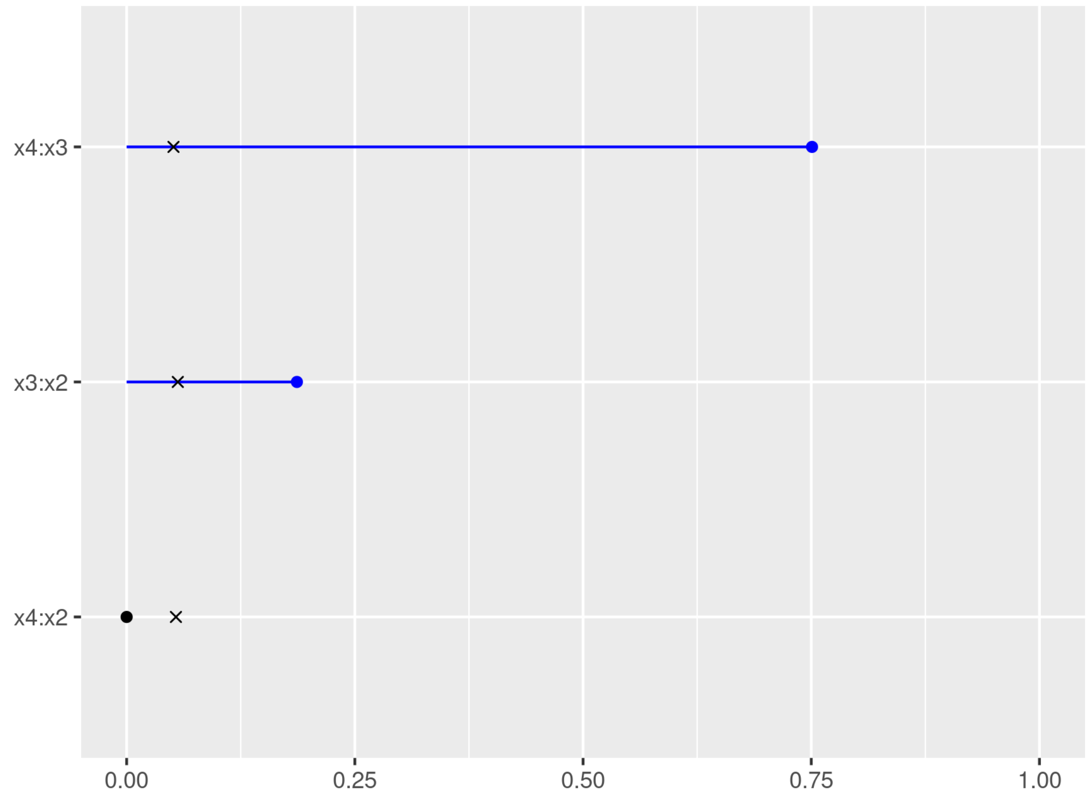


Figure 4.8: *Two-way interaction strength - Scenario 2*

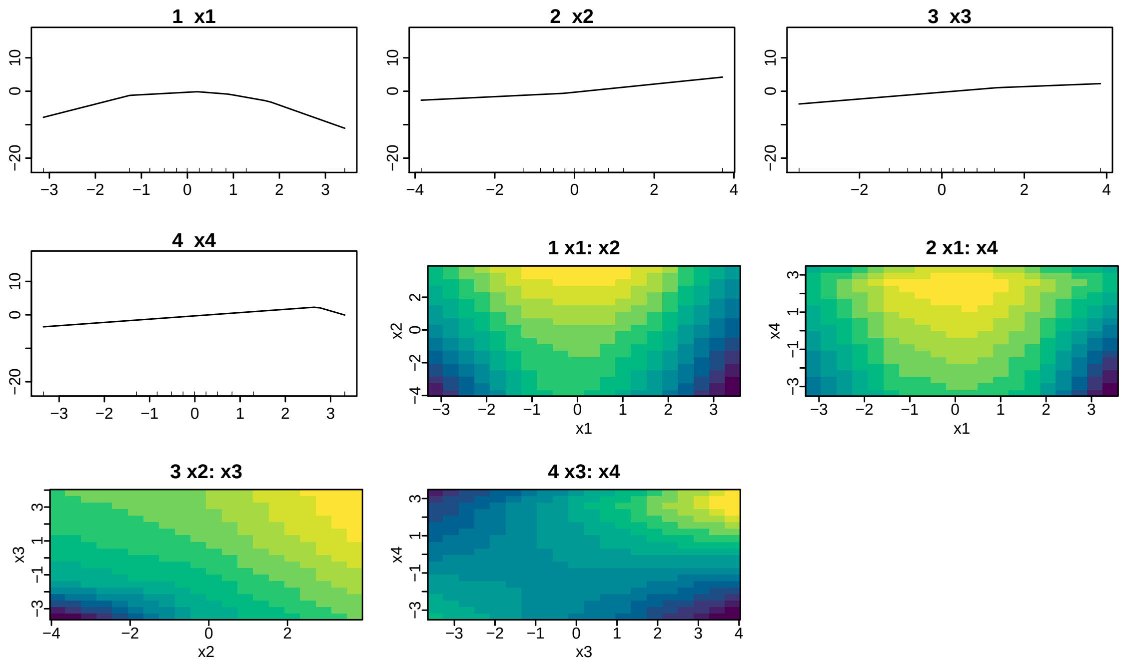


Figure 4.9: *Partial dependence plots of the initial model- Scenario 3*

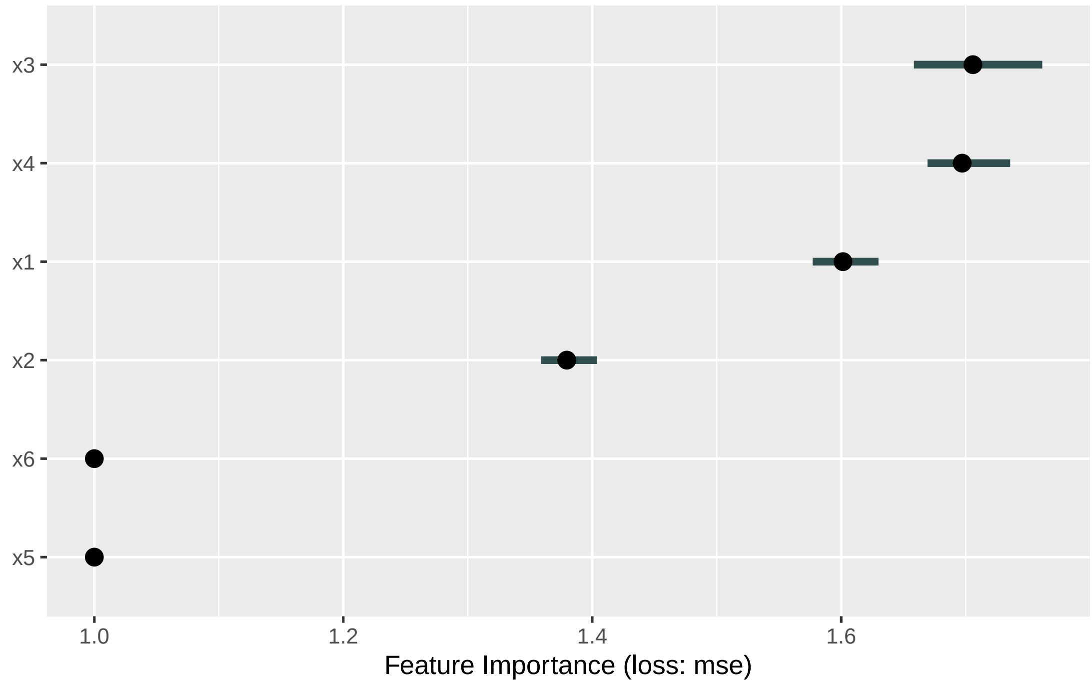


Figure 4.10: *Variable importance of the initial model - Scenario 3*

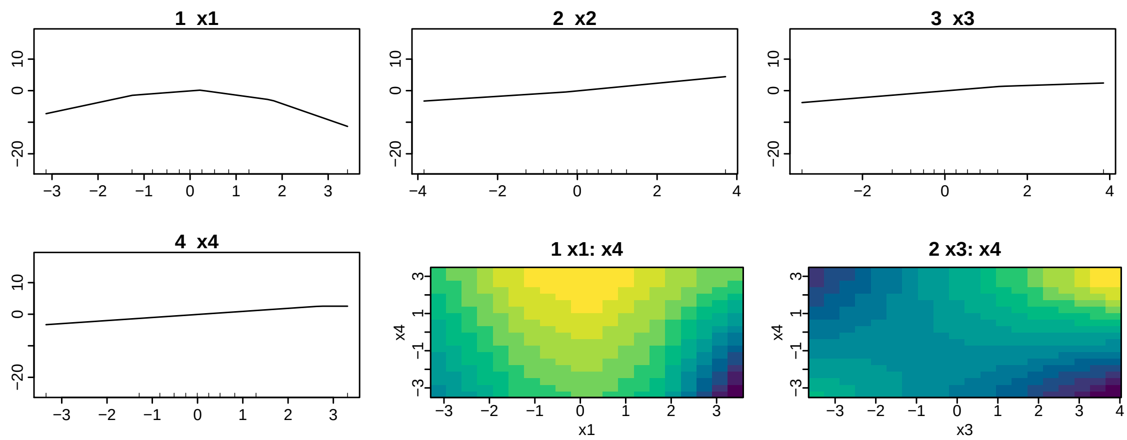


Figure 4.11: *Partial dependence plots of the final model- Scenario 3*

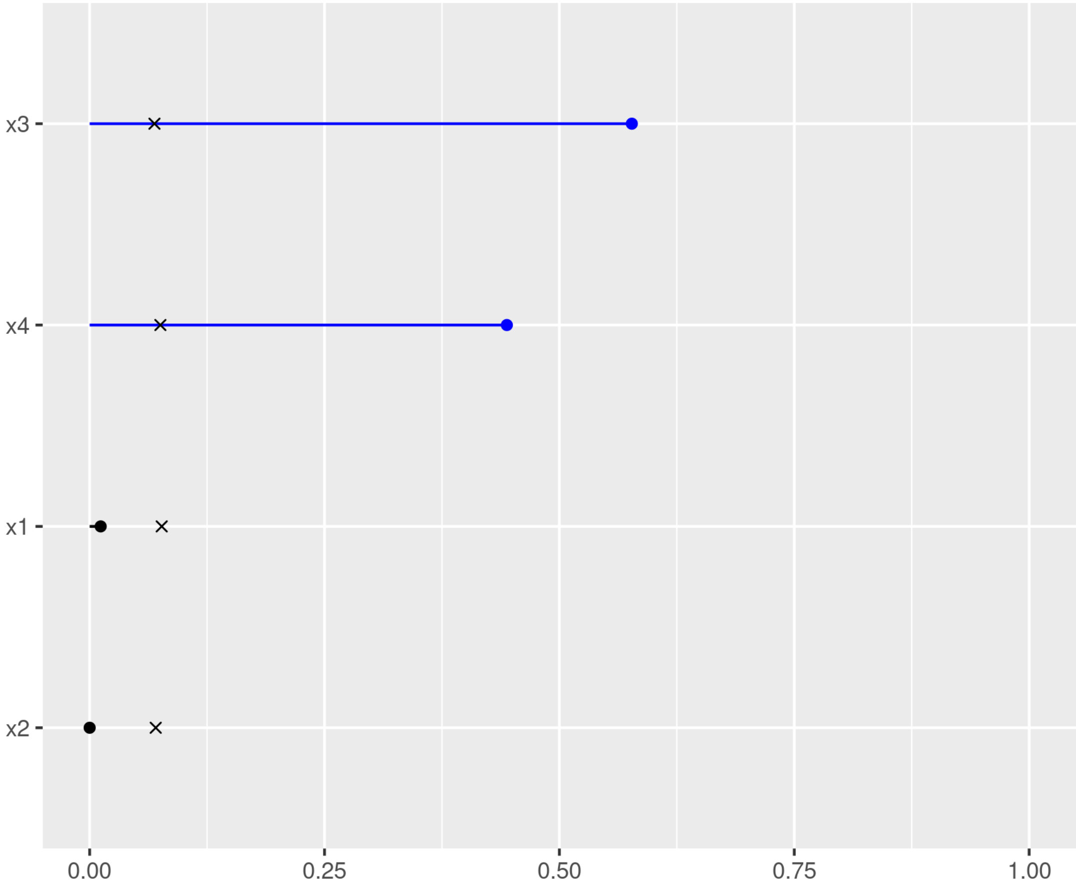


Figure 4.12: *Overall interaction strength of all variables - Scenario 3*

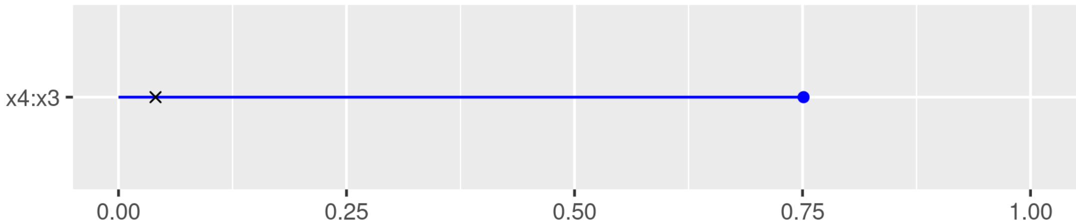


Figure 4.13: *Two-way interaction strength - Scenario 3*

4.5 Data

The data used in the analysis is sourced from the Household, Income and Labour Dynamics in Australia (HILDA) survey. In this section we describe how the analysis sample is selected and provide details on the key variables – the outcome variable, log hourly wages, and the key explanatory variables, namely cognitive skills, the Big Five personality traits and locus of control – and other explanatory variables.

4.5.1 Sample selection

The analysis sample is taken from wave 13 of the HILDA survey. The majority of the interviews took place in the second half of 2013. We focus on the prime working age population and select only those who were aged between 25 and 54. Of the 8,809 respondents in that age group, 6,136 (or 69.7 percent) were employed, usually worked 1 hour or more per week, and earned more than 1 dollar weekly from all jobs. The other 2,673 respondents worked/earned less, were unemployed, or not in the labour force. They are excluded from the analysis. Among the 6,136 individuals, 737 (or 12.0 percent) did not complete the Self-Completion Questionnaire (SCQ), which contains questions on personality, and are therefore also excluded from the analysis. The non-response rate of 12.0 percent for the SCQ is similar to that for the overall sample in wave 13, at 11.9 percent (Summerfield et al, 2016). Further excluding the 737 without the SCQ, our sample comprises 5,399 individuals, with 2,751 males (51.0 percent) and 2,648 females (49.0 percent).

4.5.2 Log hourly wages and the explanatory variables

Log hourly wages The outcome variable is log hourly wages. It is calculated as the weekly gross wages and salary from all jobs, divided by the number of hours per week usually worked in all jobs.

Cognitive skills Cognitive skills are sourced from wave 12 (2012) as they were not asked in wave 13. These are matched using the cross-wave person identifier. There are three cognitive ability tests in wave 12 of the HILDA survey: the backwards digit span (BDS), the symbol-digits modalities (SDM) test, and a shortened version of the National American Reading Test (NART). Given the three tests measure different aspects of cognitive abilities, the three test scores are included separately (i.e. three explanatory variables) in the regressions.

Big Five personality traits The personality inventory in HILDA is derived from the Trait Descriptive Adjectives (Saucier, 1994a), which in turn are based on Goldberg's Big-Five Factor Markers (Goldberg, 1992b). In total there are 36 trait descriptive adjectives. These questions are included in the Self Completion Questionnaire. For each question, the respondents self-assessed the degree to which the adjective described them on a 7-point scale, from 1 "Does not describe me at all" to 7 "Describes me very well". About 94% of the respondents replied to all 36 questions. A 5-factor model is fitted on the 36 personality items. The five factor scores are used as measures of the Big Five.

Locus of control Locus of control refers to how much control a person believes they have over events in their lives. Individuals with a strong internal locus of control believe that their life experiences are driven by their own actions and decisions, whereas individuals with an external locus of control believe that external factors determine what happen to them. Before wave 13 of the HILDA survey, questions relating to locus of control were asked in waves 3, 4, 7 and 11 (i.e. 2003, 2004, 2007 and 2011 respectively). We matched the wave 11 responses to our sample. There were 7 questions and we factor analysed them to obtain a single factor that increases with internal locus of control.

Other explanatory variables The other explanatory variables include the peron's age, gender, the highest level of education they achieved, the state/territory they reside in, their father's and mother's occupational status, and various attributes of their job, including employment status (part-time or full-time), whether a casual or permanent worker, occupations, industries, tenure in the same occupation, tenure in the same industry, firm size, and sector of employer.

4.6 Results

We will present the results first for females and then for males. For females, we find no interaction effects on earnings. For males, we find seven two-way interaction terms among five variables, namely Big 5 conscientiousness, occupation, industry, whether a casual or permanent worker, and full-time or part-time work. Of these, the three strongest two-way interactions as measured by the H_{jk} statistics are between conscientiousness and casual/permanent, between occupation and industry, and between industry and conscientiousness.

Interactions found in the literature, such as those between social and cognitive skills in the US and emotional stability and cognitive skills in Poland, are not found in the Australian labour market. However, our findings regarding the main effects of cognitive and non-cognitive skills are generally consistent with the literature.

Please note that while the outcome variable is the log of the hourly wages, in the partial dependence plots below the predicted wages have been exponentiated to aid interpretation.

4.6.1 Earnings regressions for females

First, we fit an initial MARS model using all explanatory variables. In this model (results not shown), any two-way interaction is allowed. The relative importance of the variables is ranked according to the variable importance measure, as shown in Figure 4.14. The most important predictors of earnings are occupation, highest level of education achieved, industry, and firm size. Of the three measures of

Figure 4.14: *Variable importance - females*

cognitive skills, the National American Reading Test score best predicts earnings. The other two measures, backward digit span and symbol-digit modalities, are among the weakest predictors. Of the non-cognitive skills, the strongest predictor is locus of control, followed by agreeableness, conscientiousness, and emotional stability. The other two Big 5 dimensions, extraversion and openness to experience, are not chosen in the initial model. After removing these and other unused variables, we refit the MARS model. Its partial dependence plots are shown in Figure 4.15. It has eight main effects and 31 interaction terms. Given the many interaction terms, the estimated MARS model is admittedly difficult to interpret even with the aid of partial dependence plots. But this also demonstrates the practical need to first remove any potentially spurious interactions.

To reduce the number of interaction terms, we allow only certain variables to interact. We measure the overall interaction strength of each variable by calculating

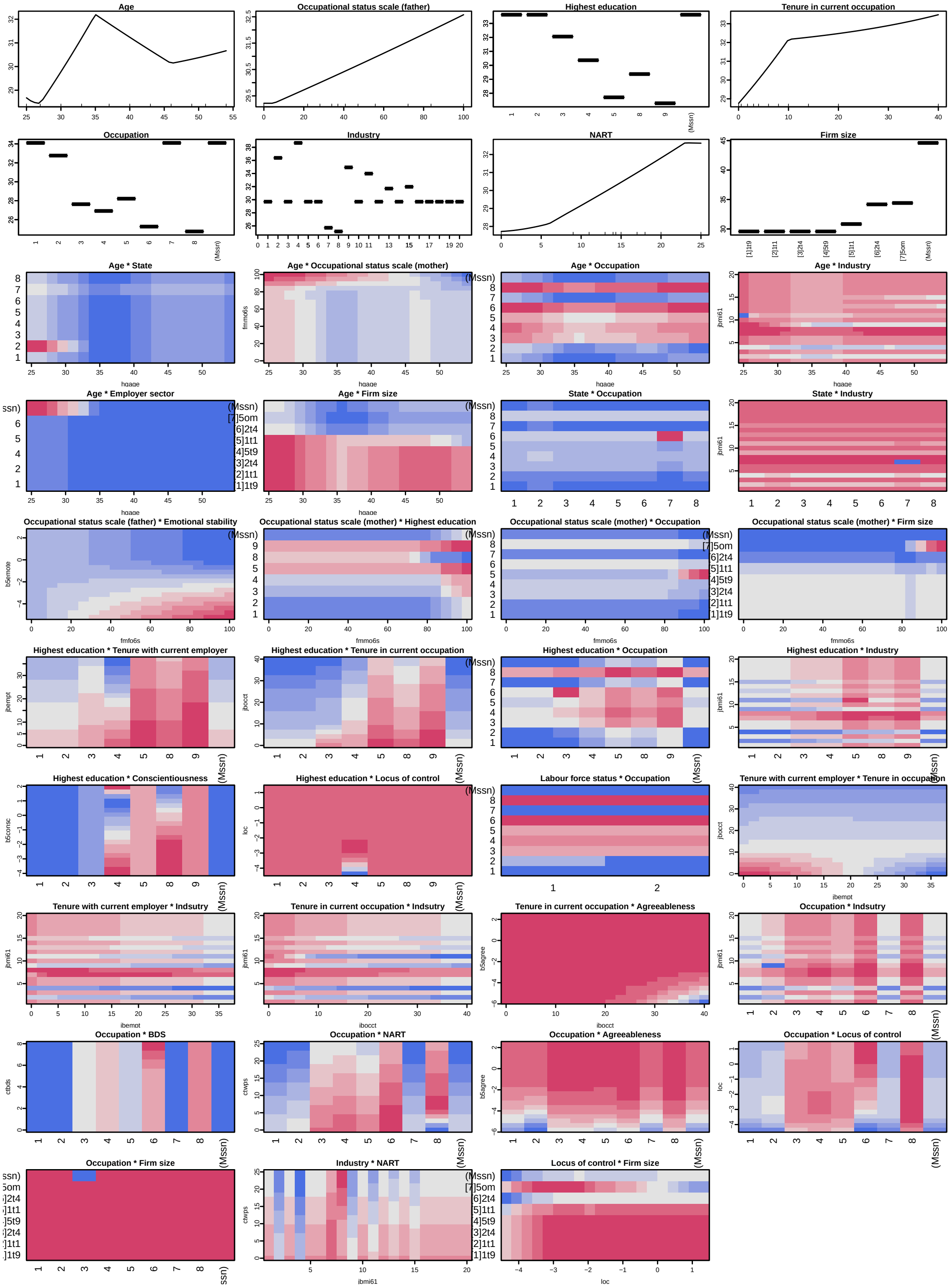


Figure 4.15: *Partial dependence plots of the first model - females*

the H_j statistic. As Figure 4.16 shows, occupation, education and industry have the largest H_j statistics, suggesting that they interact with other variables. However, their magnitude of around 0.25 appears to suggest they are not strong interactions. As some of these could be spurious, it is important to compare each H_j statistic to its magnitude under the null hypothesis of no interaction involving X_j. Two variables, highest education level achieved and internal locus of control, are the only variables with an H_j statistic greater than the 95th percentile of the null distribution, which we obtain from 500 bootstraps. We refit the MARS model, allowing only education and locus of control to interact. Figure 4.17 shows the resulting model.

Next, we test the two-way interaction between education and locus of control. Again the null distribution of the H_{jk} statistic is obtained using 500 bootstraps. According to Figure 4.18, the H_{jk} statistic is less than the 95 percentile, suggesting that the interaction may have been spurious. Given that the only interaction term is not statistically significant, we fit our final model with no interaction terms at all. This is shown in Figure 4.19.

To summarise, for female workers, we do not find any interaction effects on wages. The only non-cognitive skill with a statistically significant H_j statistic is locus of control. Regarding the other non-cognitive skills, it is possible that a type II error has been made. However, in the second stage model, of the 31 interaction terms none are between non-cognitive and cognitive skills. In other words, the interactions found in the literature are not strong enough to be even considered in our model. In fact, the magnitude of the interaction effects found in the literature is relatively modest. For example, Deming (2017) found that the interaction between cognitive skills and social skills on (log) hourly wages had a coefficient of 0.019. With both explanatory variables standardised, this translates to a magnitude of just under 2 percent increase in wages for a one standard deviation increase. Another

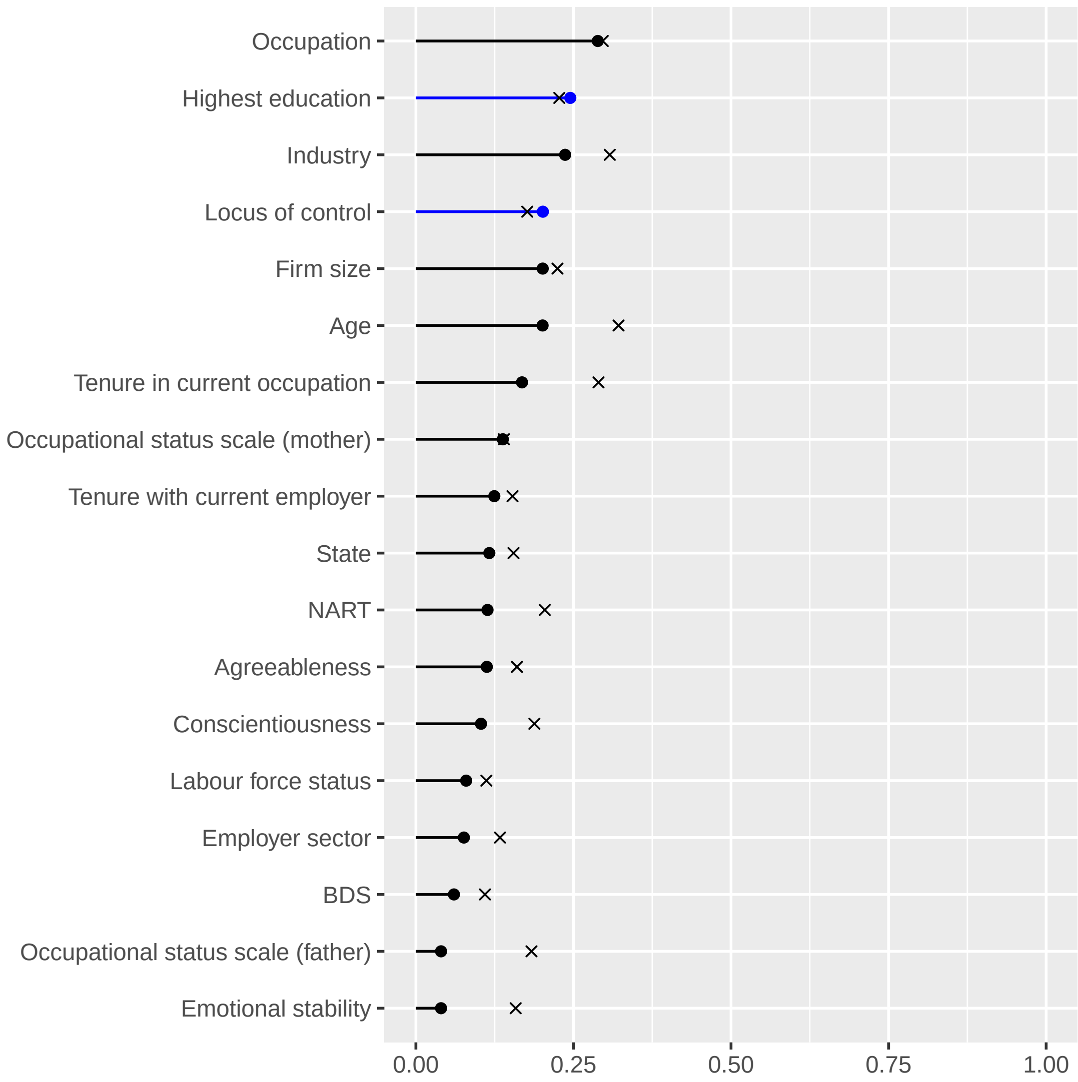


Figure 4.16: *Overall interaction strength of all variables - females*

reason we do not find interaction between social and cognitive skills is that our measure of sociability is different. In Deming (2017) and Weinberger (2014), social skill is measured by participation in sports clubs and other leadership roles but in the absence of similar variables in the HILDA survey and following Palczyńska (2021), social skill is proxied by extraversion instead. The interaction between emotional stability and cognitive skills reported by Palczyńska (2021) is not found either. This could be due to differences between the Polish and Australian labour markets.

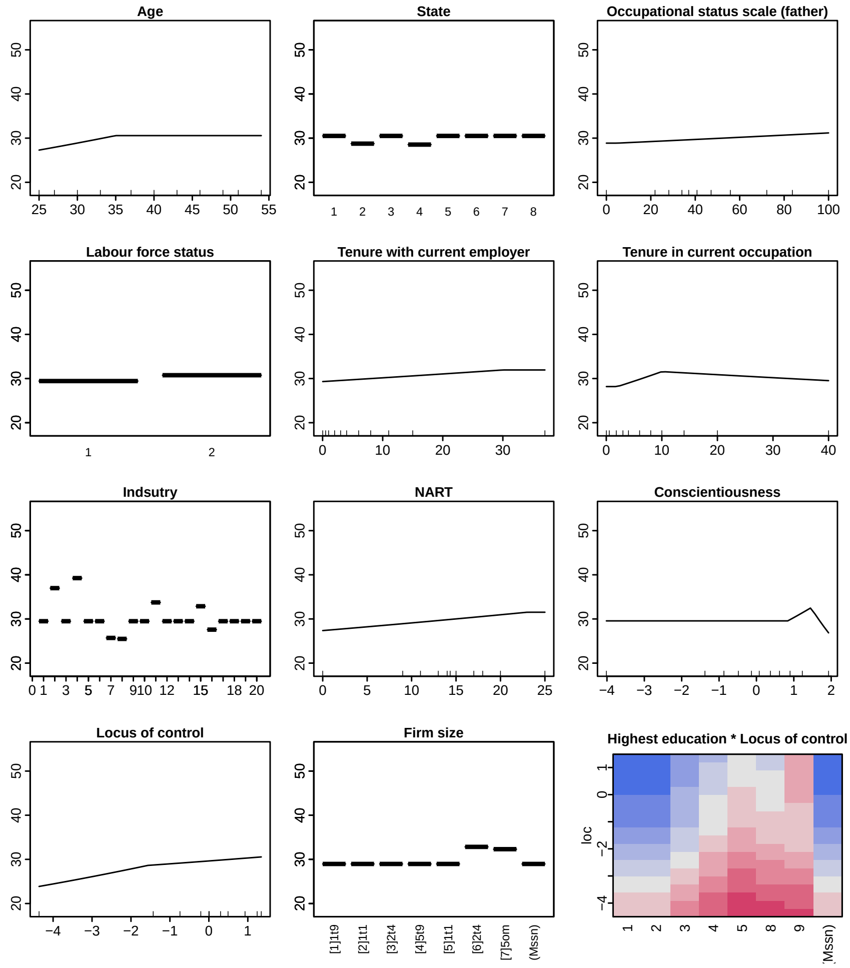


Figure 4.17: *Partial dependence plots of the second model - females*

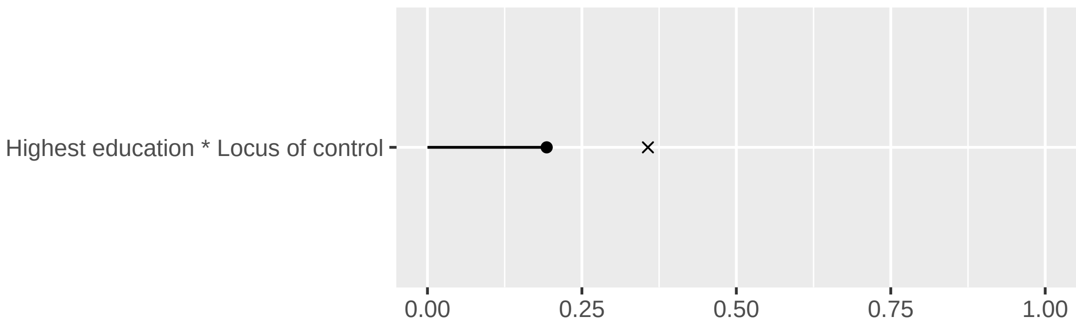


Figure 4.18: *Two-way interaction strength - females*

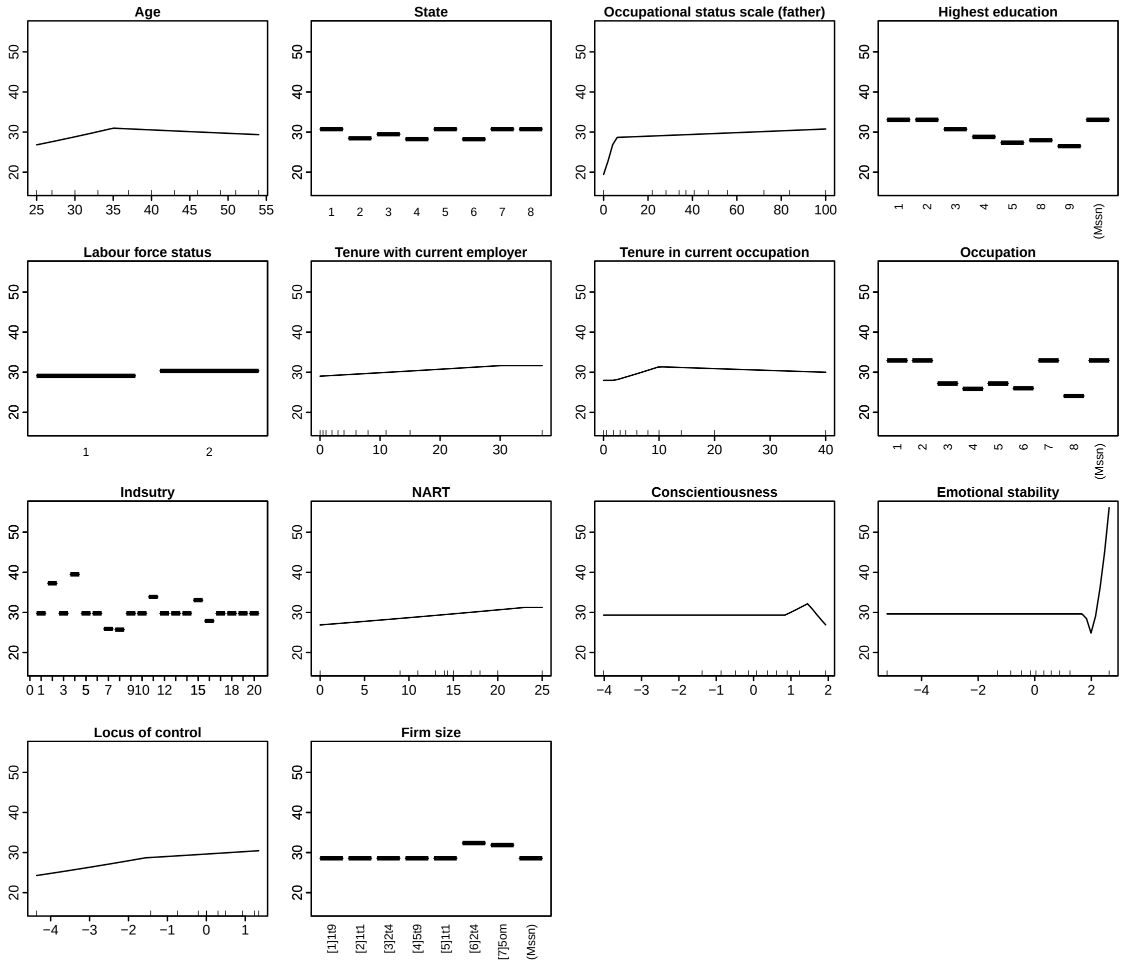


Figure 4.19: *Partial dependence plots of the final model- females*

Our findings regarding the main effects of both non-cognitive and cognitive skills are broadly in line with the literature. Of the Big Five personality traits, conscientiousness, emotional stability and agreeableness have been found to predict wages. In our final model for females, both conscientiousness and emotional stability (but not agreeableness) enter as main effects. However, in our estimated model, wages do not differ for female workers with different levels of each trait. Wages do increase with cognitive skills, as measured by the NART score, and internal locus of control.

The main effects of other explanatory variables appear consistent with general earnings statistics recorded by the Australian Bureau of Statistics. For example, we find that workers in Tasmania have lower hourly wages. This agrees with the median weekly earnings reported for each state. Workers in larger firms tend to earn more. The hourly wages are also higher for certain occupations (e.g., managers; professionals; and machinery operators and drivers) and industries (e.g., mining; electricity, gas, water and waste services; and financial and insurance services).

4.6.2 Earnings regressions for males

We again start with an initial MARS model with all 23 explanatory variables. Figure 4.21 ranks the variable by predictiveness. The top predictors of earnings for men are the same as for women (but in a slightly different order): industry, occupation, firm size, and education. Among cognitive skills, the National American Reading Test score best predicts earnings. Of the non-cognitive skills, the strongest predictor is conscientiousness, followed by locus of control. Emotional stability, extraversion, and agreeableness are relatively weak predictors, with openness to experience unused in the initial model. Figure 4.20 shows the PDP after removing the two unused variables. Like the model for females, it has a large number (38) of interaction terms.

Calculating the H_j statistics and comparing them to the 95th percentile of the respective null distribution, we see that five variables have a statistically significant H_j statistic: occupation, industry, conscientiousness, whether a casual or permanent worker, and employment status of full-time or part-time work. This is shown in Figure 4.22. In the next iteration of the model, we therefore allow only these five variables to interact. As Figure 4.23 shows, this results in eight interaction terms. Of the eight interaction terms, seven are statistically significant, as can be seen in Figure 4.24. Retaining only these interaction terms gives the final model for males, shown in Figure 4.25.

Of the seven interaction effects, of most interest to us are the two that involves conscientiousness. Generally, workers with a higher level of self-rated conscientiousness have a higher hourly wage. The interaction between conscientiousness and casual/permanent worker status suggests that for casual workers conscientiousness matters even more than for permanent workers. For example, for permanent workers a one standard deviation increase from the mean in Big Five conscientiousness (i.e. from 0 to 1) is associated with an increase of only $0.7/hour (from $34.0/hour), but for casual workers the increase is $1.7/hour from $35.2/hour. This is an increase by nearly 5 percent.

The returns to conscientiousness also differ by industry. For most industries, wages increase very slightly with conscientiousness. There are more wage variations with conscientiousness for those in "[4] electricity, gas, water and waste services", "[5] construction" and "[7] retail trade". For both [4] and [5], wages increase with conscientiousness but there are differences. For those in "[4] electricity, gas, water and waste services", having low conscientiousness is penalised more heavily. Those whose conscientiousness is one standard deviation below average earns $4.7/hour less than the industry average of $41.0/hour. In contrast, for those in "[5] construction", high conscientiousness is rewarded more heavily, earning $5.1/hour more than the industry average $37.4/hour. Lastly, for those in "[7] retail trade" conscientiousness is penalised slightly.

Of the other five interaction effects, two involve industry, another two involve occupation, and one is between industry and occupation. We highlight only a handful of the interaction effects. In the mining industry (Industry=2), part-time workers (Labour force status=2) have a higher hourly wage than those working full-time. For technicians and trades workers (Occupation=3), part-timers also have a higher hourly wage. Wages also differ between casual and permanent workers for certain occupations. In general permanent workers have a higher hourly wage, e.g. “[1] managers” and “[8] labourers”, but for “[2] professionals” casual workers earn more per hour.

Our procedure avoids testing all possible interactions indiscriminately. However, given the many hypothesis tests still conducted - 21 tests for overall interactions and 8 tests for two-way interactions, the problem of multiple testing is a valid concern. In particular, one or more of the seven two-way interactions identified could be false positives. Methods such as the Bonferroni correction to control the type I error rate are applicable here. However, no such corrections have been applied partly because the goal of this study is primarily to find out whether the interaction effects on wages reported in the literature are also present in the Australian labour market. Correcting for multiple testing will not change our finding that they are not. Regarding the seven two-way interactions we did find for males, their validity is perhaps better tested by attempts to replicate them using more recent labour market data than by finding the right trade-off between type I and type II errors.

In terms of the main effects of the cognitive and non-cognitive skills, we find that hourly wages increase with cognitive skills as measured by the SDM ant NART scores. Wages also increase with the Big Five conscientiousness. Those with an above-average level of internal locus of control also tend to have higher wages. Workers who self-rated to be emotionally stable have a lower hourly wage. This is unexpected as the literature generally finds a positive relationship. Lastly, we notice a large increase in wages as Big Five agreeableness goes below -3, but this may not be reliable given the 5th percentile is -1.7 and so very few workers have scores below -3.

The main effects of the other explanatory variables are generally sensible. For example, hourly wages are higher for workers in Western Australia but lower for those in Tasmania. Wages increase with firm size and those employed in not-for-profit organisations have a slightly lower hourly wage.

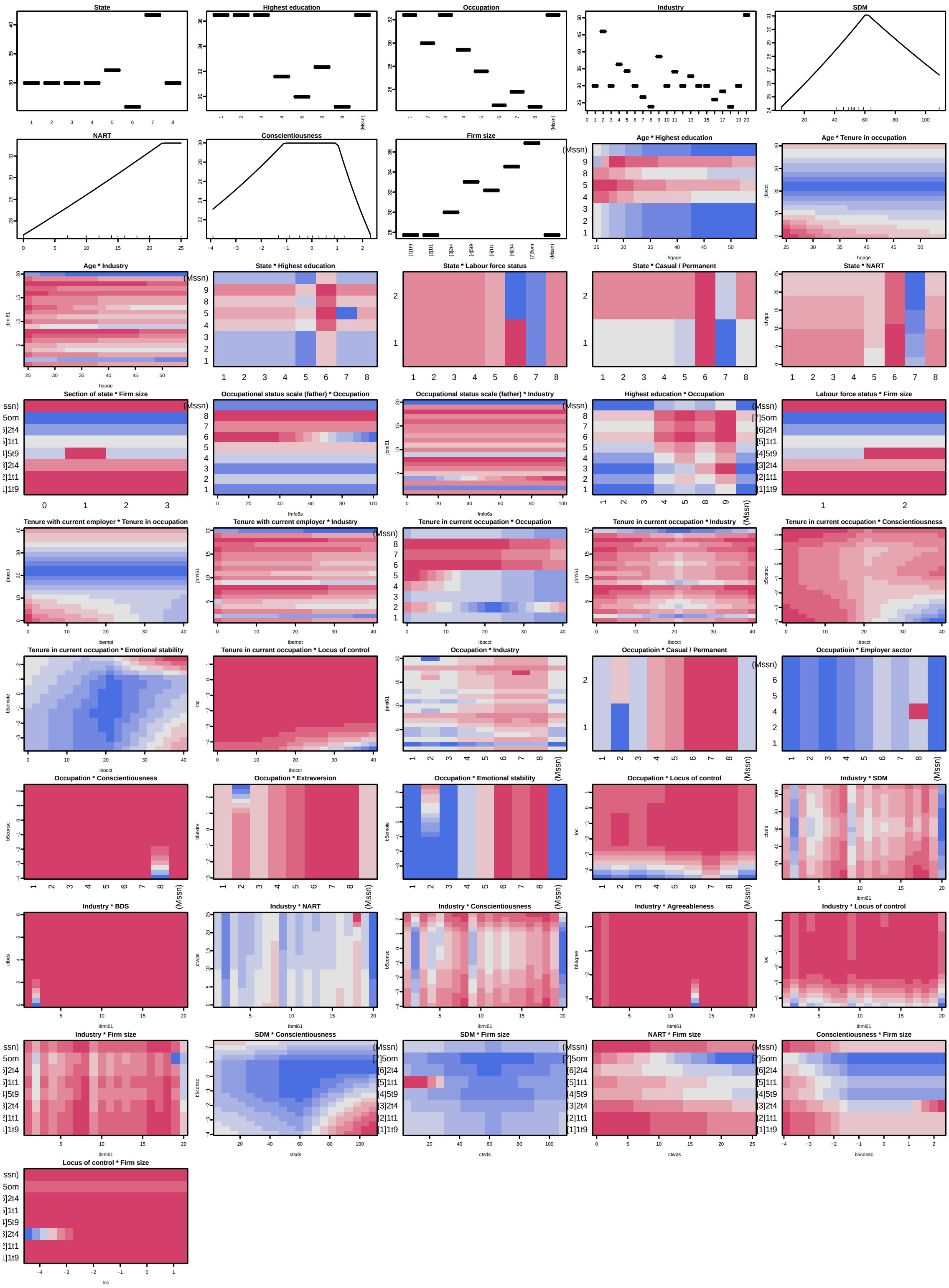


Figure 4.20: *Partial dependence plots of the first model- males*

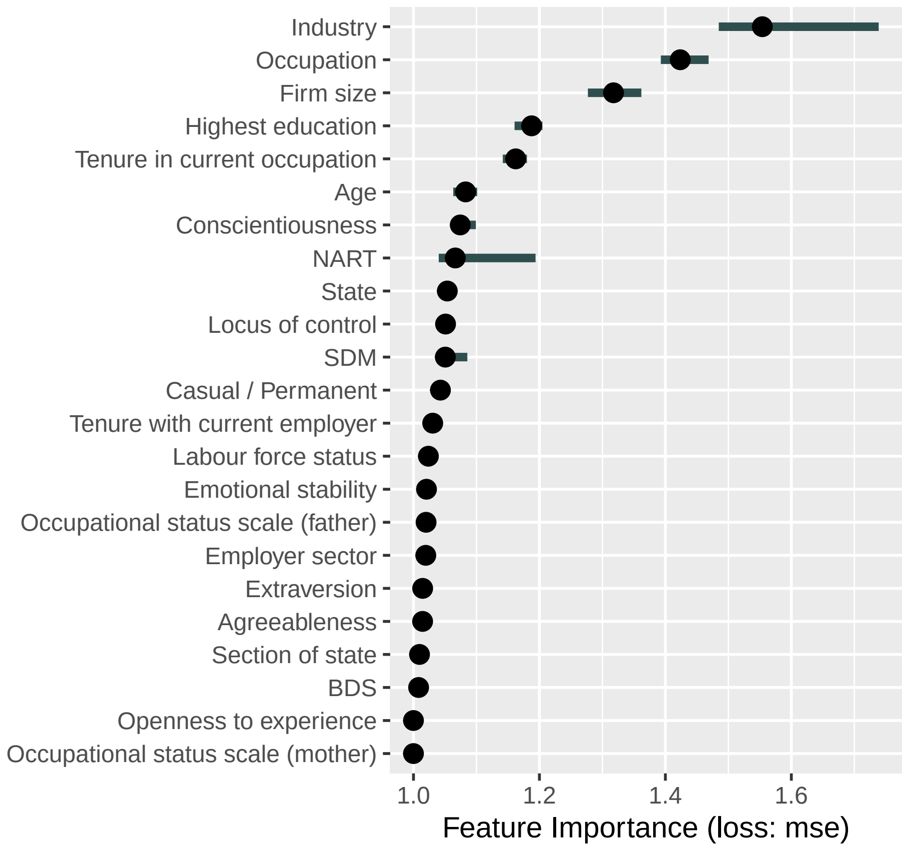


Figure 4.21: *Variable importance - males*

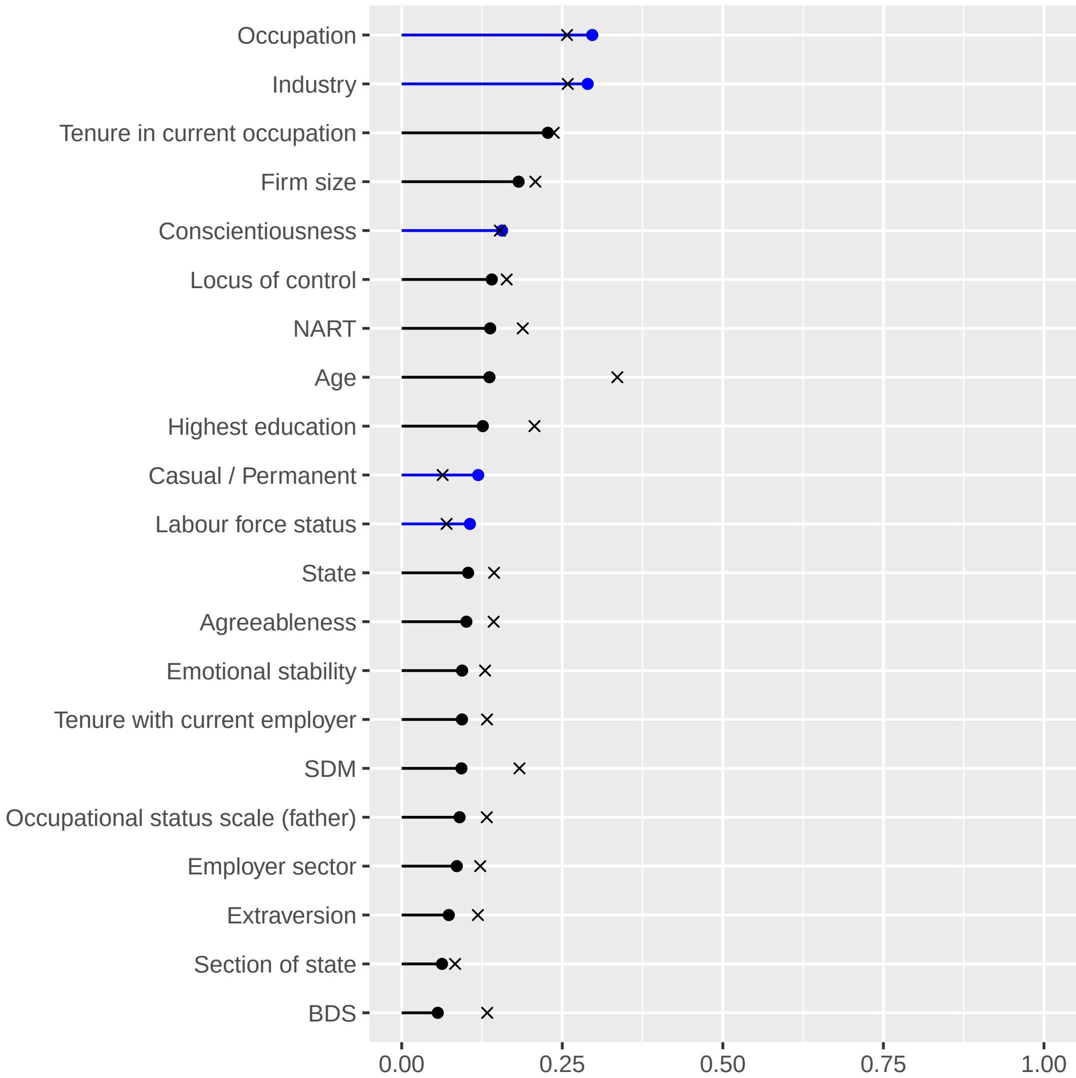


Figure 4.22: *Overall interaction strength of all variables - males*

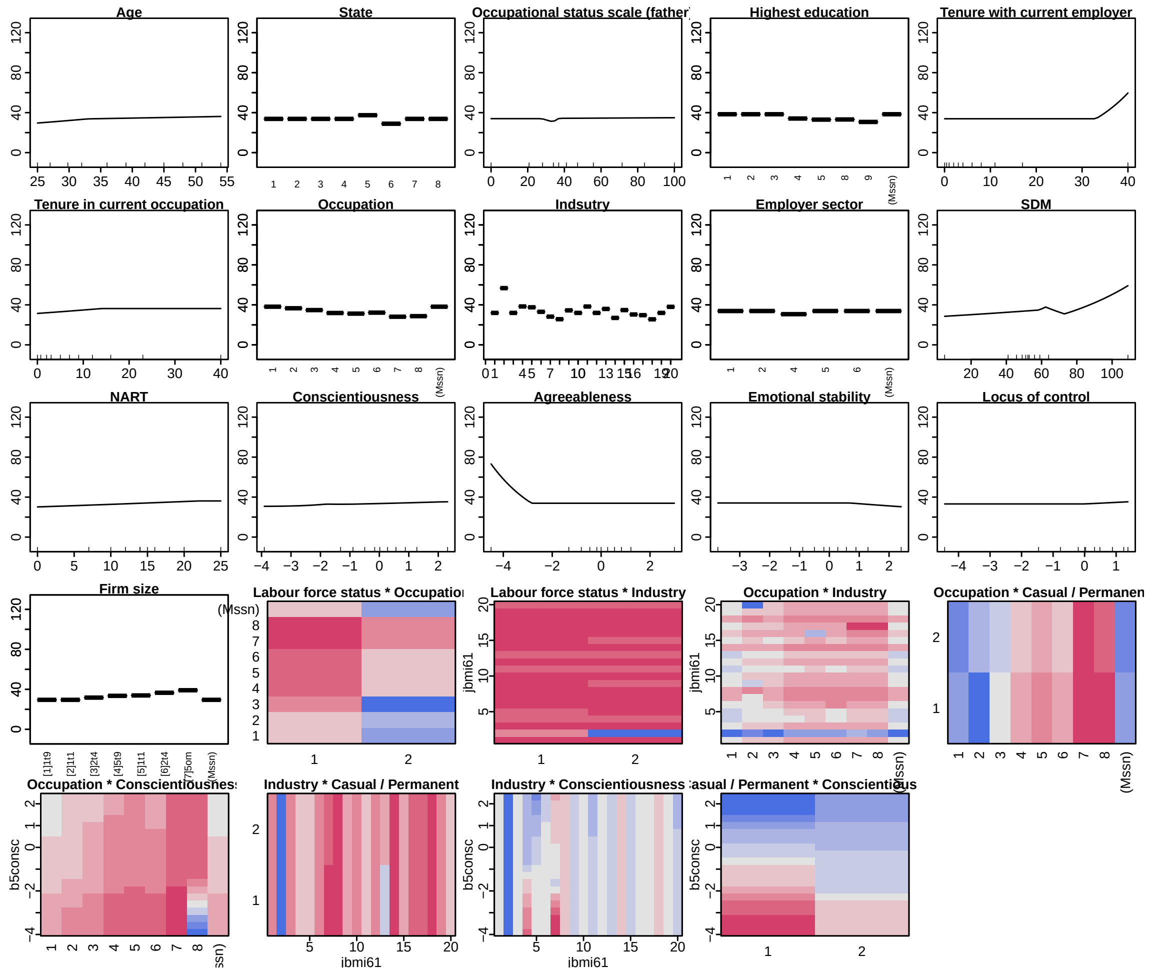


Figure 4.23: *Partial dependence plots of the second model- males*

Figure 4.24: *Two-way interaction strength - males*

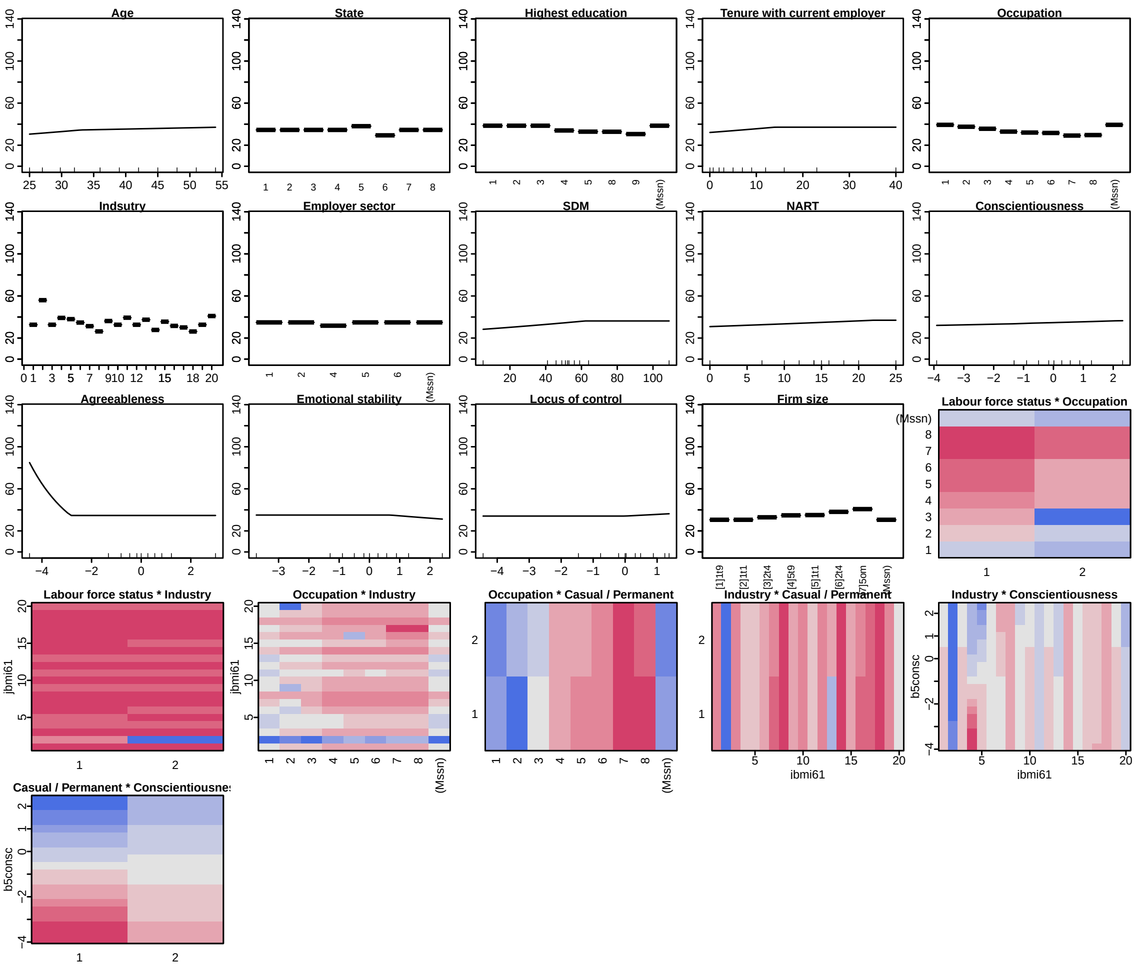


Figure 4.25: *Partial dependence plots of the final model - males*

4.7 Conclusion

In this chapter, we explore the interaction effects of cognitive and non-cognitive skills on wages in Australia. Previous studies have focused on specific two-way interactions, for example between cognitive and social skills (Deming, 2017; Weinberger, 2014) and between cognitive skills and emotional stability (Palczyńska, 2021). In an attempt to evaluate these interactions analysed before and discover new ones, we combine the use of interaction statistics with a flexible regression model to estimate the wage regression separately for male and female workers. Unlike Weinberger (2014) and Deming (2017), we find no interaction effects between cognitive and social skills. The interaction between cognitive skills and emotional stability, reported in Palczyńska (2021), is also not found in our data. While we find no interaction effects on wages for females, for males we find seven two-way interactions, including interactions between conscientiousness and whether a casual/permanent worker, and between conscientiousness and industry. That we find different interactions than Weinberger (2014), Deming (2017) and Palczyńska (2021) could be due to the use of different methods or could reflect differences in the respective labour markets. To answer this question, future research could adopt a similar approach to ours and more systematically study the interaction effects on earnings across different labour markets.

Chapter 5

Conclusion

This chapter provides a summary of the key findings and their implications. Also discussed are limitations of the analysis and potential avenues for future research.

5.1 Key findings and their implications

The three research questions this dissertation aims to answer are:

1. What effects does university education have on non-cognitive skills?
2. Why have researchers found personality types despite the abundance of research replicating the Big Five personality traits?
3. Are non-cognitive skills complimentary with any cognitive skills in the labour market?

Regarding the first question, we find that university education fosters sociability and the tendency to cooperate. Importantly, university students from low socioeconomic backgrounds experienced the most growth in their level of agreeableness. Our findings suggest that university education fosters agreeableness likely by introducing students to new peer groups and extracurricular activities. These findings may have important implications for policies that aim to address disparity in skills between adolescents from different socioeconomic backgrounds. While the return on investment may be lower compared to that of early childhood interventions, it is an opportunity to help adolescents lacking in valuable non-cognitive skills before they enter the labour market.

In chapter 3, we show that recent studies replicating personality types can be reconciled with earlier research that has replicated the Big Five personality traits. First, we show that many key statistics that are routinely used to guide factor extraction, show internal consistency, and demonstrate factor replicability only depend on the data through the sample covariance matrix and do not imply the absence of clusters in the data. We then compare the relative model fit of a multitude of factor analysis models and Gaussian mixture models on two personality data sets and find that a mixture model provides a better relative fit. However, even the best mixture model could not capture all nuances in the personality data. This suggests that alternative latent variable models may be needed.

Regarding the last question, we could not replicate in the Australian labour market the interaction effects between cognitive and social skills, as reported by Deming (2017) and Weinberger (2014). Instead, we find different interaction effects of non-cognitive skills on wages for male workers, namely between conscientiousness and casual/permanent worker status, and between conscientiousness and industry. In other words, no interaction effects between cognitive and non-cognitive skills are found in our analysis sample. For female workers, we do not find any interaction effects on wages.

5.2 Limitations and future research

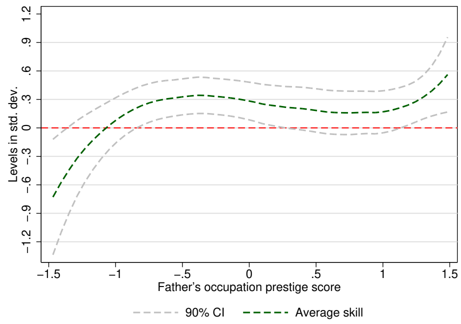
In chapter 3 we compare different factor analysis and Gaussian mixture models on two personality data sets. While the findings are consistent on both data sets, there are many more household surveys and other publicly available data sets on the Big Five. Given the extensive replication studies on the Big Five, future research should extend our analysis to these other data sets. Directly comparing the relative model fit of factor analysis and Gaussian mixture models on Big Five personality data most commonly used in economic research would help inform the debate on whether a variable-centric or person-centric perspective is empirically better. More importantly, our findings suggest neither factor analysis nor Gaussian mixture models can fit the data very well. Given that factor analysis and Gaussian mixture models are the two classes of models most used to analyse personality data, the alternative latent variable model known as mixtures of factor analysers (McLachlan, Peel, and Bean, 2003; Ghahramani and Hinton, 1996) appears to be a natural extension that future research should evaluate. Mixtures of factor analysers postulate that the data are generated by a finite mixture of factor analysis models, thereby allowing each cluster to have a different factor analysis model.

While the use of MARS and interaction test statistics can more effectively identify and test for interaction effects on wages, bootstrapping the null distribution of the interaction test statistics requires significant amount of computation. The bootstrap is also specific to the particular regression model used. However, the method by Friedman and Popescu (2008) is to our knowledge the only formal hypothesis test for interactions. An alternative that could help reveal interaction effects in a regression model is Goldstein et al. (2015). The proposed method is based on individual conditional expectation plots that could help visualize heterogeneous relationships between the dependent and independent variables. So, instead of hypothesis testing, the identification of interactions is based on visualization of the regression model. This would be helpful if the goal is to uncover strong and obvious interaction effects in the model.

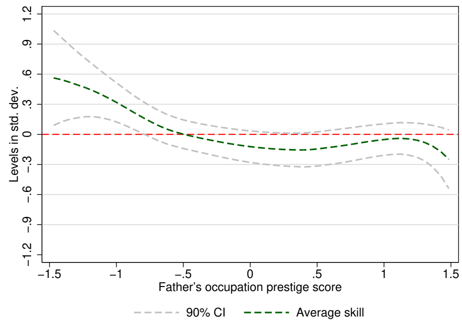
We find different interaction effects than Deming (2017) and Weinberger (2014). This could be due to differences in the different labour market analysed or due to the different empirical method being used. Future research could extend our findings by applying similar methods, i.e. MARS and interaction test statistics, to other labour markets. This could help answer the question whether the different interaction effects found on wages are because of local labour market dynamics.

Appendix A

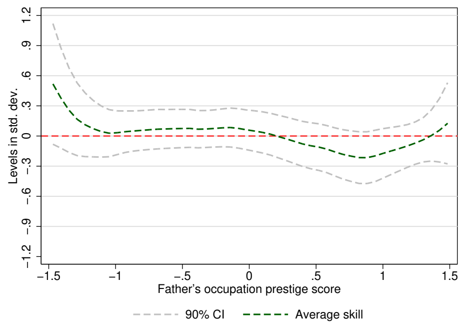
Supplementary material



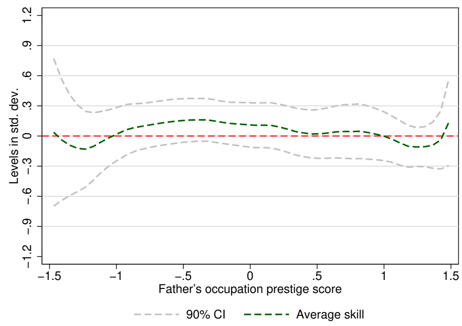
(a) Extraversion



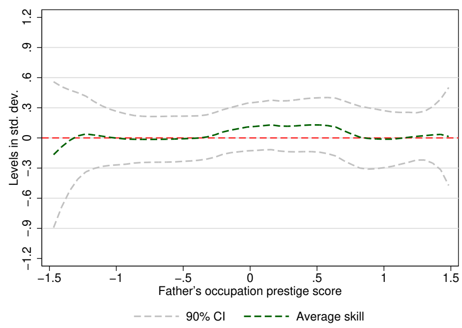
(b) Agreeableness



(c) Conscientiousness



(d) Emotional stability



(e) Openness to experience

Figure A.1: *Nonparametric estimates of the heterogenous effect of university education by father's occupational prestige score (90% confidence intervals). Sample includes all youth who went to university at some point in time (N=192). Dropped from the sample are the bottom and top 1% of Father's occupational prestige score distribution (10 observations on each side). Note: To estimate the interaction effect of university graduation and FOPS we followed the same method as in Schurer, S. (2015), Lifecycle patterns in the socioeconomic gradient of risk preferences, Journal of Economic Behavior and Organization, 119, 482-495. We first estimated a linear regression of personality on all control variables, excluding father's occupation class controls on the sample of individuals in the university group. We then predicted the residual in personality change. The standardized residuals are then plotted against the continuous measure of father's occupational class using bivariate kernel regression methods (In Stata - lpolyci - command.)*

Table A.1: *Balance of covariates between university (Uni) and non-university (Non-Uni) groups*

	Non-Uni (1)	Uni (2)	p-value (3)
Extraversion 2013	4.495	4.654	0.088
Agreeableness 2013	5.301	5.404	0.176
Conscientiousness 2013	4.904	5.027	0.156
Emotional Stability 2013	4.905	4.971	0.503
Openness to Experience 2013	4.258	4.506	0.007
Extraversion 2005	4.719	4.589	0.140
Agreeableness 2005	5.054	5.198	0.063
Conscientiousness 2005	4.447	4.673	0.009
Emotional Stability 2005	4.819	4.894	0.420
Openness to Experience 2005	4.215	4.538	0.000
Father's occupation prestige score	44.626	58.226	0.000
Age	24.804	24.583	0.061
Female	0.496	0.672	0.000
Country of birth: Australia	0.953	0.938	0.452
Country of birth: English-speaking	0.021	0.005	0.081
Country of birth: Other	0.026	0.057	0.096
Major urban area of residence	0.614	0.792	0.000
Regional area of residence	0.386	0.208	0.000
Got married	0.157	0.109	0.107
Separated from spouse	0.311	0.203	0.004
Pregnancy	0.300	0.063	0.000
Birth/adoption of new child	0.240	0.047	0.000
Serious personal injury/illness	0.264	0.208	0.135
Serious injury/illness to family member	0.470	0.385	0.052
Death of close relative/family member	0.473	0.432	0.360
Death of a close friend	0.358	0.245	0.005
Victim of a property crime	0.272	0.307	0.377
Fired or made redundant	0.292	0.177	0.001
Changed jobs	0.744	0.833	0.011
Promoted at work	0.384	0.385	0.970
Changed residence	0.783	0.771	0.737
HH income 1-9,999	0.037	0.021	0.266
HH income 10,000-19,999	0.034	0.016	0.156
HH income 20,000-29,999	0.055	0.068	0.551
HH income 30,000-39,999	0.107	0.036	0.001
HH income 40,000-49,999	0.104	0.083	0.406
HH income 50,000-59,999	0.078	0.078	0.993
HH income 60,000-79,999	0.078	0.063	0.477
HH income 80,000-99,999	0.133	0.089	0.098
HH income 100,000-124,999	0.115	0.167	0.101
HH income 125,000-149,999	0.110	0.177	0.035
HH income $>$ 150,000	0.149	0.203	0.114
Government School	0.726	0.547	0.000
Catholic non-Government school	0.164	0.240	0.039
Other non-Government school	0.099	0.203	0.002
Other school	0.010	0.010	0.998
Years in paid work between 2005-2013	6.256	6.393	0.033
Differences in physical health 2005-2013	1.663	−0.647	0.115
Number of observations	383	192	

Source: HILDA, waves 5 and 13

Note: Includes all respondents aged 15 to 19 in wave 5. Column (3) reports the p-value of a t-test of equality of means between Uni and Non-Uni group means.

Table A.2: *Balance of covariates between university (Uni) and non-university (Non-Uni) groups, after matching with Coarsened Exact Matching method*

	Non-Uni (1)	Uni (2)	p-value (3)
Extraversion 2005	4.787	4.763	0.853
Agreeableness 2005	5.369	5.247	0.287
Conscientiousness 2005	4.612	4.728	0.407
Emotional Stability 2005	5.002	5.085	0.540
Openness to Experience 2005	4.455	4.539	0.533
Father's occupation prestige score	50.267	53.082	0.403
Age	24.481	24.509	0.867
Female	0.632	0.632	1.000
Country of birth: Australia	0.972	0.981	0.653
Country of birth: English-speaking	0.019	0.009	0.563
Country of birth: Other	0.009	0.009	1.000
Major urban area of residence	0.660	0.774	0.068
Regional area of residence	0.340	0.226	0.068
Serious personal injury/illness	0.321	0.198	0.042
Serious injury/illness to family member	0.425	0.368	0.402
Death of close relative/family member	0.519	0.462	0.412
Death of a close friend	0.396	0.245	0.018
Changed jobs	0.774	0.840	0.225
Changed residence	0.755	0.764	0.873
HH income 1-9,999	0.019	0.000	0.157
HH income 10,000-19,999	0.019	0.000	0.157
HH income 20,000-29,999	0.000	0.038	0.044
HH income 30,000-39,999	0.094	0.028	0.045
HH income 40,000-49,999	0.113	0.104	0.826
HH income 50,000-59,999	0.075	0.132	0.178
HH income 60,000-79,999	0.094	0.085	0.811
HH income 80,000-99,999	0.160	0.142	0.703
HH income 100,000-124,999	0.151	0.198	0.368
HH income 125,000-149,999	0.113	0.113	1.000
HH income $>$ 150,000	0.160	0.160	1.000
Government School	0.755	0.575	0.006
Catholic non-Government school	0.132	0.245	0.035
Other non-Government school	0.113	0.170	0.239
Other school	0.000	0.009	0.318
Years in paid work between 2005-2013	6.177	6.344	0.123
Differences in physical health 2005-2013	−0.271	−0.375	0.967
Number of observations	113	113	

Source: HILDA, waves 5 and 13

Note: Includes all respondents aged 15 to 19 in wave 5. Column (3) reports the p-value of a t-test of equality of means between Uni and Non-Uni group means.

Table A.3: *Full estimation results: Determinants of 2013 Big-Fiver personality traits*

	Extrv	Agree	Consc	Emote	Openn
UNI	0.289***	0.0192	0.0257	0.0283	0.107
	(0.0866)	(0.0781)	(0.0795)	(0.0874)	(0.0867)
FOPS	−0.0345	0.113*	−0.0655	−0.0445	0.0560
	(0.0522)	(0.0616)	(0.0555)	(0.0555)	(0.0575)
UNI × FOPS	0.0176	−0.225***	−0.0248	−0.0527	−0.0328
	(0.0782)	(0.0854)	(0.0803)	(0.0832)	(0.0817)
Age	0.0663	0.0400	−0.0248	0.0866*	0.0110
	(0.0423)	(0.0471)	(0.0451)	(0.0476)	(0.0465)
Female	−0.0782	0.460***	0.141*	−0.134*	−0.0908
	(0.0768)	(0.0794)	(0.0796)	(0.0801)	(0.0774)
COB: Eng speaking	0.139	0.427**	0.101	0.189	−0.185
	(0.218)	(0.215)	(0.263)	(0.220)	(0.195)
COB: Other	−0.275	−0.284	−0.441**	−0.347	0.237
	(0.205)	(0.176)	(0.174)	(0.248)	(0.148)
Non major urban	0.0531	0.0527	−0.0751	−0.199**	−0.0365
	(0.0801)	(0.0874)	(0.0898)	(0.0926)	(0.0832)
Got married	−0.151	−0.0111	0.269**	0.0360	0.0128
	(0.105)	(0.118)	(0.107)	(0.122)	(0.118)
Separated from spouse	0.202**	−0.0630	0.000683	−0.0167	0.0311
	(0.0863)	(0.0955)	(0.0843)	(0.0922)	(0.0892)
Pregnancy	0.338**	0.00715	0.0265	0.281	0.0979
	(0.157)	(0.261)	(0.152)	(0.207)	(0.239)
Birth/adoption of new child	−0.241	−0.236	−0.0832	−0.316	−0.167
	(0.164)	(0.256)	(0.171)	(0.221)	(0.241)
Serious personal injury/illness	−0.0138	0.0361	−0.0784	−0.172*	0.123
	(0.0820)	(0.0888)	(0.0840)	(0.0899)	(0.0848)
Serious injury/illness to family member	0.0640	0.0616	0.0119	−0.00894	0.0691
	(0.0725)	(0.0792)	(0.0767)	(0.0812)	(0.0746)
Death of close relative/family member	0.0880	−0.0147	0.266***	0.234***	−0.0916
	(0.0694)	(0.0815)	(0.0760)	(0.0788)	(0.0773)
Death of a close friend=1	0.225***	0.157*	0.0375	0.145*	0.0279
	(0.0729)	(0.0891)	(0.0815)	(0.0812)	(0.0838)
Victim of a property crime=1	0.0481	−0.0349	−0.0446	−0.0104	0.0662
	(0.0797)	(0.0971)	(0.0886)	(0.0908)	(0.0885)
Fired or made redundant=1	−0.0741	−0.0404	−0.166*	−0.112	0.0509
	(0.0848)	(0.0939)	(0.0859)	(0.0901)	(0.0926)
Changed jobs	0.0602	0.111	−0.00645	0.138	−0.0313
	(0.0857)	(0.108)	(0.0986)	(0.100)	(0.0922)
Promoted at work	0.00841	0.0615	0.170**	−0.0458	0.0436
	(0.0745)	(0.0837)	(0.0810)	(0.0814)	(0.0799)
Changed residence	0.0137	0.0469	0.126	−0.0922	−0.0809
	(0.0909)	(0.0995)	(0.0916)	(0.106)	(0.0972)
Extraversion 2005	0.528***	0.0346	−0.0424	−0.0229	0.0246
	(0.0365)	(0.0391)	(0.0367)	(0.0384)	(0.0393)
Agreeableness 2005	−0.0346	0.389***	0.0177	0.0735	−0.0542
	(0.0472)	(0.0569)	(0.0535)	(0.0486)	(0.0477)
Conscientiousness 2005	−0.0433	−0.0200	0.443***	0.0386	−0.0312
	(0.0426)	(0.0435)	(0.0421)	(0.0414)	(0.0400)
Emotional stability 2005	0.0536	0.0386	−0.00558	0.405***	0.101**
	(0.0400)	(0.0404)	(0.0393)	(0.0420)	(0.0425)
Openness to experience 2005	0.101**	−0.00588	−0.00109	0.0426	0.495***
	(0.0423)	(0.0423)	(0.0431)	(0.0439)	(0.0423)
Catholic non-government school	0.0658	0.0160	−0.00412	−0.174	0.0780
	(0.0963)	(0.0958)	(0.101)	(0.106)	(0.0966)
Other non-government school	0.129	−0.123	−0.0852	−0.167	0.193*
	(0.107)	(0.113)	(0.108)	(0.104)	(0.112)
Other school	−0.904***	−0.246	−0.733***	0.0202	−0.330
	(0.210)	(0.211)	(0.276)	(0.377)	(0.408)
HH income 10,000-19,999	0.137	−0.313	−0.110	0.111	−0.229
	(0.254)	(0.369)	(0.289)	(0.313)	(0.299)
HH income 20,000-29,999	−0.140	−0.177	−0.0381	0.537*	−0.208
	(0.251)	(0.252)	(0.252)	(0.293)	(0.238)
HH income 30,000-39,999	−0.177	−0.0995	0.119	0.280	0.0191
	(0.222)	(0.245)	(0.238)	(0.275)	(0.222)
HH income 40,000-49,999	−0.0567	−0.401	−0.0298	0.475*	−0.318
	(0.210)	(0.263)	(0.247)	(0.275)	(0.223)
HH income 50,000-59,999	0.113	−0.350	0.178	0.386	−0.253
	(0.227)	(0.248)	(0.239)	(0.297)	(0.208)
HH income 60,000-79,999	0.192	−0.134	0.373	0.219	−0.121
	(0.212)	(0.243)	(0.235)	(0.271)	(0.213)
HH income 80,000-99,999	0.156	−0.250	0.0980	0.305	0.000626
	(0.213)	(0.241)	(0.237)	(0.266)	(0.202)
HH income 100,000-124,999	0.00277	−0.269	0.176	0.365	−0.270
	(0.206)	(0.230)	(0.228)	(0.261)	(0.197)
HH income 125,000-149,999	0.0268	−0.369	0.376	0.462*	−0.357*
	(0.211)	(0.237)	(0.230)	(0.262)	(0.204)
HH income > 150,000	0.234	−0.306	0.182	0.570**	−0.433**
	(0.208)	(0.247)	(0.230)	(0.263)	(0.198)
Years in paid work between 2005 and 2013	0.0528	0.0246	−0.0487	0.00917	−0.0900**
	(0.0390)	(0.0435)	(0.0427)	(0.0407)	(0.0375)
Difference in physical functioning between 2005 and 2013	0.0270	0.00173	0.0586	0.0635	−0.0279
	(0.0401)	(0.0435)	(0.0366)	(0.0504)	(0.0387)
Constant	−3.187***	−2.339***	−2.308***	−2.949***	−2.183***
	(0.405)	(0.437)	(0.467)	(0.465)	(0.424)
Observations	575	575	575	575	575
R^2	0.403	0.269	0.334	0.272	0.311

Source: HILDA, waves 5 and 13.
Note: Includes all respondents aged 15 to 19 in wave 5. Each dependent variable is standardised to mean 0 and standard deviation 1. Omitted are dummy variables that indicate whether life event was imputed (if missing), or if health or work experience information was missing.
* p < 0.1, ** p < 0.05, *** p < 0.01. Standard errors in parentheses.

Table A.4: *Estimated means of Big Five personality traits by university and non-university group using Coarsened Exact Matching*

	Uni	Non-Uni	Diff
Panel A: Completed/about to complete uni degree (113 pairs)			
Extraversion	0.19	–0.07	0.26*
Agreeableness	0.05	0.23	–0.18
Conscientiousness	0.09	0.06	0.03
Emotional stability	0.15	–0.01	0.16
Openness to experience	0.17	0.09	0.08
Panel B: Completed uni degree (94 pairs)			
Extraversion	0.24	–0.04	0.29*
Agreeableness	0.07	0.34	–0.27**
Conscientiousness	0.26	0.11	0.15
Emotional stability	0.23	0.03	0.20
Openness to experience	0.13	0.20	–0.07

Note: Reported are mean personality traits by university and non-university group and differences in means. Matching is based on: Age (above versus below or equal to 25 years), Sex (0,1), Father's Occupational Prestige Score (above versus below or equal to the mean of 49.8), Parent's household income (above versus below or equal $60,000-79{,}999$), urban region (0,1), and each personality trait in 2005 (above versus below or equal to 4). Out of 144 individuals who completed university, we find a perfect match for 94 individuals; remaining individuals are discarded. Out of 192 individuals who entered or completed university for at least two years, we find a perfect match for 113 individuals; remaining individuals are discarded.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table A.5: *Estimated effects of university participation (UNI) on Big-Five personality traits: Specification check of heterogeneity in the effect of university education (Fixed Effects model)*

	Extrv (1)	Agree (2)	Consc (3)	Emote (4)	Openn (5)
PANEL A: Father's Occupational Prestige Score (FOPS, standardised)[a]					
Estimated coefficients[a]					
UNI=1	0.253***	−0.033	−0.118	0.009	−0.075
	(0.096)	(0.094)	(0.097)	(0.115)	(0.111)
UNI=1 × FOPS	0.022	−0.153**	0.015	−0.145*	−0.082
	(0.077)	(0.072)	(0.074)	(0.077)	(0.078)
Marginal effect for each group[b]					
Low SES	0.231*	0.120	−0.133	0.154	0.007
High SES	0.274***	−0.186*	−0.103	−0.136	−0.157
Panel B: FOPS by quartile indicator variables					
Estimated coefficients					
UNI (Base: 1st quartile)	−0.058	0.392**	−0.253	−0.244	0.285
	(0.234)	(0.192)	(0.212)	(0.238)	(0.259)
UNI × 2nd	0.491	−0.265	0.010	0.293	−0.380
	(0.315)	(0.274)	(0.263)	(0.328)	(0.357)
UNI × 3rd	0.354	−0.654***	0.300	0.586**	−0.424
	(0.262)	(0.228)	(0.249)	(0.283)	(0.307)
UNI × 4th	0.332	−0.565**	0.121	0.004	−0.475*
	(0.258)	(0.219)	(0.237)	(0.255)	(0.281)
Marginal effect for SES groups[b]					
Bottom 25th	−0.058	0.392**	−0.253	−0.244	0.285
Second	0.433**	0.127	−0.243	0.049	−0.095
Third	0.296**	−0.262*	0.047	0.342*	−0.139
Top 25th	0.274**	−0.173	−0.132	−0.240*	−0.189
PANEL C: FOPS categories by specific skill groups					
Estimated coefficients					
UNI (Unskilled)	0.088	0.288*	−0.291	−0.172	0.074
	(0.208)	(0.170)	(0.183)	(0.226)	(0.272)
UNI × Trade/Tech/Sales	0.264	−0.413**	0.308	0.447*	−0.197
	(0.231)	(0.196)	(0.207)	(0.253)	(0.296)
UNI × Prof./Manag.	0.138	−0.482**	0.105	−0.170	−0.230
	(0.239)	(0.206)	(0.217)	(0.245)	(0.293)
Marginal effect for SES groups[b]					
Unskilled	0.088	0.288*	−0.291	−0.172	0.074
Trade/Tech/Sales	0.353***	−0.125	0.017	0.275*	−0.122
Prof/Manager	0.227*	−0.194	−0.186	−0.342**	−0.155
PANEL D: Father's post-secondary training					
Estimated coefficients					
UNI (No post-secondary training)	0.445**	0.094	−0.167	0.068	−0.317*
	(0.175)	(0.142)	(0.188)	(0.182)	(0.164)
UNI=1 × Employer-provided	−0.118	−0.145	−0.097	−0.333	0.271
	(0.274)	(0.311)	(0.300)	(0.306)	(0.401)
UNI=1 × Teacher college/TAFE	−0.379	−0.200	0.109	−0.104	0.537**
	(0.231)	(0.202)	(0.232)	(0.274)	(0.237)
UNI=1 × University	−0.211	−0.290*	0.085	−0.167	0.205
	(0.201)	(0.172)	(0.209)	(0.200)	(0.189)
Marginal effect for SES groups[b]					
No post-secondary	0.445**	0.094	−0.167	0.068	−0.317*
Employer-provided	0.327	−0.051	−0.264	−0.264	−0.045
Teacher college/TAFE	0.066	−0.107	−0.058	−0.035	0.220
University	0.234*	−0.196	−0.082	−0.099	−0.112

Note: Each model is estimated with a first-difference model in which personality trait k is the dependent variable. Each model controls for the full set of time-varying control variables, and time-variant, individual-specific heterogeneity. Sample size is N=575 and T=2, and 192 individuals are in the university educated group (UNI=1). [a] Father's occupational prestige score (FOPS) is a continuous measure standardized to mean 0 and standard deviation 1. A one-standard deviation increase in FOPS is 23.48 points on a scale from 0 to 100. [b] Marginal effects of the interaction terms are calculated in STATA using the - lincom - command. Each dependent variable is standardised to mean 0 and standard deviation 1.

Source: HILDA, waves 5 and 13, respondents ages 15 to 19 in wave 5.

* p < 0.1, ** p < 0.05, *** p < 0.01. Standard errors in parentheses.

Table A.6: *Estimated effects of university participation on Big-Five personality traits by sex (Value-added model)*

	Extrv (1)	Agree (2)	Consc (3)	Emote (4)	Openn (5)
Male sample (N = 256)					
Panel A: Without interaction term					
UNI=1	0.293**	0.117	0.009	0.064	0.251
	(0.140)	(0.157)	(0.157)	(0.140)	(0.157)
R^2	0.45	0.31	0.38	0.26	0.35
Panel B: With interaction term					
UNI=1	0.272*	0.172	0.067	0.075	0.251
	(0.141)	(0.156)	(0.155)	(0.146)	(0.158)
UNI=1 × FOPS	0.089	−0.235	−0.245*	−0.046	−0.002
	(0.125)	(0.157)	(0.128)	(0.122)	(0.142)
Marginal effect for adolescents from high and low SES families					
High	0.361**	−0.063	−0.178	0.029	0.250
Low	0.183	0.407*	0.311	0.121	0.253
R^2	0.45	0.31	0.39	0.26	0.35
Female sample (N = 319)					
Panel C: Without interaction term					
UNI=1	0.290**	−0.053	0.020	−0.039	0.120
	(0.125)	(0.095)	(0.100)	(0.114)	(0.111)
R^2	0.43	0.20	0.38	0.36	0.38
Panel D: With interaction term					
UNI=1	0.291**	−0.012	0.017	−0.038	0.138
	(0.127)	(0.093)	(0.099)	(0.118)	(0.112)
UNI=1 × FOPS	−0.006	−0.278***	0.020	−0.007	−0.120
	(0.110)	(0.096)	(0.105)	(0.117)	(0.101)
Marginal effect for adolescents from high and low SES families					
High	0.285*	−0.290**	0.037	−0.044	0.018
Low	0.298*	0.266*	−0.003	−0.031	0.258
R^2	0.44	0.22	0.38	0.36	0.38

Note: Each model controls for the full set of control variables. The value-added model includes lagged personality traits as additional control variables. Sample includes all respondents aged 15 to 19 in wave 5. FOPS stands for Father Occupational Prestige Score. A one-standard deviation increase in occupational prestige is 23.48 points on a scale from 0 to 100. Each dependent variable is standardised to mean 0 and standard deviation 1.
Source: HILDA, waves 5 and 13.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Standard errors in parentheses.

Bibliography

Allport, GW and HS Odbert (1936). Trait-names: A psycho-lexical study. *Psychological monographs* **47**(1), i.

Almlund, M, AL Duckworth, JJ Heckman, and T Kautz (2011a). “Personality Psychology and Economics”. In: *Handbook of the Economics of Education*. Ed. by EA Hanushek, SJ Machin, and L Woessmann. Vol. 4. Elsevier. Chap. 1, pp.1–181.

Almlund, M, A Lee Duckworth, JJ Heckman, and T Kautz (2011b). “Personality psychology and economics”. In: *Handbook of the Economics of Education*. Ed. by EA Hanushek, S Machin, and L Woessmann. Vol. 4. Amsterdam: North-Holland, pp.1–181.

Arrow, K (1973). Higher Education as a Filter. *Journal of Public Economics* **2**(3), 193–216.

Arrow, K (1997). “The benefits of education and the formation of preferences”. In: *The social benefits of education*. Ed. by JR Behrman and N Stavey. Ann Arbor: The University of Michigan Press, pp.11.

Asendorpf, JB (2002). *The puzzle of personality types*.

Asendorpf, JB (2003). Head-to-head comparison of the predictive validity of personality types and dimensions. *European Journal of Personality* **17**(5), 327–346.

Asendorpf, JB, P Borkenau, F Ostendorf, and MA Van Aken (2001). Carving personality description at its joints: Confirmation of three replicable personality prototypes for both children and adults. *European Journal of personality* **15**(3), 169–198.

Asendorpf, JB and JJ Denissen (2006). Predictive validity of personality types versus personality dimensions from early childhood to adulthood: Implications for the distinction between core and surface traits. *Merrill-Palmer Quarterly (1982-)*, 486–513.

Ashton, MC and K Lee (2007). Empirical, theoretical, and practical advantages of the HEXACO model of personality structure. *Personality and social psychology review* **11**(2), 150–166.

Ashton, MC, K Lee, M Perugini, P Szarota, RE De Vries, L Di Blas, K Boies, and B De Raad (2004). A six-factor structure of personality-descriptive adjectives: solutions from psycholexical studies in seven languages. *Journal of personality and social psychology* **86**(2), 356.

Australian Bureau of Statistics (2001). *Australian Standard Classification of Education (ASCED)*. ABS Catalogue 1272.0. Canberra: Australian Bureau of Statistics.

Australian Bureau of Statistics (2013). 4102.0 - Australian Social Trends, July 2013, Available at: https://www.abs.gov.au/AUSSTATS/abs@.nsf/Lookup/4102.0Main+Features20July+2013#Top [Accessed 01 Jun. 2017].

Bartholomew, DJ (2007). "Three faces of factor analysis". In: *Factor analysis at 100: historical developments and future directions*. Ed. by R Cudeck and RC MacCallum. Lawrence Erlbaum Associates. Chap. 2, pp.9–22.

Bartholomew, DJ, M Knott, and I Moustaki (2011). *Latent variable models and factor analysis: A unified approach*. John Wiley & Sons.

Becker, A, T Deckers, T Dohmen, A Falk, and F Kosse (2012). The relationship between economic preferences and psychological personality measures. *Annual Review of Economics* **4**, 453–478.

Bigoni, M, S Bortolotti, M Cesari, D Gembetta, and F Pancotto (2016). Amoral familism, social capital, or trust? The behavioural foundations of the Italian North-South Divide. *The Economic Journal* **594**, 1318–1341.

Black, N and S Kassenboehmer (2017). Getting weighed down: The effect of childhood obesity on the development of socioemotional skills. *Journal of Human Capital* **11**(2), 263–295.

Bleidorn, W, TA Klimstra, JJA Denissen, PJ Rentfrow, J Potter, and SD Gosling (2013). Personality Maturation Around the World: A Cross-Cultural Examination of Social-Investment Theory. *Psychological Science* **24**(12), 2530–2540.

Block, J (1995). A contrarian view of the five-factor approach to personality description. *Psychological bulletin* **117**(2), 187.

Block, J (2010). The five-factor framing of personality and beyond: Some ruminations. *Psychological Inquiry* **21**(1), 2–25.

Block, JH and J Block (1980). "The role of ego-control and ego-resiliency in the organization of behavior". In: *Development of cognition, affect, and social relations*. Psychology Press, pp.49–112.

Borgatta, EF (1964). The structure of personality characteristics. *Behavioral science* **9**(1), 8–17.

Brown, AL, J Bransford, RA Ferrara, and JC Campione (1983). "Learning, remembering, and understanding". In: *Handbook of child psychology: Cognitive development*. Ed. by JHF amd E. H. Markman. Vol. 3. New York: Wiley, pp.77–176.

Carneiro, P, K Hansen, and J Heckman (2003). Estimating Distributions of Treatment Effects with an Application to the Returns to Schooling and Measurement of the Effects of Uncertainty on College Choice. *International Economic Review* **44**(2), 361–422.

Caspi, A and PA Silva (1995). Temperamental qualities at age three predict personality traits in young adulthood: Longitudinal evidence from a birth cohort. *Child development* **66**(2), 486–498.

Chen, H and JH Friedman (2017). A new graph-based two-sample test for multivariate and object data. *Journal of the American statistical association* **112**(517), 397–409.

Chetty, R, JN Friedman, N Hilger, E Saez, D Whitmore Schanzenbach, and D Yagan (2011). How Does Your Kindergarten Classroom Affect Your Earnings? Evidence from Project Star. *The Quarterly Journal of Economics* **126**(4), 1593–1660.

Cobb-Clark, DA and S Schurer (2012). The stability of Big-Five personality traits. *Economics Letters* **115** (1), 11–15.

Cobb-Clark, DA and S Schurer (2013). Two economists' musings on the stability of locus of control. *The Economic Journal* **123**(570), F358–F400.

Costa Jr, PT, JH Herbst, RR McCrae, J Samuels, and DJ Ozer (2002). The replicability and utility of three personality types. *European Journal of Personality* **16**(S1), S73–S87.

Cramer, AO, S Van der Sluis, A Noordhof, M Wichers, N Geschwind, SH Aggen, KS Kendler, and D Borsboom (2012). Dimensions of normal personality as networks in search of equilibrium: You can't like parties if you don't like people. *European Journal of Personality* **26**(4), 414–431.

Cunha, F, J Heckman, and S Navarro (2005). Separating Uncertainty from Heterogeneity in Life Cycle Earnings. *Oxford Economic Papers* **57**(2), 191–261.

Cunha, F and JJ Heckman (2008). Formulating, identifying and estimating the technology of cognitive and noncognitive skill formation. *Journal of Human Resources* **43**(4), 738–782.

De Fruyt, F and EB Karevold (2021). "Personality in adolescence." In: *Handbook of personality: Theory and research*. Ed. by OP John and RW Robins. The Guilford Press. Chap. 13, pp.303–321.

Del Bono, E, M Francesconi, Y Kelly, and A Sacker (2016). Early maternal time investment and early child outcomes. *The Economic Journal* **126**(596), F96–F135.

Delbanco, A (2014). *College: What it was, is, and should be*. Princeton University Press, Princeton, New Jersey.

Deming, DJ (2017). The growing importance of social skills in the labor market. *The Quarterly Journal of Economics* **132**(4), 1593–1640.

DeVitis, JL (2013). *Contemporary colleges and universities*. New York, NY: Peter Lang Publishing, Inc.

Digman, JM and J Inouye (1986). Further specification of the five robust factors of personality. *Journal of personality and social psychology* **50**(1), 116.

Digman, JM and NK Takemoto-Chock (1981). Factors in the natural language of personality: Re-analysis, comparison, and interpretation of six major studies. *Multivariate behavioral research* **16**(2), 149–170.

Docherty, D (2012). Employability: University education isn't just about developing skills. *The Guardian*. [online], Available at: https://www.theguardian.com/higher-education-network/blog/2012/apr/04/employability-university-education-developing-skills [Accessed 27 Jul. 2016].

Doornik, JA and H Hansen (2008). An omnibus test for univariate and multivariate normality. *Oxford Bulletin of Economics and Statistics* **70**(s1), 927–939.

Elkins, R, S Kassenboehmer, and S Schurer (2017). The Stability of Personality Traits in Adolescence and Young Adulthood. *Journal of Economic Psychology* **60**, 37–52.

Epskamp, S, D Borsboom, and EI Fried (2018). Estimating psychological networks and their accuracy: A tutorial paper. *Behavior research methods* **50**, 195–212.

Fiorini, M and MP Keane (2014). How the allocation of children's time affects cognitive and noncognitive development. *Journal of Labor Economics* **32**(4), 787–836.

Fisher, A, C Rudin, and F Dominici (2019). All Models are Wrong, but Many are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously. *J. Mach. Learn. Res.* **20**(177), 1–81.

Fisher, A (2016). How to Make a Liberal Arts Degree a Career Asset, Available at: https://www.fortune.com/2016/05/12/liberal-arts-degree-career-asset/ [Accessed 01 Jun. 2017].

Fiske, DW (1949). Consistency of the factorial structures of personality ratings from different sources. *The Journal of Abnormal and Social Psychology* **44**(3), 329.

Fletcher, J (2013). The effects of personality traits on adult labor market outcomes: Evidence from siblings. *Journal of Economic Behavior and Organization* **89**(C), 122–135.

Fletcher, J and S Schurer (2017). Origins of adulthood personality: The role of adverse childhood experiences. *B.E. Journal of Economic Analsys & Policy* **17**(2).

Freudenstein, JP, C Strauch, P Mussel, and M Ziegler (2019). Four personality types may be neither robust nor exhaustive. *Nature human behaviour* **3**(10), 1045–1046.

Friedman, JH (1991). Multivariate adaptive regression splines. *The annals of statistics* **19**(1), 1–67.

Friedman, JH and BE Popescu (2008). Predictive learning via rule ensembles. *The annals of applied statistics*, 916–954.

García, JL, JJ Heckman, DE Leaf, and MJ Prados (2020). Quantifying the life-cycle benefits of an influential early-childhood program. *Journal of Political Economy* **128**(7), 2502–2541.

Gensowski, M (2018). Personality, IQ, and lifetime earnings. *Labour Economics* **51**, 170–183.

Gerlach, M, B Farb, W Revelle, and LA Nunes Amaral (2018). A robust data-driven approach identifies four personality types across four large data sets. *Nature human behaviour* **2**(10), 735–742.

Ghahramani, Z and GE Hinton (1996). *The EM algorithm for mixtures of factor analyzers*. Tech. rep. Technical Report CRG-TR-96-1, University of Toronto.

Gilbert, R, J Balatti, P Turner, and H Whitehouse (2004). The generic skills debate in research higher degrees. *Higher Education Research & Development* **23**(3), 375–388.

Goldberg, LR (1992a). The development of markers for the Big-Five factor structure. *Psychological Assessment* **4**(1), 26.

Goldberg, LR (1992b). The development of markers for the Big-Five factor structure. *Psychological assessment* **4**(1), 26.

Goldberg, LR (1993a). The structure of phenotypic personality traits. *American psychologist* **48**(1), 26.

Goldberg, LR (1993b). The structure of phenotypic personality traits. *American Psychologist* **48**, 26–34.

Goldstein, A, A Kapelner, J Bleich, and E Pitkin (2015). Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *journal of Computational and Graphical Statistics* **24**(1), 44–65.

Golino, HF and S Epskamp (2017). Exploratory graph analysis: A new approach for estimating the number of dimensions in psychological research. *PloS one* **12**(6), e0174035.

Graduate Careers Australia (2014). *The report of the graduate outlook survey: Employers' perspectives on graduate recruitment*. Graduate Outlook 2013 Report. Melbourne: Graduate Careers Australia.

Green, DA and WC Riddell (2003). Literacy and earnings: an investigation of the interaction of cognitive and unobserved skills in earnings generation. *Labour Economics* **10**(2), 165–184.

Hastie, T, R Tibshirani, JH Friedman, and JH Friedman (2009). *The elements of statistical learning: data mining, inference, and prediction*. Vol. 2. Springer.

Heckman, JJ, S Moon, R Pinto, P Savelyev, and A Yavitz (2010). Analyzing social experiments as implemented: A reexamination of the evidence from the HighScope Perry Preschool Program. *Quantitative Economics* **1**, 1–46.

Heckman, JJ, R Pinto, and P Savelyev (2013). Understanding the mechanisms through which an influential early childhood program boosted adult outcomes. *American Economic Review* **103** (6), 2052–86.

Heckman, JJ, T Jagelka, and T Kautz (2021). "Some contributions of economics to the study of personality." In: *Handbook of personality: Theory and research*. Ed. by OP John and RW Robins. The Guilford Press. Chap. 42, pp.852–892.

Heckman, JJ, J Stixrud, and S Urzua (2006a). The Effects of Cognitive and Noncognitive Abilities on Labor Market Outcomes and Social Behavior. *Journal of Labor Economics* **24**(3), 411–482.

Heckman, JJ, LJ Lochner, and PE Todd (2006). "Earnings Functions, Rates of Return and Treatment Effects: The Mincer Equation and Beyond". In: ed. by E Hanushek and F Welch. Vol. 1. Handbook of the Economics of Education. Elsevier, pp.307–458.

Heckman, JJ, J Stixrud, and S Urzua (2006b). The Effects of Cognitive and Noncognitive Abilities on Labor Market Outcomes and Social Behavior. *Journal of Labor Economics* **24**(3), 411–482.

Heineck, G (2011). Does it pay to be nice? Personality and earnings in the United Kingdom. *ILR Review* **64**(5), 1020–1038.

Heineck, G and S Anger (2010). The returns to cognitive abilities and personality traits in Germany. *Labour Economics* **17** (3), 535–546.

Herzberg, PY and M Roth (2006). Beyond resilients, undercontrollers, and overcontrollers? An extension of personality prototype research. *European Journal of Personality* **20**(1), 5–28.

Hilbig, BE, A Gloeckner, and I Zettler (2014). Personality and prosocial behavior: Linking basic traits and social value orientations. *Journal of Personality and Social Psychology* **107**(3), 529–539.

Hofstee, WK, HA Kiers, B De Raad, LR Goldberg, and F Ostendorf (1997). A comparison of Big-Five structures of personality traits in Dutch, English, and German. *European Journal of Personality* **11**(1), 15–31.

Hopwood, C, M Donnellan, D Blonigen, R Krueger, M McGue, W Iacono, and S Burt (2011). Genetic and environmental influences on personality trait stability and growth during the transition to adulthood: A three wave longitudinal study. *Journal of Personality and Social Psychology* **100**(3), 545–556.

Horn, JL (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika* **30**(2), 179–185.

Iacus, SM, G King, and G Porro (2011). Multivariate matching methods that are Monotonic Imbalance Bounding. *Journal of the American Statistical Association* **106**(493), 345–361.

Jacob, B (2002). Where the boys aren't: Non-cognitive skills, returns to school and the gender gap in higher education. *Economics of Education Review* **21**(6), 589–598.

John, OP (2021). "History, measurement, and conceptual elaboration of the Big-Five trait taxonomy: The paradigm matures." In: *Handbook of personality: Theory and research*. Ed. by OP John and RW Robins. The Guilford Press. Chap. 2, pp.35–82.

Jones, AM, N Rice, and P Rosa Dias (2011). Long-term effects of school quality on health and lifestyle: Evidence from comprehensive schooling reforms in England. *Journal of Human Capital* **5**(3), 342–376.

Kassenboehmer, SC, F Leung, and S Schurer (2018). University education and non-cognitive skill development. *Oxford Economic Papers* **70**(2), 538–562.

Katahira, K, Y Kunisato, Y Yamashita, and S Suzuki (2020). Commentary: A robust data-driven approach identifies four personality types across four large data sets. *Frontiers in big Data* **3**, 8.

Kautz, T, JJ Heckman, R Diris, B ter Weel, and L Borghans (2015). *Fostering and Measuring Skills - Improving Cognitive and Non-cognitive Skills to Promote Lifetime Success*. Education Working Papers 110. Paris: OECD.

Kerber, A, M Roth, and PY Herzberg (2021). Personality types revisited–a literature-informed and data-driven approach to an integration of prototypical and dimensional constructs of personality description. *PloS one* **16**(1), e0244849.

Linden, D van der, CS Dunkel, and K Petrides (2016). The General Factor of Personality (GFP) as social effectiveness: Review of the literature. *Personality and Individual Differences* **101**, 98–105.

Luedtke, O, BW Roberts, U Trautwein, and G Nagy (2011). A random walk down university avenue: Life paths, life events, and personality trait change at the transition to university life. *Journal of Personality and Social Psychology* **101**(3), 620–637.

Lundberg, S (2013). The College Type: Personality and Educational Inequality. *Journal of Labour Economics* **31**(3), 421–441.

Mareckova, J and W Pohlmeier (2017). Noncognitive Skills and Labor Market Outcomes: A Machine Learning Approach.

McCrae, RR (1982). Consensual validation of personality traits: evidence from self-reports and ratings. *Journal of Personality and Social Psychology* **43**(2), 293.

McCrae, RR and PT Costa (1987). Validation of the five-factor model of personality across instruments and observers. *Journal of personality and social psychology* **52**(1), 81.

McCrae, RR and PT Costa Jr (1997). Personality trait structure as a human universal. *American psychologist* **52**(5), 509.

McCrae, RR and A Terracciano (2005). Universal features of personality traits from the observer's perspective: data from 50 cultures. *Journal of personality and social psychology* **88**(3), 547.

McLachlan, GJ, D Peel, and R Bean (2003). Modelling high-dimensional data by mixtures of factor analyzers. *Computational Statistics & Data Analysis* **41**(3-4), 379–388.

McMillan, J, A Beavis, and FL Jones (2009). The AUSEI06: A new socioeconomic index for Australia. *Journal of Sociology* **45**(2), 123–149.

Melnykov, V, WC Chen, and R Maitra (2012). MixSim: An R package for simulating data to study performance of clustering algorithms. *Journal of Statistical Software* **51**, 1–25.

Mincer, J (1958). Investment in human capital and personal income distribution. *Journal of Political Economy* **66**, 281–302.

Moodie, G (2015). How different are higher education institutions in the UK, US and Australia. *Higher Education Quarterly* **69**(1), 3–36.

Musek, J (2007). A general factor of personality: Evidence for the Big One in the five-factor model. *Journal of research in personality* **41**(6), 1213–1233.

Norman, WT (1963). Toward an adequate taxonomy of personality attributes: Replicated factor structure in peer nomination personality ratings. *The Journal of Abnormal and Social Psychology* **66**(6), 574.

Norton, A and I Cherastidtham (2014). *Mapping Australian higher education, 2014-15*. Grattan Institute Report 2014-11. Melbourne: Grattan Institute.

Nyhus, EK and E Pons (2005). The effects of personality on earnings. *Journal of Economic Psychology* **26**, 363–384.

OECD (2012). *Education at a Glance*. Paris: OECD Publishing.

OECD (2013). *Education at a Glance 2013: OECD Indicators*. Paris: OECD Publishing.

Palczyńska, M (2021). Wage premia for skills: the complementarity of cognitive and non-cognitive skills. *International Journal of Manpower* **42**(4), 556–580.

Paris, S, MY Lipson, and KK Wixson (1983). Becoming a strategic reader. *Contemporary Educational Psychology* **8**, 293–316.

Post, SG (2005). Altruism, happiness, and health: It's good to be good. *International Journal of Behavioural Medicine* **2**, 66–77.

Roberts, BW (1997). Plaster or plasticity: Are adult work experiences associated with personality changes in women? *Journal of Personality* **65**, 205–232.

Robins, RW, OP John, A Caspi, TE Moffitt, and M Stouthamer-Loeber (1996). Resilient, overcontrolled, and undercontrolled boys: three replicable personality types. *Journal of Personality and Social psychology* **70**(1), 157.

Roth, M and G von Collani (2007). A head-to-head comparison of big-five types and traits in the prediction of social attitudes: Further evidence for a five-cluster typology. *Journal of Individual Differences* **28**(3), 138–149.

Saucier, G (1994a). Mini-Markers: A brief version of Goldberg's unipolar Big-Five markers. *Journal of personality assessment* **63**(3), 506–516.

Saucier, G (1994b). Mini-markers: A brief version of Goldberg's unipolar Big-Five markers. *Journal of personality assessment* **63**(3), 506–516.

Saucier, G (2003). An alternative multi-language structure for personality attributes. *European Journal of Personality* **17**(3), 179–205.

Saucier, G and LR Goldberg (2001). Lexical studies of indigenous personality factors: Premises, products, and prospects. *Journal of personality* **69**(6), 847–879.

Schultz, T (1961). Investment in human capital. *American Economic Review* **51**, 1–17.

Schurer, S (2017). Does education strengthen life skills of adolescents? *IZA World of Labor* **366**(June).

Schurer, S, M Alspach, J MacRae, and G Martin (2016). The medical cost of mood disorders: A coarsened exact matching approach. *The Economic Record* **92**(296), 81–93.

Specht, J, B Egloff, and SC Schmukle (2011). Stability and Change of Personality Across the Life Course: The Impact of Age and Major Life Events on Mean-Level and Rank-Order Stability of the Big Five. *Journal of Personality and Social Psychology* **101**(4), 862–882.

Spence, MA (1973). Job Market Signaling. *Quarterly Journal of Economics* **87**(3), 355–374.

Stiglitz, J (1975). The Theory of Screening, Education, and the Distribution of Income. *American Economic Review* **65**(3), 283–300.

Summerfield, M, S Freidin, M Hahn, P Ittak, N Li, N Macalalad, N Watson, R Wilkins, and M Wooden (2013). *HILDA User Manual–Release 12*.

Terenzini, PT, L Springer, PM Yaeger, ET Pascarella, and A Nora (1996). First-generation college students: Characteristics, experiences, and cognitive development. *Research in Higher Education* **37**(1), 1–22.

Thomas, JW and WD Rohwer Jr. (1986). Academic Studying: The Role of Learning Strategies. *Educational Psychologist* **1-2**, 19–41.

Todd, PE and KI Wolpin (2003). On the specification and estimation of the production function for cognitive achievement. *The Economic Journal* **113**(485), F3–F33.

Tupes, EC and RE Christal (1992). Recurrent personality factors based on trait ratings. *Journal of personality* **60**(2), 225–251.

Vallone, D (2013). Majors, lecture theatres and classes all day: what is college really like?, Available at: https://www.independent.co.uk/student/study-abroad/so-what-is-liberal-arts-and-why-does-everyone-in-america-study-it-8933110.html [Accessed 01 Jun. 2017].

Vedel, A (2016). Big Five personality group differences across academic majors: A systematic review. *Personality and individual differences* **92**, 1–10.

Ware Jr, JE (2000). SF-36 health survey update. *Spine* **25**(24), 3130–3139.

Weinberger, CJ (2014). The increasing complementarity between cognitive and social skills. *Review of Economics and Statistics* **96**(5), 849–861.

Weiss, A (1995). Human Capital vs. Signaling Explanations of Wages. *Journal of Economic Perspectives* **9**(4), 133–154.