Aalto University
School of Business

THE PREDICTIVE POWER OF PSYCHOMETRIC TRAITS ON LABOUR MARKET OUTCOMES

An Examination of the Perry Preschool Program's Causal Analysis

Master's Thesis
Joonas Ylönen
Aalto University School of Business
Master's Programme in Economics
Spring 2022

Aalto University
School of Business

Aalto University, P.O. BOX 11000, 00076 AALTO

www.aalto.fi

Abstract of master's thesis

Author Joonas Ylönen

Title of thesis The Predictive Power of Psychometric Traits on Labour Market Outcomes

Degree Master of Science (Economics and Business Administration)

Degree programme Master's Programme in Economics

Thesis advisor(s) Marko Terviö

Year of approval 2022 **Number of pages** 70 **Language** English

Abstract

Even though, the association between cognitive ability and labour market outcomes is well-established, only recently have economists started to examine the role of non-cognitive traits. During the past two decades, mainly due to the wide acceptance of the Big Five personality factor model, the association between non-cognitive traits and labour market outcomes has started to receive greater attention.

In this thesis, I attempt to review the main findings made within both the psychological and the economic literature. The main concepts and terminology are largely gathered from former literature, whereas the econometrical tools utilized in associational studies and the causal analysis are derived from the latter literature. This thesis offers both descriptive and causal findings while placing substantial emphasis on the analysis of a longitudinal intervention study, namely the Perry Preschool Program – one of the most influential studies in this branch of literature.

I show that there exists a widely accepted association between non-cognitive traits and labour market outcomes. Further, the Perry Program analysis finds a significant causal effect between a change in the non-cognitive traits of the treatment group and the later exhibited labour market outcomes. However, the results are likely not, at least to the full extent, generalizable. What is evident though, is that the mechanism through which non-cognitive traits affect labour market outcomes remains unknown. Interestingly, despite this, this mechanism seems to differ between men and women.

Moreover, I shed light on the utilization of econometrical tools and assumptions required to prove an associational or causal relationship between non-cognitive traits and labour market outcomes.

Keywords psychometrical traits, personality, intelligence, earnings, early education, Big Five, Perry Preschool Program

Aalto-yliopisto
Kauppakorkeakoulu

Aalto-yliopisto, PL 11000, 00076 AALTO

www.aalto.fi

Maisterintutkinnon tutkielman tiivistelmä

Tekijä Joonas Ylönen

Työn nimi Psykometristen ominaisuuksien selittävä voima työmarkkina lopputulemiin

Tutkinto Kauppatieteiden maisteri

Koulutusohjelma Taloustieteen maisteriohjelma

Työn ohjaaja(t) Marko Terviö

Hyväksymisvuosi 2022 **Sivumäärä** 70 **Kieli** englanti

Tiivistelmä

Kognitiivisen kyvykkyyden välinen yhteys työmarkkina lopputulemiin on suurelta osin vakiintunut, tästä huolimatta taloustieteilijät ovat kuitenkin vasta hiljattain alkaneet tutkia ei-kognitiivisten piirteiden osuutta työmarkkina lopputulemiin. Tämän yhteys on alkanut saamaan kasvavaa huomiota vasta viimeisen kahden vuosikymmenen aikana pitkälti laaja-alaisen persoonallisuus teoriakehyksen hyväksymisen kautta, nimeltään Big Five.

Tämä opinnäytetyö yrittää tuoda esiin keskeisimpiä löydöksiä niin psykologian kuin myös taloustieteellisen kirjallisuuden saralta, edellä mainittuun yhteyteen liittyen. Käytetty terminologia ja avain konseptit ovat pitkälti tuotu psykologisen kirjallisuuden piiristä, kun puolestaan taloustieteellisestä kirjallisuudesta on hyödynnetty ekonometrisia työkaluja, joiden avulla yhteys ei-kognitiivisten taipumusten ja työmarkkinalopputulemien välillä kyetään osoittamaan. Opinnäytetyö painottaa kirjallisuuden kulmakivenä toimivaa pitkittäis- ja interventiotutkimusta, nimeltään Perry Preschool Program.

Näytän, että ei-kognitiivisten piirteiden ja työmarkkina lopputulemien väliltä löytyy laajasti hyväksytty yhteys. Lisäksi Perry Program:in analyysi osoittaa merkittävän kausaalivaikutuksen ei-kognitiivisissa piirteissä tapahtuvan muutoksen ja työmarkkina lopputulemien väliltä. Tämä tulos ei kuitenkaan ole todennäköisesti, ainakaan kokonaisuudessaan, yleistettävissä. Selvää on kuitenkin se, että reitti, jota pitkin ei-kognitiiviset piirteet vaikuttavat työmarkkina lopputulemiin on yhä tuntematon. Mielenkiintoa asiaan lisää se, että tästä huolimatta, tämä tuntematon reitti kuitenkin vaikuttaisi eroavan miesten ja naisten välillä.

Lisäksi tutkielmani tuo tarkasti esiin niin tavan, jolla ekonometrisiä työkaluja hyödynnetään, sekä sen millaisia oletuksia kausaaliyhteyden osoittaminen vaatii.

Avainsanat psykometriset ominaisuudet, persoonallisuus, älykkyys, ansiotulot, varhaiskasvatus, Big Five, Perry Ohjelma

Table of Contents

1. Introduction 1
2. Background for psychometrical terminology 5
2.1 Intelligence 5
2.2 Personality 7
2.2.1 Changes in personality over the life cycle 12
2.2.2 Faking 12
3. The association between personality traits and earnings 15
3.1 Common associational findings 16
3.1.1 Evidence from Germany 19
3.2 Evidence from the GED Testing Program 21
3.3 Evidence from a laboratory experiment 26
3.4 The conclusions of the association-based relationship between personality and earnings 29
4. The Perry Preschool Program 31
4.1 Background 31
4.2 The Perry Program's early results 33
4.3 Data 35
4.4 Imputation 36
4.4.1 Proxy creation 36
4.4.2 Imputation methods 38
4.4.3 Imputation results on earnings estimates 41
4.5 The treatment effect on labour market outcomes 44
4.5.1 Criminal activity 44
4.5.2 Employment 46
4.5.3 Earnings 48
5. Perry Program validity concerns and result interpolation 52
5.1 Internal validity 52
5.2 Generalizability 55
5.3 The economic returns of the Perry program 57
5.4 Result interpolation 58
5.5 Perry Program Conclusions 59
6. Discussion 60
7. Conclusions 66

List of Figures

Figure 1, Personal earnings and the Big Five: evolution over time, from Figure 4 Alderotti (2021) ... 18
Figure 2 Labor Market Differences, Ages 20–39 (Males, All Levels of Postsecondary Education) from Figure 5.6 Heckman et al. (2014) ... 24
Figure 3 Annual Earnings by Type of Female GED Recipient (All Races, Background and Ability-Adjusted) from Figure 5.45 Heckman et al. (2014) ... 25
Figure 4 Personality Traits Density Distribution by Gender from Figure 1 Cubel et al. (2016) ... 28
Figure 5 Perry Preschool Program: IQ, by Age and Treatment Group from Figure 14A Cunha et al. (2006) ... 34
Figure 6 : Probability of Two or More Violent Criminal Convictions over the Life Course for Males from Figure 7 Heckman (2019) ... 46
Figure 7 Proportion of Employed Population: Male, Percentage from Figure E.3 Heckman et al. (2010) Web Appendix ... 47
Figure 8 Mean Annual Earnings over the Life Course for Males from Figure 8 Heckman (2019) ... 49

List of Tables

Table 1, Modified from Table 1.3 Almlund (2011) ... 8
Table 2 Earnings with different imputation methods, Modified from Table G.5 Heckman et al. (2010) Web Appendix ... 41

1. Introduction

What is the fundamental driver behind labour market outcome disparity, is one of the great questions of economics. There is a wide consensus between economists that human capital, through its great influence on labour productivity, accounts for most of the differences we encounter. Yet, despite its crucial role, our causal understanding regarding the formation of human capital remains largely unknown. In an attempt to provide insight into this great question, this thesis will be conducted as a literature review and will focus on two questions: firstly, what is the predictive power of psychometric traits on labour market outcomes? and secondly, if psychometric traits do predict labour market outcomes, can they be altered? In an attempt to answer the former questions, studies using different methods have been included in order to utilize the strengths and account for the weaknesses a single method may encounter. More precisely this question will be examined from the perspectives of meta-analytical literature review, quasi-natural association analysis and laboratory experiment.

The latter question is explored by a thorough examination of a famous analysis of an ongoing longitudinal intervention study on early education, namely the Perry Preschool Program, which serves as one of the most influential studies in this branch of research. This thesis places great emphasis upon the evaluation of the Perry Program because it presents causal evidence about the power of psychometrics on labour market outcomes. Moreover, later analyses of the program have provided evidence that the non-cognitive abilities of treatment group participants were changed, while these changes led to better labour market outcomes. Further, the causal analysis presented by Heckman et al. (2010), introduced an array of econometrical tools to address inherent study compositional challenges, evaluation of which will be at the centre of this thesis.

Intuitively it is obvious that psychometrical traits, i.e., intelligence, and non-cognitive abilities, e.g., an individual's personality, do both greatly affect the opportunities open to a person, and hence, influence the choices they make, which in turn, generate the experienced life outcomes. However, within the economic literature, the inclusion of non-cognitive abilities in the analysis has proven to be especially challenging. This stems partly from not having sufficient metrics to accurately account for non-cognitive abilities and partly because this void has been mostly filled by behavioural economics, where behaviour is believed to be

determined almost entirely by situational constraints and incentives (Almlund Chapter 1, 2011). However, this once popular view of extreme situationism is largely outdated in mainstream psychological literature. It has been replaced by the notion of somewhat, even though the ever-changing, stable non-cognitive character of a person. This framework will lay the foundation upon which the later thesis rests.

Within the psychological literature, a person's cognitive ability can be reliably derived from a standardized intelligence quotient (IQ)-test, where one's total score on standardized tests mirrors one's cognitive capability. It should be noted that IQ can be divided into two sub-categories: fluid intelligence and crystallized intelligence. The former describes the ability to solve novel problems, whereas the latter describes the knowledge and developed skills. The relative weighting of fluid and crystallized intelligence slightly varies between IQ tests. Despite these small inconsistencies between different IQ tests, they can be compared without any significant alterations. On top of this, they are widely used, and hence, we do possess a great deal of available data on IQ. By utilizing these data sets, the association between cognitive ability and socio-economic success has been deeply studied in the last century. Therefore, this linkage is well documented in both psychological and economic literature today (Becker 1976).

When we focus on labour market outcomes, the importance of IQ increases with job complexity: cognitive skills are more important for professors and senior managers than for unskilled labour. In a sense, the role of IQ in the labour market can be viewed through the lens of thresholds. If a person possesses an IQ lesser than the required threshold for the job, there is hardly anything that could compensate for this lack of cognitive ability.

Despite the great predictive power of IQ, a large proportion of labour market outcomes is left unexplained even after the role of pure cognitional ability has been taken into account. Further, the previous literature has placed a substantial emphasis on cognitive ability compared to other traits, which has left a vital part of the equation unexplained. However, during the past two decades, this gap has been recognised within the field of economics, which has led to a wider adaptation of measures measuring, especially personality. Partly because individual earning differences can only partly be explained with conventional background characteristics, and partly since measures of personality have been adopted into

large individual-level datasets. Hence, in this thesis, I will focus on the relative new literature on non-cognitive traits and their association with labour market outcomes.

The main findings from this literature review can be divided into three parts: first, non-cognitive traits are associated with a variety of labour market outcomes, especially the Big Five personality factor Consciousness is associated with significantly better labour outcomes (Almlund Chapter 7, 2011; Cubel 2016), even when cognitive ability and educational level are controlled for (Alderotti 2021), whereas Neuroticism is constantly associated with more unfavourable labour market outcomes (Almlund Chapter 7, 2011; Cubel 2016). Moreover, Heckman et al. (2010) do find a *causal* relationship between non-cognitive traits and labour market outcomes in their Perry Program analysis. The causal evidence presented in Heckman et al. (2010), shows that participation in the treatment group increased lifetime[1] earnings estimate by 10 or even 35 per cent, while also increasing the employment rate by 20 per cent among the treatment group participants compared to the control group counterparts. However, the generalizability of this *causal* relationship does still undergo constant debate. (e.g., Xie et al. 2020)

Secondly, the inability to explain the avenue through which non-cognitive traits affect labour market outcomes unites both the association and causal studies. Even though, Heckman et al. (2010) *causally* demonstrated that an early education intervention significantly altered life outcomes, while the mean levels of IQ remained unchanged, they are unable to pinpoint the mechanism from which these outcome differences arise. Hence, the fundamental challenge this line of research encounters is the inability to distinguish the role each trait or ability plays in certain behaviour. In other words, performance on a single task is dependent on multiple parts of a person's non-cognitive and cognitive functions which is why singling out one part of a person's non-cognitive realm has proven to be difficult. This, in turn, creates a fundamental identification problem, that most of the papers have simply ignored.

In addition, the majority of the published papers, especially before the last decade, suffer from relatively low levels of statistical power mainly due to small sample sizes. More recently, as discussed above, the wider inclusion of personality metrics into individual-level datasets has

[1] Lifetime earnings estimates do include information regarding the age 40 follow-up.

slightly alleviated this concern. This evolution provides great opportunities for future research.

Thirdly, there seem to exist significant gender differences in both the distribution of non-cognitive, e.g., personality traits (Cubel 2016), and the effects and associations between certain non-cognitive traits and labour market outcomes. (Cubel 2016, Heckman 2010, 2014) More specifically, as illustrated in figure 4, even within a randomly selected laboratory experiment sample of university students, significant gender differences in the distribution of Neuroticism and Agreeableness can be found. (Cubel 2016) This finding is consistent with the current consensus within the psychological literature that the greatest differences in personality between genders can be found in the average level and distribution of Neuroticism and Agreeableness. Women are more likely to express extreme values of the aforementioned traits. (Weisberg 2011) Moreover, in the meta-analytical literature review, Alderotti et al. (2021) emphasize that their results indicate that with male-only samples the positive association between Openness and earnings was particularly significant, while the negative association between Neuroticism and earnings was significantly smaller with female-only samples

Further, for men, the increased labour market outcomes in the Perry Program seem to stem from a decreased level of criminal activity, and consequently, a lower rate of incarceration, whereas, for women, the increased labour market outcomes seem to stem from increased educational attainment and the decreased rate of teenage pregnancies. (Heckman 2010) Similarly, even though, GED receiving women's hourly wage does not differ from high school dropouts, they are more likely to participate in labour markets, whereas GED receiving men’s labour market outcomes do not significantly differ from those high school dropouts. (Heckman 2014). To conclude, there exists no conclusive evidence of the mechanisms through which personality affects labour market outcomes, but the mechanism seems to differ between men and women.

The rest of this thesis is organized into three sections, where chapter 2 serves as an introduction to psychometrical terminology, upon which the later analysis rests. Chapter 3 introduces evidence on the association between personality and labour market outcomes with different approaches in order to gather a broader understanding of this complex phenomenon.

Later chapters introduce, analyse, and discuss the Perry Preschool program causal study analysis in detail, which has served as the most influential causal study on the relationship between personality and labour market outcomes. While also highlighting and providing an in-depth examination of the heavy utilization of econometric tools, and the underlying assumptions needed to produce *causal* results within this field.

2. Background for psychometrical terminology

In this thesis, the term psychometrical traits refer to a combination of two distinct categories: intelligence and personality. Ideally, the association and effect of both intelligence and personality on labour market outcomes would be done separately, however, a sharp distinction between these two categories is challenging to make, hence most of the relevant literature studies them both simultaneously. Consequently, one must have a basic understanding of both of these underlying concepts upon which the later work builds to critically evaluate the correlational and causal studies presented in this thesis. The following definition section mostly follows the structure presented by Almlund et al. (2011) (especially chapters 2 and 5), a foundational and the most thorough literature review combining psychological and economic literature up to date.

2.1 Intelligence

As described in the introduction, the crucial role of intelligence in producing better (labour market) outcomes has been thoroughly demonstrated. But questions regarding what intelligence means, where is it derived from, how is it utilized today and what limitations it faces, are rarely thoroughly examined. To get an answer to these questions, I will shortly lay out the relevant historical evolution of our modern intelligence tests.

Initially, in 1904 the French government commissioned Alfred Binet to devise a test to identify retarded[2] children in need of special instructions. In response, the first IQ test was created. Later Binet's test came to be known as Stanford-Binet Intelligence Scale after Terman (1916) translated and revised the version of the test for the American children. In

[2] Since 1950 the term mentally retarded has been pointed to people with two standard deviations lower IQ of 70 than the populational average 100. (Flynn 2000) Nowadays term mentally retarded has been placed with more politically correct words like mentally disadvantaged, mentally challenged, etc.

addition to mere translation, Terman developed a procedure to compare children within the same age group against one another, by which the individual receives an intelligence quotient (IQ) score. This score is normally distributed, and the average is 100. Historically IQ tests have, therefore, been primarily designed to predict academic performance. The utilization of IQ tests in other fields of life outcomes has later followed.

The utilization of the Stanford-Binet test received critique for overemphasizing verbal skills, which is dependent on cultural exposure and formal education. As an answer to this critique the Wechsler Adult Intelligence Scale (WAIS) and Wechsler Intelligence Scale for Children (WISC) were created in the 1950'. Similarly, to Stanford-Binet, both WAIS and WISC are normally distributed at each age. For the past few decades, WAIS and WISC have been the most widely used IQ tests, however, numerous slightly differing IQ formats do exist and are used, which makes the comparison between different studies somewhat hard to navigate.

To ease the interpretation between different IQ tests, it is useful to distinguish two distinct IQ categories: *fluid intelligence* and *crystallised intelligence*. The former is defined as the ability to actively solve novel problems, for example, to solve verbal analogies. The skills involved are usually not taught and are believed to represent a person's raw information procession power. (Gottfredson & Saklofske 2009) Whereas the latter is defined as the use of knowledge acquired through schooling and other life experiences. For example, tests of general information, word comprehension and numerical abilities are among the measures of crystallized intelligence. Interestingly, fluid intelligence tends to peak in early adulthood and then decline, whereas crystallized IQ tends to steadily increase throughout the life cycle.

In practice, an intelligence test partially captures both realms of intelligence but the relative weighting of these two realms of intelligence does differ between the tests, depending on the role of prior experiences in performance. An example of a fluid intelligence test is a Raven Progressive Matrices, generally referred to as an aptitude test. Whereas an example of a crystallized intelligence test, widely used in national level datasets, and hence, by economists, is Armed Forces Qualifying Test (AFQT) generally referred to as the achievement test.

Partially explained by the historical evolution, the validity of an IQ test is generally established by comparing test scores with other test scores or with school grades, not success in life in general. (Almlund Chapter 5, 2011) Following a similar logic, the AFQT scores are

validated by the combination of success in military training schools and performance on standardized tasks like fixing a rifle. This route of validation may sound inadequate to an economist, yet it is considered standard practise among psychologists. Nevertheless, the predictive power of the ability and achievement test on occupational and life outcomes is well established. (Kuncel 2010, from Almlund Chapter 5, 2011)

2.2 Personality

Personality (traits) has received increased attention among economists during the last two decades, as a partial answer to increasing the predictive power of individual ability proxies when combined with IQ. (Almlund Chapter 2, 2011) However, unlike with cognition, the set of terminologies and conventions for personality (traits) still experience rigorous debates within the psychological literature. Even the definition of personality undergoes constant debate due to its diverse nature. In order to make sense of this complex branch of literature, personality traits are defined, in the context of this thesis, as "the relatively enduring patterns of thoughts, feelings, and behaviours that reflect the tendency to respond in certain ways under certain circumstances" (Roberts 2009).

Despite the ongoing debate regarding the definition of personality, a proper theoretical framework to describe and measure has emerged. Psychologists have widely, even though not universally, accepted the taxonomy, i.e., categorization of personality traits, called the Big Five. This Five-Factor model has its origins in a lexical hypothesis (1936), which assumes that most important individual differences are encoded in our natural language. Building on this assumption, during the 1980s, multiple independent psychologists analysed personality describing words found in the English dictionary[3] with factor analysis, a technique used to reduce many variables into fewer number factors[4]. Subsequently, as a result of this analysis, it has been concluded that personality traits can be organized into five superordinate factors:

[3] 17 953 words to be precise. Later from these personality describing words only adjectives were considered.
[4] Factor analysis extracts maximum common variance from all variables and puts them into a common score, or in this case a common category. However, to be valid several assumptions must be met. There must exist linear relationship, there is no multicollinearity, it must include relevant variables into analysis and there exists a true correlation between variables and factors. The Big Five framework endures critical assessment of these assumptions. (Lee et al. Chapter 25, 2007)

Openness to Experience, Consciousness, Extraversion, Agreeableness and Neuroticism, explained in more detail in table 1.

Table 1, Modified from Table 1.3 Almlund (2011)

Big Five Personality Factor	**APA Dictionary Description**	**Facets (and Correlated Trait Adjective)**
Conscientiousness	"the tendency to be organized, responsible, and hardworking"	Competence (efficient), Order (organized), Dutifulness (not careless), Achievement striving (ambitious), Self-discipline (not lazy), Deliberation (not impulsive)
Openness to Experience	"the tendency to be open to new aesthetic, cultural, or intellectual experiences"	Fantasy (imaginative), Aesthetic (artistic), Feelings (excitable), Actions (wide interests), Ideas (curious), Values (unconventional)
Extraversion	"an orientation of one's interests and energies toward the outer world of people and things rather than the inner world of subjective experience; characterized by positive affect and sociability"	Warmth (friendly), Gregariousness (sociable), Assertiveness (self-confident), Activity (energetic), Excitement seeking (adventurous), Positive emotions (enthusiastic)
Agreeableness	"the tendency to act in a cooperative, unselfish manner"	Trust (forgiving), Straightforwardness (not demanding), Altruism (warm), Compliance (not stubborn), Modesty (not show-off),

		Tendermindedness (sympathetic)
Neuroticism / Emotional Stability	Emotional stability is "predictability and consistency in emotional reactions, with absence of rapid mood changes." Neuroticism is "a chronic level of emotional instability and proneness to psychological distress."	Anxiety (worrying), Hostility (irritable), Depression (not contented), Self-consciousness (shy), Impulsiveness (moody), Vulnerability to stress (not self-confident)

These five factors can later be reduced to lower-level facets, as shown in table 1. These lower-level facets can then be further subdivided into more narrow traits, which are then used in the personality surveys from which a description of an individual's personality is derived. (Costa and McCrae 1992) For example, impulsivity, one of the lower-level facets of a factor Neuroticism, can be further subdivided into narrower context-dependent traits, like impulsivity toward junk food and impulsivity towards drinking. In this way, these narrow lower-level traits will be coupled into broader lower-level traits, which will be further coupled to form the individual's Neuroticism factor. Like IQ, these factors are analysed against other people's levels of the same Big Five factor, however, unlike IQ, these differences are examined linearly.

Further, on a more technical level, the procedure of picking a task to measure a particular trait is called operationalization, which is closely related to construct validity. Construct validity refers to a mechanism to check whether the chosen trait constructed in the stage of operationalization correlates with measures that are supposedly representing the trait. Almlund et al. (Chapter 3, 2011), do warn that within this process lies a great danger of circularity, and thus, a considerable amount of judgement is required in the process of operationalizing a trait in question, and in independently validating it. Since this process

requires expertise far beyond the level, I'm capable of critically analysing, this thesis depends on the trustworthiness of the pre-existing published work.

Unsurprisingly, the Big Five model is not without its critics. One should note that Big Five is silent about motivation (i.e., what people value or desire), and for that reason, motivation is largely ignored in this thesis. Despite this, the Big Five factor Consciousness includes the facet "achievement striving", which serves as an indirect proxy for motivation, even though it is far from being fully comprehensive[5]. In addition, the Big Five has received critique because the first-factor analysis studies and personality studies were conducted mainly on English-speaking samples. Even though the Big Five framework has appeared to replicate across many cultures, concerns have still been raised, and additional factors have been proposed to account for cultural differences. The most popular of which is the HEXACO six-factor model of personality that utilizes the lexical approach similarly to the Big Five, but unlike the Big Five it utilizes several different languages around the world. (Ashton et al., 2004). From this cultural-inclusive framework rises the sixth personality factor called Honesty-Humility, which accounts for differences in fairness and modesty, while the factors agreeableness and emotionality also slightly differ. Although the expressed concerns are relevant, and the HEXACO model offers a detailed way to address them, the proposed alterations, are rather cosmetic compared to the Big Five model.

Secondly, the Big Five model has been criticized for being atheoretical. (Block 1995, from Almlund Chapter 5, 2011) As DeYound et al. (2010)[6] point out, "the finding that descriptions of behaviour as measured by tests, self-reports, and reports of observers cluster reliably into five groups (factors) has not been satisfactorily explained by any scientifically grounded theory". The psychobiological model of personality serves as an alternative and independent model to the Big Five since its seven factors of personality are derived from neurobiological, developmental and genetic studies, which provide a theoretical foundation upon which the psychobiological model of personality is built on. (Feher et al., 2021) In other words, alternatives to the Big Five do exist but they either only add cosmetic improvements or

[5] In fact, as Almlund (Chapter 3, 2011) argues "some motivation theorists believe that one's deepest desires are unconscious and, therefore, may dispute the practice of measuring motivation using self-report questionnaires".

[6] From Almlund Chapter 5 (2011)

require a greatly increased level of information about the individuals, which makes their utilization in broad individual level questionnaires implausible and nearly impossible.

Thirdly, there exists a practical problem: personality questionnaires are not uniform, and they largely differ between studies. The plurality of personality questionnaires partly reflects the more heterogenous nature of personality compared to cognitive ability, as discussed above. This also reflects the incomplete agreement among the personality psychologists on the Big Five model, as well as the lack of consensus among researchers about identifying and organizing lower-order facets of the Big Five factors, as Peterson et al. (2007) point out.

For example, the impulsivity facet discussed above has classically been categorized as a facet of Neuroticism (Costa and McCrae 1992), yet others claim it to be a facet of Conscientiousness, and others suggest it to be a blend of Conscientiousness, Extraversion and Neuroticism. (Depue and Collis 1999, from Peterson et al. 2007) These arguments partly arise from the possible use of alternative methodologies for verifying tests, not guaranteed to produce the same taxonomies. In the context of this thesis further analysis of these methodological differences exceeds the boundaries of this thesis.

Additionally, the variation between traits related to one factor might dilute the predictive validities of facets when analyses only consider factor-level aggregate scores. Hence, estimations based on factors-level can efface possible relationships between specific facets and outcomes. For instance, Paunonen and Ashton (2001) demonstrated how facets 'need for achievement' and 'need for understanding' predicted course grades among undergraduates more accurately than the corresponding factors Conscientiousness and Openness to Experience, respectively. Peterson et al. (2007), have raised the same concern. Peterson et al. (2007) suggest that in order to be able to partially account for this factor-level analysis obscuring variation in relevant facets, researchers focus should be shifted to the analysis of the Big Five's second-highest order facets. In the economic literature, this suggestion has been widely gone unnoticed.

Thus, even though the Big Five suffers from obscurity regarding the precise definition of lower-level facets concerning superordinate factors, it does, nonetheless, provide a common language for researchers and organizing personality research. Furthermore, the theory of the

Big Five has been used both widely and successfully for three decades in empirical settings, which makes it the most prominent framework to examine the diverse nature of personality.

2.2.1 Changes in personality over the life cycle

The life cycle malleability of personality introduces an additional layer of challenge into the association analysis between personality and earnings. During the early years of life, a change in absolute levels of personality referred to as mean-level changes are obvious and dramatic. For example, children in experience significant progress in their ability to exercise self-control as they move from infancy into toddler and preschool years. (McCabe et al. (2004), from Almlund Chapter 3, 2011) As people reach adulthood, significant changes in mean levels of personality are rare.

In 2006, a meta-analysis of longitudinal studies examining cumulative lifetime changes in the Big Five (Roberts, Walton and Viechtbauer 2006), concluded that typically people become more socially dominant (a facet of Extraversion), conscientious, agreeable and emotionally stable (non-neurotic) as they age, whereas Openness to Experience tends to rise early in life and then decrease in old age. On an individual level, however, personality can change significantly due to, e.g., dramatic life experiences (trauma). Within the psychological literature, it is believed that genetic factors are largely responsible for stability in personality in adulthood, excluding the typical personality changes explained above, whereas environmental factors are mostly responsible for the perceived change.

2.2.2 Faking

The empirical personality psychological literature has identified a consistent potential threat to self-descriptive personality measurements, faking. This intentional context-dependent manipulation of personality test responses may stem from two potential routes, impression management or self-deception. (Paulhus 1984) For example, in the case of a hiring process, an individual may be inclined to deliberately exaggerate their strengths and downplay their weaknesses in areas that can be viewed as positive in the context of a certain occupation. Similarly, an individual may be inclined to perceive him or herself in a more virtuous (or worse) manner, referred to as self-deception, even in anonymous situations, where no real-life consequences can rise from the answers. These, in turn, may corrupt measurements designed

to proxy latent factors, as Almlund et al. (Chapter 4, 2011) highlight. It could be seen as obvious that in the test-retest reliability creation format when the retest is not immediate, the self-deception inflicted faking is likely to be more consistent through time, compared to the impression management, where an alternative circumstance might call fourth differing virtues, leading to faking in alternative personality facets.

To correct for faking, psychologists have created scales to measure deliberate lying, yet these efforts seem not to improve predictive validity. However, empirical evidence suggests that faking has a relatively small effect on predicting job performance. Hence, while in the current form of personality tests, accounting for faking may be insufficient, merely acknowledging that the problem potentially exists, goes a long way, in drawing conclusions from the personality measurement tests.

2.3 IQ and achievement scores capture both cognitive and personality traits

Overall, a sharp distinction between intelligence and personality is challenging to make. The Big Five factor, Openness to Experience, for example, has facets of curiosity ("ideas") and imagination ("fantasy"), which are associated with intellect and measured intelligence. (McCrae and Costa 1997a). Furthermore, certain aspects of human experience seem to require interaction between an individual's cognitive abilities and aspects of specific personality traits. E.g., creativity seems to stem from an interaction between intelligence and Openness to Experience. Even the inventor of IQ tests, Alfred Binet, acknowledges this intrinsic embeddedness of personality and IQ by noting:

"… admits of other things than intelligence; to succeed in his studies one must have qualities which depend on attention, will, and character; for example, certain docility, a regularity of habits, and especially continuity of effort. A child, even if intelligent, will learn little in class if he never listens, if he spends his time in playing tricks, in giggling, is playing truant."

However, the great benefit of the Big Five model is that the correlation between the five factors is relatively low, and further, apart from Openness to Experience, the correlation with IQ is low. Hence, this mostly independent nature serves as the main driver behind the wide adaptation of the Big Five into the human behaviour outcome analysis. (Almlund Chapter 2,

2011) It should be noted that the correlation between Openness to Experience with IQ should also not be exaggerated.

More recently, numerous studies have studied the role of effort, incentives, and personality in IQ test scores. These studies indicate that incentives, like money or candy, can substantially improve performance on IQ tests, especially among low-IQ individuals. In fact, among the low-IQ individuals, the addition of incentives could increase their performance on IQ tests up to a full standard deviation. (Holt and Hobbs 1979 from Almlund Chapter 5, 2011). However, effects of this magnitude have not been systematically replicated, but most studies indicate statistically significant increases in IQ test scores among the low-IQ individuals when incentives are introduced. In addition, when rewards are higher, individuals take substantially more time answering IQ questions, indicating an increased level of effort in general. (ter Weel et al. 2008). Moreover, individuals, especially men, with lower levels of the Big Five-factor Conscientiousness are particularly affected by the inclusion of incentives. In other words, conscientious individuals tend to already operate at a higher level of effort in the absence of incentives, whereas individuals with a lower level of conscientiousness operate at their best after an incentive is introduced.

On top of this, as brought forth by Almlund (Chapter 5, 2011), IQ tests generally require some level of general knowledge, which is largely acquired through schooling and life experience, which are partly, achieved through curiosity, intrinsic motivation and persistence. Hence, similarly to the expressed effort in IQ tests, personality can indirectly affect IQ scores through the knowledge acquired by an individual higher in the Big Five factors of Conscientiousness and Openness to Experience.

As with insufficient intrinsic motivation, personality can decrease an individual's IQ test scores through its effect on test anxiety, which impairs individuals' performance on a test. Since individuals higher in the Big Five-factor Neuroticism experience test anxiety significantly more often, the factor Neuroticism seems to affect the IQ test scores as well.

In practice, general large-scale databases are commonly utilized, where information about individual-level IQ test scores does not exist. For this reason, it is common to proxy an individual's IQ test scores with standardized achievement test scores due to its high, 0,8, correlation with IQ. (Murray 2002) Similarly to IQ test scores, personality and incentives can

affect the standardized achievement test scores. However, the role of factual knowledge in achievement tests is greatly amplified compared to IQ tests. This, in turn, exposes achievement test scores more heavily to the effect of personality.

Thus, as extensively discussed above, a measure of pure intelligence does not exist, and proxies of intelligence are exposed to measurement error. Moreover, personality test scores, themselves, are laden with intrinsic measurement error because they are relying on self-reports and third-party descriptive analysis.[7] Despite their intrinsic impurity, they can be utilized as proxies with great practical relevance, that provides us a greater understanding of the human condition.

Secondly, the relationship between personality and outcomes can suffer from reverse causality. This may be especially problematic between personality measurements and outcomes, as Almlund et al. (Chapter 4 2011), emphasize. For example, self-esteem might increase income, and income might increase self-esteem. Further, the problem itself, might not be accounted for even if personality is measured before the predicted outcomes. E.g., the anticipation of a future pay raise may by itself increase present self-esteem. Within the psychological literature, these inadequacies are occasionally addressed by using early measures of personality before the outcomes are measured to predict later outcomes. However, since personality does slightly change over time, this method only replaces the problem of reverse causality with errors in variables problem. Hence, in chapter 3 to account for these inadequacies the study with the most rigorous econometrical tools is introduced.

3. The association between personality traits and earnings

There are numerous routes through which personality may affect earnings. Personality is likely to affect educational attainment as described above, labour market status, participation in criminal activity, occupational choice and even the compensation scheme selected within an occupation. Thus, an all-inclusive comprehensive analysis regarding the mechanisms through which personality may affect earnings is far beyond the scope of this thesis. For this

[7] A branch of Freudian motivation theorists does argue that individual's deepest desires are unconscious and, therefore self-report questionnaires are, by definition, unable to measure the concept of motivation (Weinberger 1989).

reason, the main role of this section of the thesis is to serve as an academic introduction to the topic of the association between personality traits and earnings. To support this purpose, section 2 introduces literature from differing points of view, focusing on the mere association between personality traits and earnings. Chapter 2.1 introduces the most recent, and the first, meta-analytical literature review on the existing literature to demonstrate the current cohesion among the scholars around the subject, chapter 2.2 introduces a unique non-intervention study with immense sample size and chapter 2.3 examines a laboratory experiment focusing on the role of individuals personality in productivity when presented with an artificial task.

3.1 Common associational findings

The relationship between personality traits and labour market outcomes has been studied within both the psychological literature, mostly in the 90s, and economic literature mainly during the last two decades. Common findings from the former are a strong positive association of conscientiousness and emotional stability (opposite of neuroticism) on job performance, whereas the labour market outcome associations of other personality traits seem to be confined to certain occupations. E.g., extraversion has a positive association with occupations involving social interactions. Further, the association seem to exist within certain job aspects as well, e.g., openness to experience is associated to training proficiency. (Cubel 2016) Common findings within the economic literature do suggest neuroticism and agreeableness be correlated with lower earnings, while conscientiousness is associated with more favourable labour market outcomes. Additionally, there seems to exist heterogeneity in the effects and distribution of personality traits between genders (see also chapter 2.3), which can partly contribute to the explanation of the gender wage gap.

In support of these common findings, I introduce a recent study by Alderotti et al. (2021) that offers us a meta-analytical review of the empirical literature on the association between Big Five personality traits and earnings by using meta-analysis and meta-regression techniques. Even though only primary studies are able to address specific research questions, a meta-analytical literature view can quantitatively synthesize intriguing overarching themes present within the literature. For example, figure 1 presents the over time evolution of the statistical power of the results in each empirical study.

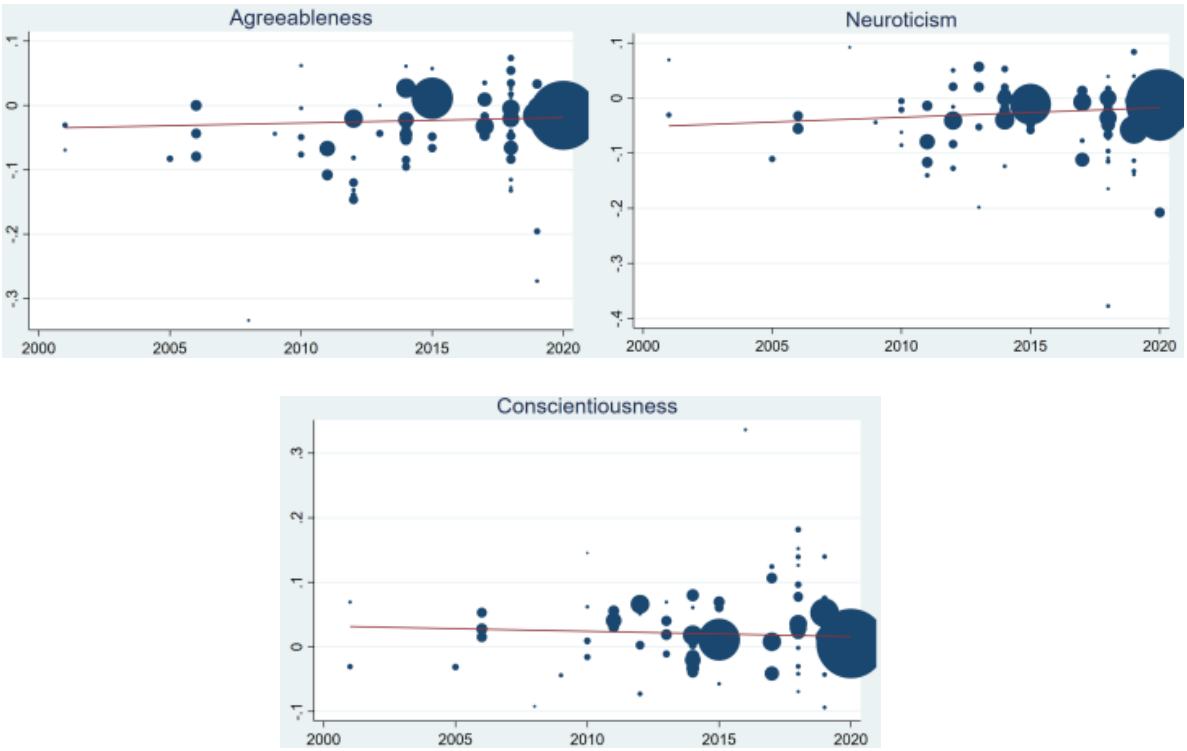
The meta-analytical review included 63 peer-reviewed articles published between 2001 and 2020. To be eligible the paper needed to be indexed on the Scopus database (the largest abstraction and citation database of peer-reviewed literature), the paper needed to be written in English and it needed to contain relevant words in the title, abstract or keywords. Only papers in which the dependent variable was a direct measure of the level of earnings, including life-cycle income estimations and only papers with Big Five personality traits were included.

Before meta-analysis and meta-regressions can be made, Alderotti et al. (2021) needed to address a major challenge with the coding process: part of the included papers did only report whether the estimated coefficients were significant at conventional confidence levels, without including standard errors or t-statistics. (Heckman et al. (2010) do raise and address this same issue, see chapter 4). By creating and then later utilizing an effect size index[8], Alderotti et al. (2021) are allowed to focus on the correlation between the Big Five, and the dependent variable (earnings) while controlling for the confounding factors deemed relevant by the authors of the primary studies.

Figure 1 below offers us a unique illustration of the literature's results evolution over time in the primary literature on earnings and Big Five. Each circle accounts for one primary study. The area of the circles is proportional to the statistical power of the studies, computed as the inverse of the square of effect sizes standard errors and it has been used to weight the linear trend (red line). It becomes evident that the more recent studies exhibit substantially greater statistical power compared to those published further in the past. This is partly explained by the popularity of the Big Five, which led to the inclusion of its short version into a few national surveys, which in turn, offers greater sample sizes and samples with greater national representativeness, allowing more precise estimates. On the other hand, the studies conducted further in the past typically rely on smaller sample sizes, generally gathered by the researchers themselves. Additionally, I only include Agreeableness, Neuroticism and Conscientiousness because they are most consistently linked to labour market outcomes from the Big Five framework, and hence the evolution in these three is highlighted.

[8] See more detailed description of this process from the original Alderotti et al. (2021) paper.

Figure 1, Personal earnings and the Big Five: evolution over time, from Figure 4 Alderotti (2021)



Notes: The figure presents the over time evolution of results of the primary literature between personal earnings and the Big Five. The area of an individual circle is proportional to the statistical power of the corresponding primary study.

What is interesting is that, as the statistical power rises, the found associations between personality traits and earnings seem to trend towards zero. This emerging trend seems to be especially true for Openness and Neuroticism, while the results of other traits tend to be more stable across time. Moreover, the observed between-study heterogeneity within the primary literature result seems to be striking. Both of these observations require further examination.

To conclude, the first quantitative literature review provides evidence to support the significant positive association between Consciousness and earnings, even when cognitive ability and the level of education are controlled for. This finding is consistent with the notion that the positive effect of Consciousness may not be fully mediated by cognitive skills and education. On the other hand, the association between earnings and Neuroticism seem to be significant and negative. On a more descriptive level, the more recently published studies tend

to possess greater statistical power, partly due to the wide acceptance, and hence, use of the Big Five framework.

3.1.1 Evidence from Germany

Guido Heineck and Silke Anger (2009) present joint evidence on the relationship between individual-level cognitive ability, personality, and earnings in Germany. They utilize a sample size of 1580 persons, gathered from the German Socio-Economic Panel Study (SOEP), a representative longitudinal micro-database. To be eligible, the SOEP database had to include a person's both the information of personality and cognitive ability. Moreover, after the German reunification in 1990, the SOEP database extended to include also Eastern Germans, which is why the data utilized in the study is restricted to include only to years 1991 to 2006 (the last datapoint at the time of the study).

Due to the limitations set by the large-scale panel survey, measures of cognitive ability were derived from two ultra-short IQ tests, that followed the WAIS framework described in chapter 2.1. Whereas measures of personality in the SOEP database followed the Big Five framework, with 15 items (questions) that spread evenly to the five factors. Additionally, the study uses locus of control (LOC) measurement, which refers to an individual's perception of the underlying main causes of events in one's life. In other words, it indicates whether a person believes that one has control over life events or whether they are caused by external forces out of one's control. In this study, the external locus of control is the most robust predictor of wage differentials. Moreover, as one might expect, individuals who believe that the outcomes they experience are out of their control have 4 per cent lower wages, which is consisted for both genders.

When we turn our attention to the distributions of Big Five personality traits and cognitive ability, they are spread evenly between the genders, in all other factors than Neuroticism and Agreeableness, which as I will later discuss in more detail, is a common finding within the literature. Moreover, these gender differences seem to manifest themselves in the wage premiums. For females, OLS regression indicates a 2 per cent wage premium for one standard deviation increase in openness, whereas for males, a one standard deviation increase in openness indicates a wage penalty between 2 and 4 per cent. Within the psychological literature, openness is theoretically linked to negative labour market success because openness

is linked to autonomy and non-conformity, as Heineck and Anger argue. What is surprising, however, is that among females, conscientiousness is not associated with greater labour market outcomes. Male, on the other hand, is associated with a wage premium of about 1,5 per cent for a one standard deviation increase in conscientiousness, which is totally in line with the common findings in the literature.

In contrast to existing literature, no statistically significant relationship between neuroticism and wage is found. Heineck and Anger argue that this may be caused by the fact that they control for an individual's attitude towards reciprocal behaviour[9], and since neuroticism and reciprocity are linked, it may be that what is in other related research linked to neuroticism might be reflected as indicators of reciprocity here. The wage premium for females who score high in positive reciprocity earns 3 per cent more and similarly, males earn a little lower premium of about 1-2 per cent. It is interesting to note that male workers' negative reciprocity seems to be rewarded. For women, this seems not to be the case. From this finding, Heineck and Anger conclude that reciprocity and agreeableness are likely to be measured for the same underlying personality trait. Hence, they continue, it may therefore be that the effect of this underlying trait affects outcomes through reciprocity rather than through agreeableness.

When addressing the possibility of nonlinear associations, the fact of belonging to the bottom 25 per cent in Agreeableness among females is associated with a statistically significant 7 per cent wage premium, whereas for males, the same is true with males, whose same wage premium is 3 per cent. The fact of belonging to the top 25 per cent in Agreeableness is not associated with any statistically significant gain or loss in wage. Furthermore, nonlinearities are present in LOC score analysis. Workers who score in the bottom 25 per cent of the LOC scale experience a wage penalty of up to 20 per cent compared to workers who score in the 25 % of the scale. Hence, the role of nonlinear association calls fourth rigorous additional research.

The common challenge of reverse causality between personality and earnings in this study is taken into account by an approach, where each personality trait is regressed on age and the age squared. The residuals from these regressions, then, are free from age affects. In this way,

[9] A social norm that involves in-kind exchanges between people, responding to another's action with another equivalent action. Positive reciprocity is e.g., returning a favour, whereas negative reciprocity is e.g., punishing a negative action. (Heineck and Anger 2010)

while far from being perfect, personality traits are constant over time, which allows matching this information to all preceding waves of SOEP and applying appropriate panel estimators, which together with additional steps, beyond this thesis, helps in partly taking account the concern of reverse causality.

3.2 Evidence from the GED Testing Program

General Educational Development (GED) testing program is not an intervention, and therefore, it distinguishes itself from most of the other literature. The GED test serves as an alternative standardized achievement test by which high school dropouts can prove to attain the general knowledge level of a high school graduate[10]. This can be attained by completing a seven-hour long GED test. The GED test accounts for 12 per cent of approved high school certificates within the U.S., and they are especially popular among ethnic minority groups (Heckman 2012). The GED greatly increases an individual's ability to attain a postsecondary education, e.g., college. Despite this, GED recipients perform substantially worse in labour markets and higher education compared to high school graduates, even when cognitive ability, personal and family background characteristics are accounted for.

Compared to high school dropouts without GED, GED recipients have an overwhelming advantage in the pursuit of post-secondary education. As one might guess, at the age of 40 opportunity differences have turned into outcomes. Nearly 70 per cent of high school graduate males have attended some college, whereas nearly 40 per cent of the GED recipients have attended some college, while the rate is a few per cents among high school dropouts. In this sense, GED recipients resemble more high school graduates than high school dropouts. However, the relative share of people who have acquired a bachelor's degree, are 30, 3 and 1, conversely. (Kautz et al. 2014, Figure 5.13) This outcome difference is both significant and surprising. It begs the question how do these three distinct groups: high school graduates, GED certification recipients and high school dropouts differ from one another? As a partial explanation to this question, the underlying psychometrical metrics are analyzed.

[10] Psychometrical validation for the GED is acquired through a correlational comparison to alternative achievement tests. The correlations are relatively high, e.g., 0.74 correlation with the AFQT test, which on psychometrical grounds indicates a valid test.

Firstly, the distribution of cognitive ability differs among male individuals who have no post-secondary education[11]. High school graduate males do possess the highest level of cognitive ability but the difference among GED recipients is minimal, whereas the high school dropouts do possess significantly lower levels of cognitive ability. In other words, the cognitive ability profile of GED recipients resembles, to a great extent, high school graduates, and they significantly differ from high school dropouts. Similar distributional differences are apparent for women, even though, they slightly differ. In the women sample, the GED recipients and high school graduate intelligence distributions are almost identical, whereas the difference compared to dropouts is more distinct.

However, observed differences in cognitional ability do not fully explain differences in postsecondary schooling. How then, do the non-cognitional (personality) traits differ from one another? On a variety of behavioural dimensions, e.g., early adolescent drug use, crime, sex and violence, GED recipients resemble other dropouts more than high school graduates. GED recipients are in several categories, statistically significantly *more likely* to engage in risky behaviours than other dropouts, while the reverse is not true on any outcome measure.

Behavioural differences between high school graduates and high school dropouts including GED recipients are not only limited to risky behaviour. Generally, behavioural patterns requiring persistence (a facet of Conscientiousness) display similar differences. For example, both the average length of the marriage and the average length of staying out of jail is statistically significantly lower among the high school dropouts and GED recipients compared to high school graduates. Therefore, we can safely assume GED recipients' non-cognitional traits to a great extent mirror high school dropouts, and that these traits greatly differ from the high school graduates.

Up to this point, we have established that GED recipients' cognitive abilities are similar to high school graduates, while their non-cognitive abilities are indistinguishable from the high school dropouts. How are these contradictory forces associated with labour outcomes? Figure 2 illustrates the annual earnings of male GED recipients and high school graduates compared to high school dropouts for different age groups. The information regarding employment

[11] The examination is only limited to this subgroup to create representative counterparts.

status, length and earnings are derived from the National Longitudinal Study of Youth 1979 (NLSY) survey data.

In figure 2 the first set of bars (Raw) refers to regressions where age, race, and region of residence have been used as controls. Similarly, the second set of bars (Abil) shows the outcomes after additionally adjusting for AFQT scores, which is used as a proxy for cognitive ability. The third set of bars (BG) demonstrates the outcome after additionally introducing a full array of family background controls. The bars attached to each column represent ranges of statistical error in the estimate. The regressions allow for heteroskedastic errors. Additionally, regressions exclude individuals earning more than $300 000 (2005 USD), working more than 4 000 hours.

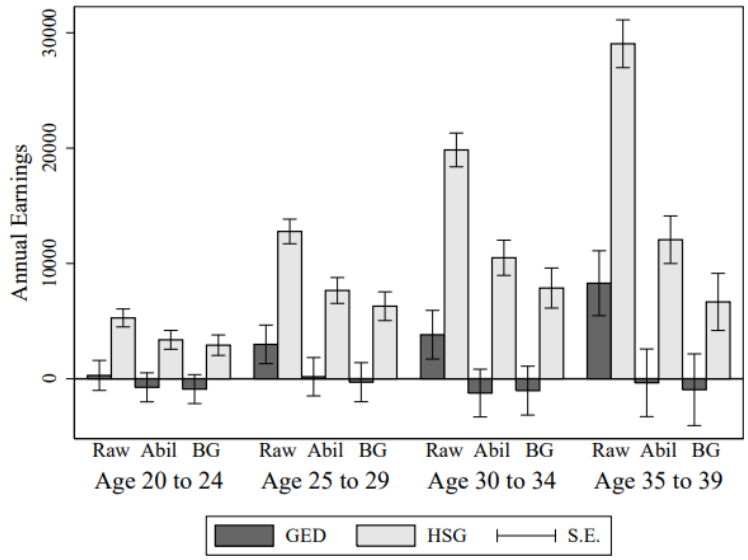
As illustrated in figure 2, high school graduates do earn significantly more than their GED recipient counterparts. This gap seems to widen as people age. When only age, race and region of residence are controlled for, the annual earnings profiles of GED recipients and high school graduates are greater than other dropouts, whose annual earnings mark the 0 on the Y-axis. However, after cognitive ability (AFQT) is accounted for, GED recipients are indistinguishable from dropouts, whereas high school graduates have higher annual earnings. Regressions on hourly wages yield similar results as well.

Even though most of the patterns found for women are in parallel with presented male ones, interesting differences can be found. As mentioned above, the GED recipient women share, to a great degree, cognitive and non-cognitive profiles of their male counterparts. Similarly, for males, after controlling for ability, no difference in hourly wages exists. However, unlike men, GED receiving women have higher annual earnings compared to high school dropouts, driven by their greater level of labour force participation.

On top of this, different groups can be distinguished among the GED receiving women. Approximately 40 per cent of GED receiving women drop out of high school to have a child. This group of GED receivers displays an identical level of cognitive ability compared to other GED recipients, yet their average participation levels in risky behaviours, excluding sex, are lower. This invites an investigation into the dynamics within the women subsample. Figure 3 similarly portrays these dynamics to figure 2, where the Y-axis represents the earnings levels

among the women high school dropouts. These regressions include the aforementioned BG controls after additionally adjusting for risky and criminal behaviour.

Figure 2 Labor Market Differences, Ages 20–39 (Males, All Levels of Postsecondary Education) from Figure 5.6 Heckman et al. (2014)



Controls: "Raw" – age, region of residence, year and race; "Abil"- raw controls and AFQT adjusted for schooling at the time of test; "BG" -ability controls, broken home status at age 14, family income in 1979, mother's highest grade completed, urban residence at age 14, residence in the South at age 14, and factors based on adolescent risky behaviour and criminal behaviour.

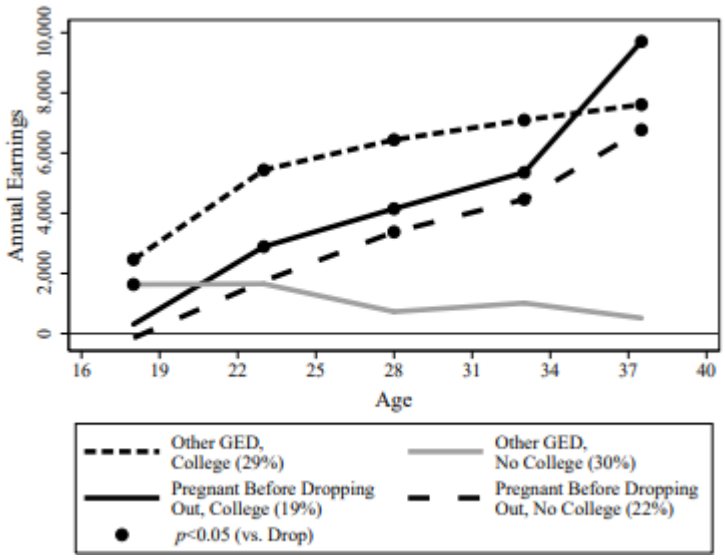
Notes: All regressions allow for clustered standard errors at the individual level.

In figure 3, GED recipient women are divided into four categories: GED recipients who are pregnant before dropping out of high school, those who attend college, at any time before age 40 (thick line), GED recipients who are pregnant before dropping out of high school who do not attend college, at any time before age 40 (loose dash line), other GED recipients who attend college at some time before 40 (close-knit dash line) and all other GED recipient women (grey line). Great trend differences emerge between other GED recipients with no college attendance and the rest.

This graph illustrates how women who receive a GED and women who drop out of high school due to pregnancy benefit from the GED certification, whereas the remaining women acquire similar earnings to high school dropouts. Heckman et al. (2014) hypothesise that these differences might arise due to inherent uncontrolled character differences between the women

who drop out because of pregnancy versus women who drop out for any other reason. On the other hand, they continue, the birth of a child might change the preferences in life among the women subsamples

Figure 3 Annual Earnings by Type of Female GED Recipient (All Races, Background and Ability-Adjusted) from Figure 5.45 Heckman et al. (2014)



Controls: Age, region of residence, year, race, AFQT adjusted for schooling at the time of test, broken home status at age 14, family income in 1979, mother's highest grade completed, urban residence at age 14, residence in the South at age 14, and factors based on adolescent risky behaviour and criminal behaviour

Notes: Respondents are classified as GED recipients if they earn a GED before the age of 40. The sample excludes people once they have been to jail. Regressions exclude those who report earning more than $300,000 (2005$). All regressions allow for heteroskedastic errors and, when appropriate, clustering at the individual level.

However, from the evidence at hand, it is impossible to identify the mechanism through which this association might take place. This association may be due to a selection effect, or it may reflect a causal effect of certification. Despite this, what can be noted is the following: for women, obtaining a GED seems to coincide with their decision to enter the labour force, which increases their annual earnings, even though, their hourly wage remains unchanged.

To conclude, on the other hand, GED recipients possess higher cognitive ability than high school dropouts and are seemingly comparable to a representative sample of high school graduates. However, their non-cognitive traits and risky behaviour outcomes are similar or

even more harmful than high school dropouts. Labour market outcomes between male GED recipients and high school dropouts are nearly indistinguishable after the cognitive ability has been accounted for. GED receiving women have higher annual earnings than high school dropouts rather due to higher labour market participation rates than higher hourly wages. This association highlights the role and relationship between non-cognitive traits and labour market outcomes.

3.3 Evidence from a laboratory experiment

Previously in chapter 2, we have explored the common findings within the literature from the meta-analysis point of view, and with a quasi-natural experiment displayed with a GED sample. To add to the scope of the analysis, we will examine the first experimental study directly aimed to unbundle the relationship between personality traits and labour market outcomes. This intriguing laboratory experiment was conducted by Maria Cubel et al. (2016), where 359 voluntaries chose university students from Wales were assigned to participate in a labour productivity assignment, where the relationship between their personality and productivity was examined. The experiment consisted of five stages: instructions, the performance of the task, break, the performance of the same task and demographic and Big Five questionnaires. Instructions were given orally to simulate employer-employee work environment hierarchy and to evade emotional responses to the instructor's characteristics. Interaction between participants was not allowed during either the task or the break. In this experiment a 44-item (question) Big Five Inventory was presented to participants at the end of the experiment, able to measure all factors.

The task itself was to answer as many additions of five 2-digit random numbers as possible in 10 minutes. Once an answer was submitted, it could not be changed, and the next sum showed up immediately. This task was chosen because on the other hand measures productivity as a function of both cognitive and non-cognitive abilities, while it does not require high levels of the former, and yet, the latter in the form of perseverance, focus and determination are needed, similarly to many occupations. Furthermore, there exist no gender differences in arithmetic or algebra solving.

Included in the experiment was a payoff scheme, where participants would receive 20 experimental dollars (0.35 USD$) for a right answer and lose 4 experimental dollars (0.07

USD$) for a wrong answer that could be later exchanged for real currency. The average participant earned 25.8 USD This payment scheme was explained in the instructions. Individual-level productivity was primarily measured by payments received. In addition, the number of answers and the number of correct sums were also examined. On average, subjects answered 50.5 sums, of which, 45.4 were correct.

Additionally, an estimation of the relationship between personality traits and productivity is derived from ordinary least squares (OLS) regression, where personal characteristics (e.g., gender, university major and family background are accounted for. Despite its simplicity, similar regression functions are within most of the papers in the literature, and hence, in this thesis as well.

Low statistical power, innate to the sample size of this experiment, is accounted for by introducing only parsimonious specifications. The results derived by this OLS are robust to controlling for parental background and average grade in college, the latter of which serves as an imperfect measure of cognitive ability[12]. The effect of personality traits is assumed to be linear, an assumption that in this experiment serves a purpose, but is in contradiction with the psychological consensus.

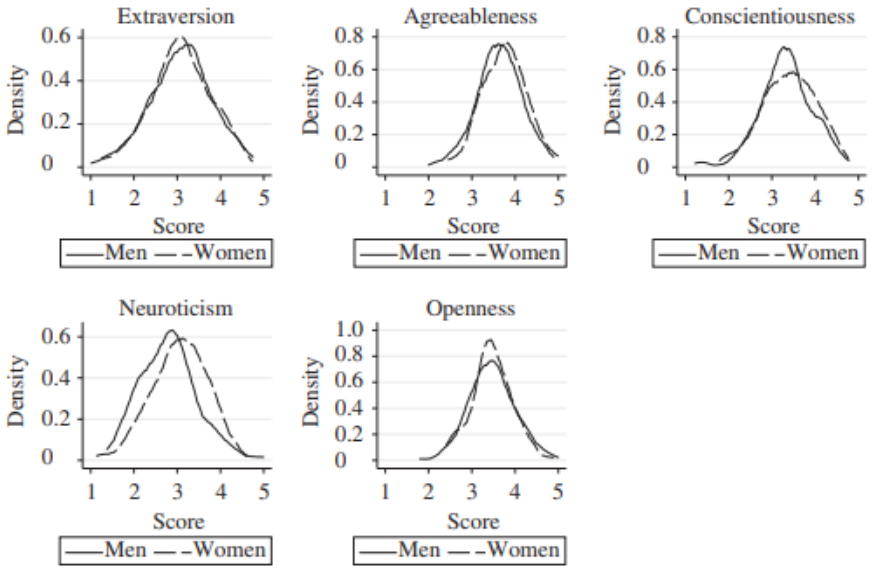
In the baseline estimates of the association of personality on total earnings, the Big Five personality traits are jointly significant. Similarly, to the notions in chapter 1.3., subjects with higher levels of Neuroticism perform significantly worse on this task, and further, an increase of one standard deviation in the level of neuroticism is associated with a 2.9 per cent decrease in task performance. Hence, the study conductors argue, Neuroticism seem to contribute to wage differences through productivity. Conscientiousness, on the other hand, does have a positive association with task performance. An increase of one standard deviation in the levels of conscientiousness is associated with an increase of 2.6 per cent in earnings. The association for the remaining factors is not statistically significant. Yet, on a closer look, very high and very low levels of Agreeableness, experienced more often in the female sample (Figure 4),

[12] As discussed in chapter 1.3 collage GPA is farm from optimal IQ test, heavily influenced by personality traits, especially Conscientiousness, which can potentially introduce the problems related to omitted variable bias. However, Cubel et al. argue with substantial empirical evidence, that this should not denote their findings.

seem to be detrimental for performance. The finding is already in contradiction with the personality's linear effect assumption used in the OLS regression.

Overall, men and women do not differ in the performance of the task. What is notable, however, as illustrated in figure 4, significant gender differences are found in the distribution of Neuroticism and Agreeableness. This finding is consistent with the earlier literature, which invites a more comprehensive exploration into the, potentially gender-specific, relationship between personality and performance.

Figure 4 Personality Traits Density Distribution by Gender from Figure 1 Cubel et al. (2016)



Surprisingly, Neuroticism and Agreeableness seem not to impact male or female productivity differently, whereas this difference in impact for other factors seems to exist. Particularly, a standard deviation increase in Extraversion is associated with a 4 per cent *increase* in earnings for men and a 3.5 per cent *decrease* for women. At first sight, this finding is peculiar since the task at hand does not account for Extraversion in any shape or form. However, this is not a novel finding within the literature. The factor Extraversion includes facets associated with ambition (assertiveness and activity) that are, on average, more common among men and facets associated with sociability (warmth, gregariousness, and positive emotions), that are, on average, more common among women. In this experiment, similar differences in the self-reported levels of these two sets of facets exist. These differences might drive the heterogenous association of Extraversion on productivity.

Unsurprisingly, given the nature of the task, individuals with science majors perform significantly better compared to other students. This way the chosen major served as another source of heterogeneity Cubel et al. (2016) hypothesize that similarly to occupation selection, unobservable characteristics determining self-selection might exist in the major selection as well, which in turn, might condition how personality influences performance, even when controls are introduced.

The extent of external validity is close to impossible to accurately predict. The artificiality of the task, not present in any real-life occupation, and the use of only undergraduate students as a sample, not representative of the whole population are both undisputable challenges for external validity, open for critique. Despite this, the laboratory experiment offers valuable insight due to its nature. Most of the variables in place can certainly be considered, which is something this branch of literature generally lacks.

3.4 The conclusions of the association-based relationship between personality and earnings

To conclude the findings presented in section 2, psychometric traits are associated with earnings, however, these associations seem to slightly differ between genders. The meta-analysis, by Alderotti et al. (2021), indicated a significant positive association between Consciousness and earnings, even when cognitive ability and the level of education are controlled for. Similarly, Cubel et al. (2016) laboratory experiment, indicated that consciousness had a positive association with task performance. These findings are consistent with the notion expressed by Almlund et al. (Chapter 7, 2011), that the positive effect of Consciousness may not be fully mediated by cognitive skills and education.

On the other hand, Alderotti et al. (2021) argue that the association between earnings and Neuroticism seem to be significant and negative, an observation supported by the laboratory evidence by Cubel et al. (2016). Furthermore, Almlund et al. (Chapter 7, 2011) literature review does also support these associations. The GED analysis by Heckman et al. (2014) does support this notion since GED receivers possess higher cognitive ability compared to other dropouts, while their life outcomes and non-cognitive traits mirror other dropouts, indicating, that non-cognitive abilities are associated with earnings, but the underlying mechanism

remains unknown. Hence, as mentioned above, the mechanism through which these personality traits affect earnings remains a mystery, even though, compelling theories have been laid out by scientists.

It is noteworthy that, Alderotti et al. (2021), Heckman et al. (2014) and Cubel et al. (2016) all report either personality or outcome differences between genders. Alderotti et al. (2016) emphasize that their results indicate that with male-only samples the positive association between Openness and earnings was particularly significant, while the negative association between Neuroticism and earnings was significantly smaller with female-only samples. In addition, as illustrated in figure 4, significant gender differences are found in the distribution of Neuroticism and Agreeableness. This finding is consistent with the current consensus within the psychological literature that the greatest differences in personality between genders can be found in the average level of Neuroticism and Agreeableness. (Weisberg 2011)

Similarly, even though on different personality factors, Cubel et al. (2016) find that a standard deviation increase in Extraversion is associated with a 4 per cent *increase* in earnings for men and a 3.5 per cent *decrease* for women, a peculiar finding considering that the artificial task does not take Extraversion specifically into account. Despite this, they continue, it is not a novel finding within the literature. The factor Extraversion includes facets associated with ambition (assertiveness and activity) that are, on average, more common among men and facets associated with sociability (warmth, gregariousness, and positive emotions), that are, on average, more common among women. Further, in this experiment, there exist similar differences in the self-reported levels of these two sets of facets.

On the other hand, Heckman et al. (2014) find that labour market outcomes between male GED recipients and high school dropouts are nearly indistinguishable after the cognitive ability has been accounted for, whereas GED receiving women have higher annual earnings than high school dropouts rather due to higher labour market participation rates than higher hourly wages.

Together these findings fall fourth a comprehensive investigation into the causal mechanisms behind the association between personality and earnings, while also examining the possibility of different mechanisms between genders. As an answer to this great challenge, section 3

carefully lays out an intervention study analysis claiming to offer a causal relationship between personality and earnings.

4. The Perry Preschool Program

4.1 Background

The original purpose of the Perry Preschool Program study was to answer the question of whether high-quality early education can help to improve both the lives of highly disadvantaged children and the quality of life of the local community. The project started as a local attempt to answer the problem of school failure and delinquency, which were disproportionately largely present in the disadvantaged segment of the school population consisting of mainly African Americans. Only when the program seemed to yield unexpectedly favourable results, did the large-scale extrapolation potential start to obtain broader interest. Later, the Perry study established itself as one of the most influential studies in this body of longitudinal research that permits definite statements regarding the essential value of early childhood education, especially among the more disadvantaged.

The Perry Preschool Program itself was a randomized controlled trial study of 123 African American children with low intellectual ability (IQ) and low economic status families. The study was conducted between the years 1962 and 1967 to explore the long-term effects of attending versus not attending high-quality early childhood education. The participants were three and four-year-old children who lived and were selected from one school district within the Ypsilanti neighbourhood in Michigan. They were randomly assigned to treatment groups, who received childhood education and to control groups that did not receive childhood education. During these five study years, five waves (birth cohorts) of children attended the program, each of which participated for two years after enrolment[13]. The Information about these children was collected and examined rigorously on a wide range of observable characteristics ranging from family demographics, child's ability, attitudes, criminal activity, employment to teenage pregnancies, annually from ages 3 to 15, and later from four follow-up surveys conducted at 19, 27, 40 and 50.

[13] The only exception being the first wave of participants, who entered the program at the age of four and for that reason participated for only one year.

The results derived from the Perry Program have been analysed hundreds of times from countless different perspectives, the earliest of which date back to the 1960s, whereas new interpretations utilizing the modern statistical tools have recently and continue still to arise. Despite its evident study-design challenges[14], which will be at the centre of this section of the thesis, the Program has kept its relevance. This is partly because it has been one of the first longitudinal studies ever created to study the effects of early childhood education, and on the other hand, due to its low attrition rate[15], even at the later stages of the follow-up surveys. On top of this, the findings made from this study population have been consistent regardless of the source data, whether it has been the collection of self-made, parental and teacher surveys, or third-party datasets. In addition, the variables followed in the program are meaningful to society, which is why the topic itself contains great everlasting policy relevance, which has sparked up a constant flow of new study evaluations.

The Perry Program itself aimed to provide high-quality, active learning focused preschool education for children from disadvantaged backgrounds, who possessed subaverage IQs. The African American children who possessed no visible handicaps and had IQs between 60 and 85 were selected. The IQ metric itself is normally distributed, the populational average IQ is 100, hence the IQ of 85 is one standard deviation lower than the populational average IQ. To demonstrate the actual disadvantages derived from low IQ alone, people with an IQ of 60 have a hard time understanding instructions or information that they have read. Thus, the program participants were assumed to be at high risk of non-favourable later life outcomes. In addition, the sole intention of the study was to increase the average IQ of the participants, since, at the time of this research, it was believed that IQ could both be improved during the early years, and it was seen to be the only avenue through which life outcomes could be improved (Heckman 2010).

The intervention mainly focused on enhancing the ability of participants to plan, execute their plans, and reflect on their activities in social groups within a "plan-do-review" sequence. In this method, the children do first plan their activity, then they act on their plans, after which,

[14] As Berrueta-Clement (1984) puts it: "This experimental design was created during a simpler time: We had not yet learned how difficult experimental designs are in field research."
[15] The follow-up attrition rates ranged between 0 to 10%. 4 control group and 2 treatment group participants died before the age 40 follow-up.

they are required to reflect on their actions both themselves and with the help of a teacher. In addition, children and adults were treated as equal partners in the learning process. Similarly, social skill abilities including cooperation with others and conflict resolution were practised. The emphasis was, in this sense, placed upon the social realm and one's ability to conduct themselves adequately, instead of teaching the participants a particular substance. Within the literature, equivalent teaching principles are referred to as active participatory learning programs. However, in this thesis, these methods will be referred to as the focus on non-cognitive skills.

For teachers to be able to execute the teaching according to this new emphasis in the pursued early education, they received extensive managerial supervision and in-service training. The first year of teaching was carried out in a free-flowing manner, where the teaching schedules were loose, whereas the later years started to follow a more strictly scheduled routine. Hence, yet again, the treatment effect measure can be seen partly exposed to deviation within itself, leaving it vulnerable to capture of something else than the initial treatment effect.

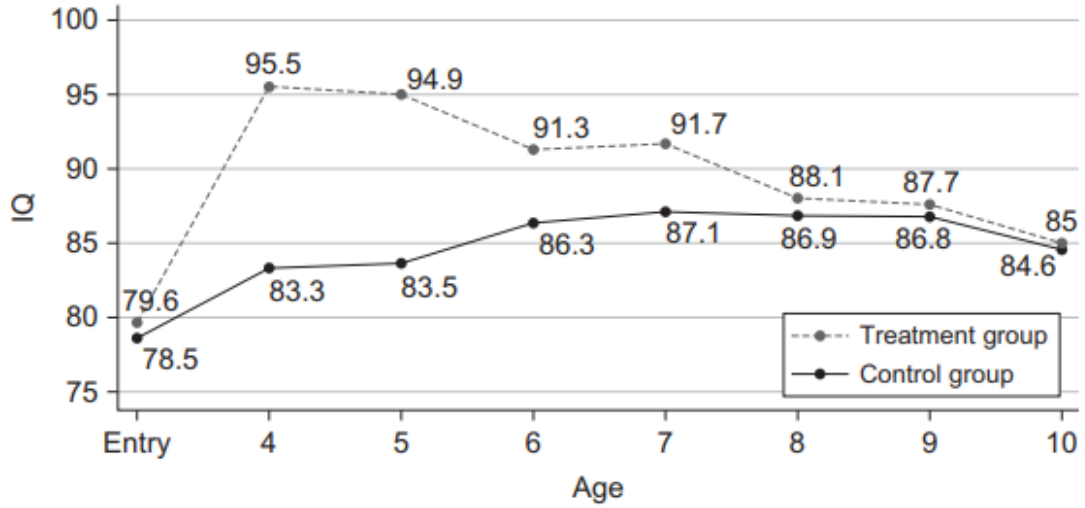
In addition, to the active early education participation, the intervention included weekly 1,5-hour teacher home visits to promote child-parent interaction. To ensure this layer of intervention, employed single mother's children were excluded from the treatment group. Thus, the maternal employment rate creates a difference between the control and treatment groups. However, when these maternal employment rates are studied in the follow-ups, the differences have disappeared, indicating that this difference was not permanent but rather temporary. There still exists a possibility that the differences in the study outcomes could partly be explained by the transitory differences in maternal employment rates. Nevertheless, statistical analyses by Berrueta-Clement (1984) showcase that this has not been the case, a notion widely accepted in later studies.

4.2 The Perry Program's early results

The Perry Preschool treatment group participants experienced significant benefits from the program on a wide range of life outcome metrics. They experienced increased levels of educational attainment, employment and earnings, while at the same time the levels of criminal activity and teenage pregnancies significantly decreased. The main avenue through which these improved life outcomes were assumed, at the time of the research design, to stem

from an increased level of IQ. Similar assumptions underlie the literature on the economics of education (Almlund Chapter 8, 2011). The Perry Program provides results that are challenging this view. Even though, the program managed to increase the treatment group's IQs initially, this treatment effect quickly faded away, as shown in figure 5.

Figure 5 Perry Preschool Program: IQ, by Age and Treatment Group from Figure 14A Cunha et al. (2006)



Notes: IQ measured on the Stanford–Binet Intelligence Scale. The test was administered at program entry and at each of the ages indicated.

The vertical axis, in figure 5, represents IQ, and the horizontal axis represents age. During the first year of the program, the IQ levels of the treatment group significantly increase. Surprisingly, however, when the participant's age progresses towards 10, the average IQs of the treatment group (dotted line) and control group (straight line) do converge. Solely for this reason, the whole study was described, by many scientists, to be a complete failure since it failed to increase the average IQ of the treatment group – the very metric the study was designed to increase.

Despite this supposed initial failure, the study continued, and the participants were followed into their 50s. Even though the IQ levels remained similar, participation in a treatment group had positively altered life outcomes, ranging from increased educational attainment and wage income to lesser participation in criminal activity and fewer teenage pregnancies. These improvements sustained well into adulthood and were captured through follow-up interviews

conducted at the ages of 15, 19, 27, 40 and 50. Hence, Heckman et al (2010), concluded that the intervention changed something other than IQ and that something produced strong treatment effects. In the next chapter, the treatment effect on labour market outcomes will be examined.

4.3 Data

The number of Perry Preschool Program participants is only 123. Heckman et al. (2009b, 2010) do systematically address several important statistical issues that arise from analysing the Perry data including its small sample size. Heckman convincingly argues that concerns over the small sample size are unfounded, an argument that will be fine-grainy examined. This powerful statement proposed by Heckman, however, requires various statistical methods, which all require a large set of assumptions. These assumptions and compromises will be discussed later in chapter 6.

Fortunately, the Perry program conducts follow-up interviews, in which, the attrition rate is low. 90 per cent of the original sample participated in an interview at the age of 40. Yet, the analysis faces challenges with inconsistent (earnings[16]) information. To tackle this problem, information on the Perry participants is gathered from two sources: subject interviews and external source data. The former data is collected from conducted surveys that were conducted annually from ages 3 to 15, and later from three follow-up surveys conducted at 19, 27, 40 and 50. The latter route of information is utilized to gather employment spell[17] length, timing, and earnings data but it suffers from data quality issues, caused by missing data points.

These data quality issues are answered by using the latter data source by replacing the missing values with surrogate estimate values, which are gathered through several imputation methods. In this way, the data quality issues created by missing values can be accounted for. The rationale for using imputation methods is described below. These imputation methods,

[16] Data quality issues stem from two sources: 1) Incomplete earnings information and 2) the correct censoring of employment spells since this information can be hard to capture when a participant does not participate in one of the follow-up surveys.
[17] Employment spell refers to the time between when a person starts and end date of an employment.

use the National Longitudinal Survey of Youth 1979 (NLSY79) dataset[18], which is similar to the Perry Program dataset, with a larger variety of variables.

4.4 Imputation

Imputation of the earnings data is crucial to the validity of the results because, in the original survey, job histories were determined retrospectively only for a fixed number of previous job spells. These earnings data inconsistencies require great emphasis, even though the attrition levels are low, due to the small sample size. For these reasons, the missing labour spells were imputed by utilizing econometric techniques.

4.4.1 Proxy creation

Before the imputation techniques can be used, a comparison group to the Perry sample is created from the NLSY79 dataset. This comparison group is created to facilitate interpolation – the estimation of the value of a function between two points (here Perry Sample, and NLSY79 comparison group) which it has tabulated. The creation of this comparison group can be justified since both longitudinal studies contain similar attributes. Both longitudinal studies are representing nearly identical birth cohorts, Perry sample includes birth years 1957-1962, whereas NLSY79 includes birth years 1956-1964. Although the NLSY79 lacks sufficient data to perfectly simulate the Perry program eligibility criteria, the comparison group can be still created from the NLSY79 dataset, yet it requires several assumptions to be eligible for further analysis.

First, the NLSY79 dataset proxy's person's ability with an Armed Forces Qualification Test (AFQT) test scores, however, it lacks IQ scores that are used in the Perry program as eligibility criteria. The AFQT test is an achievement test, not an IQ test, which raises problems of comparability. An IQ test is clinically proven to be the most accurate test on capturing a person's pure intellectual ability, leaving only a little room for non-cognitive functions to intervene. Whereas, achievement tests can mainly measure intellectual ability, leaving more room for non-cognitive functions to intervene. The correlation between IQ and

[18] The NLSY79 is a longitudinal project following 12 686 American youth born between 1957-64.

AFQT is around 0,8, according to Murray (2002). Most of this difference can be seen to be the product of non-cognitive functions.

Despite this limitation, the proxy for IQ is constructed from the AFQT data. This can econometrically be justified if achievement test scores and ability are highly correlated, which in this instance they are. To create the closest possible proxy from the AFQT data set, Heckman et al. (2010), adjust AFQT scores with age and educational level at the time of test-taking in their analysis, which follows a procedure of Carneiro et al. (2005). This adjusted score serves as a proxy for ability, which will be later utilized.

Second, the SES index in the Perry study was created by using the number of rooms in respondents' dwellings at the age of 3. This information is not available on NLSY79. To tackle this problem, Heckman et al, create a proxy for the number of rooms in the respondent's dwelling at the age of 3 by regressing the number of rooms in the Perry data set on observable characteristics such as family size to estimate a linear relationship. Further, this estimation is then used to predict the number of rooms for each NLSY79 respondent, while assuming a similar distribution of observable characterises and the corresponding number of rooms at one's dwelling at the age of 3. This estimate will then, in turn, be used to create a proxy for the SES index.

It is impossible to accurately predict how well these observable characteristics do predict the number of rooms a person enjoys at the age of 3. Nonetheless, in Berrueta-Clement (1984), the dwelling conditions in Ypsilanti are described as "one of the worst congested slum dwelling areas in the State of Michigan". This would suggest that the dwelling conditions are at least likely not to be worse in the comparison group compared to Perry participants. In addition to this, at the times before and during the Perry Preschool program, African Americans were not able to receive bank loans to improve their dwelling conditions, even if they were professionals. However, this has likely to be the case in the rest of the United States as well. Hence, the creation of this proxy can be justified.

By utilizing these two proxies for IQ and the number of rooms in one's dwelling at the age of 3, the NLSY79 comparison group can be created. To make this comparison group more representative of the Perry sample, an NLSY79 subsample is created from the NLSY79 comparison group. In this subsample, only African Americans with median or lower IQs, as

proxied with the adjusted AFQT scores, are taken into account. Alongside, the IQ proxy, broadly comparable observable characteristics, similar to the ones used in the Perry sample, are utilized. Furthermore, this subsample is used to impute the missing earnings of the Perry subjects.

4.4.2 Imputation methods

Heckman et al. (2010) argue that for the control group the imputation can be justified without additional requirements, whereas the imputation for the treatment group can only be justified if the following two assumptions are met. First, if the used observables characteristics can capture the treatment effects, then matching between the treatment group and comparison group, selected using these observables, can be justified once we control for these observables. Second, this approach requires us to operate under the null hypothesis of no treatment effect. If both assumptions are met, as Heckman et al. (2010) do assume in their analysis, the imputation can be justified. The control group and comparison group can both be assumed to be similar in the light that they both represent African American people with subaverage IQs, in addition, the observable characteristics do align as described above. Hence, the imputation procedure can be justified without any additional requirements.

When the imputation has been theoretically justified, then several imputation methods will be utilized to make imputation as widely inclusive as possible. These differing imputation methods do operate under differing methodological logic, which is why a collection of these methods provides us with the most inclusive outcome. In Heckman et al. (2010), four imputation methods are used. The imputation results derived by Belfield et al. (2006) and the comparison group results are presented to create a firmer reference point and to bring in earlier findings from the literature. In the following, these four different imputation methods will be discussed in detail.

The first imputation method utilizes Piece-wise linear interpolation, similarly to Belfield (2006), who utilised the same approach. The Piece-wise linear interpolation works in the following manner: Suppose that we would have observations of a person's earnings, in the Perry study, from the age of 18 (data point A) and 40 (data point C) but not from the age 27 (data point B). Piece-wise linear interpolation would then weigh (by the distance of age) the average of these nearest observed data points (A and C) around a missing value (B). By

completing this process, a value would be received that would be later imputed, which would serve as an estimated value.

The Belfield (2006) interpolation differs from Heckman (2010) by assessing the value of 0[19], for those individuals who are voluntarily not participating in the labour market. Whereas, Heckman (2010), do similarly assess their values to the rest of the participants. Belfield acknowledges that the imputation of the value 0 can alter the numerical estimates they propose, which is why their results derived from their Piece-wise linear interpolation are likely to be downward biased. On top of this, imputing the value 0 can be problematic in an individual-level analysis. Hence, in the light of this thesis, the assumption presented by Heckman can be assumed to be more relevant.

The second imputation method regarded as cross-section regression imputation estimates Mincer earning functions to each gender-age cross-section of the NLSY79 African American subsample, which extraction was discussed above. Mincer earnings function is a single-equation model that explains earnings as a function of schooling and working experience. Schooling is divided into five categories: high school drop-out, GED certificate, high school graduate, 2-year college graduate, and 4-year college graduate. Whereas, the measure of work experience uses actual working experience, in years. Thus, the schooling metric works through more broad categories, where work experience is examined through a more flexible framework. Further, in the regressions, run to each gender and age group, education, working experience and the square of working experience are used as regressors. These regression driven estimations are then utilized to impute missing earnings information.

One should note, however, that these regressions, run by Heckman, do display notable standard errors and large p-values (0,1 – 0,7), implying a potential threat to statistical significance. (Heckman et al. (2010) Web Appendix Table G.1 and G.2) In fact, in most of the estimates presented, the standard error equals over half of the corresponding coefficient. To address and correct this statistical issue of significant standard error, a prediction error is

[19] Belfield explains that the 0 value was chosen in part because it is conservative. In addition, for their research question they were more interested about the taxpayer analysis, for which this assumption, according to them, is justified.

added to the regression estimates. With this adjustment, the aim is not to alter the mean estimates but rather to affect their variances.

The third imputation method utilizes the Kernel matching method. In this method, each of the Perry participants is matched with the closest similar subjects found in the NLSY79 subsample. The Kernel matching process follows the standard matching processes, where distance measures are utilized. Here the distance is estimated for each pair of Perry and NLSY79 subsample subjects, but with different weights assigned by a distance measure. Estimates gathered from this process are then imputed to represent the missing earnings information on the Perry sample.

The fourth imputation method utilizes the Hause procedure, where a person's earnings can be examined in a dynamic manner, where a person's individual (year-to-year) growth rate can be included in the lifecycle earnings estimations. This procedure decomposes individual earnings processes into an individual's observed abilities, unobservable level of abilities and growth terms, together with serially correlated shocks. In other words, on top of observed abilities, the Hause procedure considers a variables current value given its past values. Hence, if a person has experienced, e.g., unemployment or has served a prison sentence, this possible path dependence can be accounted for. After the creation of a dynamic earnings function and the establishment of individual parameters, the predicted value can be obtained, which is then utilized to impute missing earnings information.

After the four imputation methods are discussed a short comparison of these methods should take place. The first three methods can be described to be conservative in so far as the differences between the treatment and control groups tend to be eliminated because they are imposed with the same earnings dynamics structure. In addition to the imposition of the same earnings dynamics structure, the fourth method also accounts for differences in the unobservable variables as well, implying a more comprehensive imputation strategy.

Table 2 Earnings with different imputation methods, Modified from Table G.5 Heckman et al. (2010) Web Appendix

Imputation	Age 19 - 27		Age 28 - 40	
	Control	Treatment	Control	Treatment
Linear Interpolation	110,839	102,793	294,622	424,764
Cross-Sectional Regression	108,920	111,651	215,464	294,211
Kernel Matching	185,239	186,923	287,290	370,722
Hause (1980)	250,992	269,418	355,526	470,969
Belfield et al. (2006)	216,353	241,501	343,611	445,923

Notes: The values represent the cumulative average earnings of a person from both the treatment and control group and from the ages of 27 and 40 in 2010 USD. For comparison, the corresponding cumulative earnings for a NLSY79 comparison group are 137,500 at the age of 27, and 325,500 at the age of 40. Discount rate is set at 0 percent; Missing data in employment is imputed by the group average; "Cross-Sectional Regression" and "Hause (1980)" series are obtained from yearly datasets, so they are not directly comparable with other series.

4.4.3 Imputation results on earnings estimates

Before the further analysis of table 2, it should be noted that earnings estimates are adjusted for employment-related fringe benefits[20]. because the missing earning information is imputed from the NLSY79 data also for the treatment group, the true earnings for the treatment group, are likely to be understated or underestimated (Heckman et al. 2010). This holds if a treatment

[20] The data is gathered from the Bureau of Labor Statistics Report on Employer Costs for Employee Compensation.

effect is present. For this reason, the estimates are likely to be partially downward biased. In addition, Heckman et al. (2010) use a discount rate of 0 per cent, whereas Belfield et al. (2006) do use a discount rate of 3 per cent. Implementing a higher discount rate will, in effect, reduce the magnitude of the estimations. Thus, the reported results in table 2, are likely to be, in this aspect, upward biased, compared to the imputation used by Belfield et al. (2006). However, these different studies contain other different assumptions, apart from imputation, leading to differing results. Below is the discussion of the results derived from table 2.

In Table 2, earning estimations from different imputation methods are compared within the Perry male sample. A similar table for women can be found in the Heckman et al. (2010) Web Appendix Table G.4. In this thesis, I'm focusing on the control and treatment group differences between the age 19 and 40 follow-ups and leave the extrapolated earnings from 41 to 65 outside the range of my analysis. Hence, the extrapolation and lifetime difference rows are not addressed. On top of that, the Cross-sectional regression and Hause (1980) series are based on yearly earnings, whereas the rest are based on monthly earnings. Thus, they are not directly comparable. In addition to the four imputation methods, the Belfield et al. (2006) row is added to describe their findings and to provide an alternative reference point. The NLSY79 comparison group refers to a low ability subsample, defined as adjusted AFQT scores below the African American median, which serves a similar purpose.

For all the used imputation methods, both Heckman et al. (2010) and Belfield et al. (2006) show significant treatment effects on the level of Perry participants earnings, especially at a later age. However, both the magnitude of this effect and the absolute levels of earnings do differ between the methods used above. In the male sample, the earnings between the ages of 19 and 27 follow-ups are quite similar in both the control and treatment groups. Even though the treatment group earnings are slightly greater on all other methods than on linear interpolation, where the control group's initial earning level is slightly greater compared to the treatment group, the effects do remain under the reach of standard error. This leaves us not able to differentiate the seen differences from zero effect.

On the other hand, in the female sample, the initial earning levels between the age of 19 and 27 follow-ups do significantly differ between the control and treatment groups. The difference in initial earnings between control and treatment groups is as high as 75 per cent, with the

Hause method. One explanation for this initial earning difference can be found in the higher number of teenage pregnancies among the control group, 1,2 pregnancies per female, compared to 0,6 pregnancies per female in the treatment group (Berrueta-Clement 1984). Pregnancies, and the fact of having children, obviously limit one's ability to earn, at least in the short run.

When the earnings are taken a closer look at the age 40 follow-up, the relative differences between the control and treatment groups increase substantially and become significant. In both the male and female participant subsamples, all the imputation methods showcase a parallel difference between the control and treatment groups earning levels. These differences differ from slightly under 30 per cent to almost 45 per cent advantage in favour of the treatment group.

The cross-section regression imputation, which creates a Mincer earning function explaining earnings as a function of schooling and experience, as discussed above, indicates a greater difference between the control and treatment group for the female subsample compared to the male counterpart. This finding is in line with the discussion above, which highlighted how, for women, the observed increase in earnings could be greatly attributed to the increase in educational attainment. In the cross-sectional regression imputation method, the role of educational attainment is emphasized more heavily compared to the other ones, which, in turn, highlights the role of the female subsample increase in educational attainment. Here, a similar treatment effect increases on the earning levels, in the male subsample, is not observed.

Overall, Table 2, visualizes the difference in the earnings levels between the control and treatment groups. For males, no significant treatment effect can be observed before the age of 28, whereas for women a positive treatment effect on earnings can be observed. However, after the age of 28, significant treatment effects can be seen in both male and female subsamples. These findings strengthen the notion proposed by Schweinhart et al. (2006) and by Heckman et al. (2009b,2010), that the magnitude of the positive treatment effects increases as the treatment participants age.

Additionally, these results demonstrate the role of imputation methods and the assumptions they rely upon. In other words, even if all the imputation methods offer parallel effects, the

absolute and relative effect within each method is greatly affected by the embedded assumptions. Hence, understanding the underlying assumptions is crucial, especially when examining causal evidence.

4.5 The treatment effect on labour market outcomes

Several Perry Program studies, Rolnick & Grunewald (2003), Schweinhart et al. (1993, 2005), Belfield et al. (2006), Heckman et al. (2007,2010) and Pinto et al. (2013), have indicated significant increases in both employment rates and earnings treatment effects among the treatment group participants. The estimated level of the treatment effect slightly varies between studies. Schweinhart et al. (2005) present lifetime earnings to be 10 to 15 per cent higher in the treatment group, whereas the Heckman et al. (2010) presents lifetime earnings estimates, on multiple slightly differing methods, to be 10 or even 35 per cent higher than the control group counterpart. However, these estimations are leaning in the same direction. Hence, the debates regarding the relevancy of the treatment effect are focused on the magnitude of the treatment effect rather than on whether the treatment effect exists.

Even though the treatment effect on earnings can be quickly expressed, one should pay attention to the underlying explanations of why these differences might occur. For this reason, the differences observed between the control and the treatment group participants on labour market outcomes will be discussed from the points of view of employment level, earnings level, and criminal activity level – the descriptive standard within the literature.

4.5.1 Criminal activity

Criminal activity levels and the following rates of arrests, convictions and the time being incarcerated greatly affects one's labour market outcomes through its direct (time being incarcerated) and indirect (the stigma associated with a felony) effects. Additionally, the evident existence of path dependence naturally amplifies these effects. The discussion regarding criminal activity is especially relevant for the Perry Program since the program was initially created to address the high level of delinquency present in the highly disadvantaged proportion of the Ypsilanti community. Against this background, one can expect noticeable

higher levels of criminal activity among the study participants compared to the normal population.

In the 50-year follow-up (Heckman 2019), the control group of males can be found to be arrested at statistically significantly higher levels for crimes classified as property, drug-related and especially arrests for violent misdemeanours. These effects on arrests do also translate into effects on convictions. A male control group participant is arrested, on average, 3.1 times, and convicted 2.5 times. Whereas a male in the treatment group averages 1.6 and 0.9, respectively. These are extraordinarily high numbers in both the control and the treatment groups. Especially, when one considers that only one in five violent crimes leads to an arrest. (Heckman 2019) However, these averages are partly obscured by the right tail of the arrest and conviction distributions since a small proportion of the participants are being arrested and convicted numerous times throughout their lives.

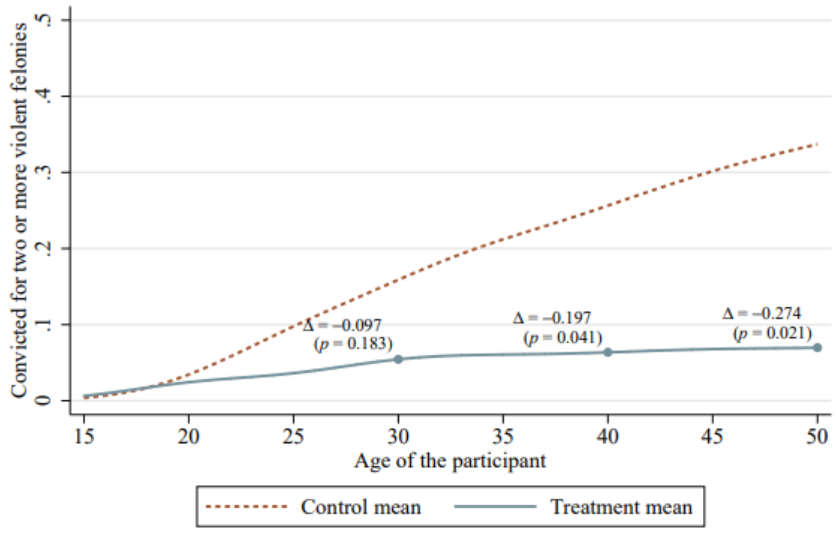
The statistically significant difference between control and treatment groups, i.e., treatment effect, observed in violent crimes levels is the most notable, as illustrated in figure 6 below. I chose this graph to demonstrate the treatment effect on serious violent offences rather than minor violent misdemeanours, to emphasize the treatment's societal impact. On top of that, the selection of two or more convictions for violent felonies limits the obscurity imposed by the right tail of the distribution, present in the averages. Whereas the “two felonies or more” indicator simultaneously indicates a disadvantaged life path on a more serious level, on both personal and societal levels. One should bear in mind that the minimum sentence for a felony in the U.S. is one year’s incarceration, while the maximum penalty is death.

In figure 6, the vertical axis represents the fraction of males to be convicted for two or more violent felonies. Whereas the horizontal axis represents the function of age, where the first age is 15 because that is the first possible age for two violent felonies to be taken place. The blue line represents the fraction of treatment group males convicted for two or more violent felonies as a function of age. The dotted red line represents the development of the control group males, respectively.

By the age of 50, about 33 per cent of the men in the control group have two or more convictions for violent felonies, while the same is true for only 7 per cent of the men in the treatment group. In addition, the treated men who are sentenced for violent felonies, at any

point, are, on average, sentenced to three fewer years in prison compared to the males in the control group. While the average sentence for the men in a control group is over four years' incarceration. Even though the number of months sentenced for violent felonies is not statistically significant, they do indirectly describe the nature of the crimes committed. These implications are both economically and socially significant, as Heckman (2019) points out.

Figure 6 : Probability of Two or More Violent Criminal Convictions over the Life Course for Males from Figure 7 Heckman (2019)

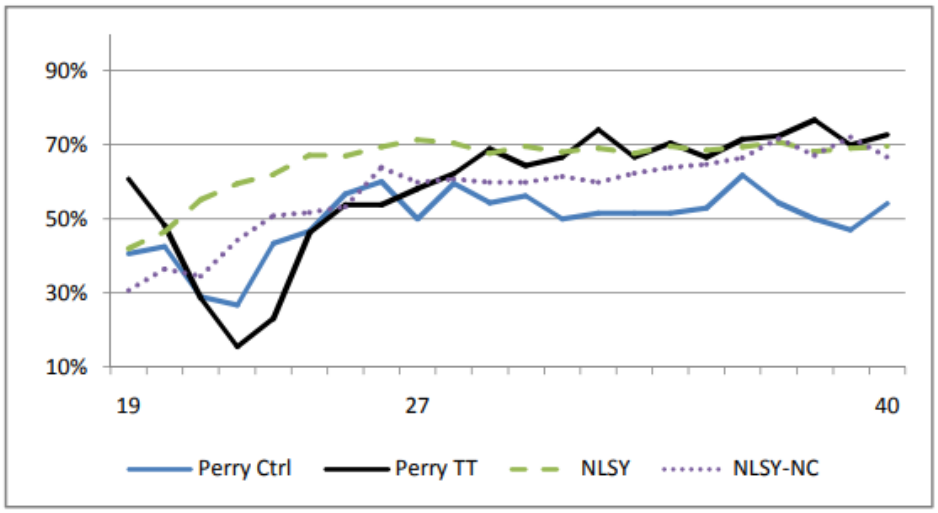


Notes: Note: Δ = augmented inverse probability weighting estimate (AIPW) of the treatment effect; p = worst-case maximum p-value based on approximate randomization test using studentized AIPW.

Hence, this descriptive peak into the criminal activity articulates clearly how the decreased imprisonment rate among the treatment group males has been seen as the single most important metric generating the positive treatment effect in the participating male subgroup. (Schweinhart 2005) For females, on the other hand, the criminal activity rates, and the differences between the control and treatment groups, can be seen as trivial in part because females do commit noticeable fewer crimes in general, Schweinhart continues.

4.5.2 Employment

Figure 7 Proportion of Employed Population: Male, Percentage from Figure E.3 Heckman et al. (2010) Web Appendix



Notes: Missing values in Perry dataset are imputed by the use of the corresponding gender group averages; "NLSY" is a "low-ability" subset of NLSY79 black subsample, where "low-ability" is defined as age and education-adjusted AFQT < the black median; "NC" denotes a subset of all those in the "North-Central" region at age 14.

Figure 7 descriptively illustrates the employment level differences among the treatment and control group participants, against comparison groups. In figure 7 the proportion of employed males, on the vertical axis, is visualized as the function of age on, the horizontal axis. The blue line, in figure 7, represents the Perry control group, the black line represents the Perry treatment group, the dotted green line represents a low-ability subset of the NLSY79 African American subsample and the dotted purple line represents a subset of the NLSY79 subsample living in the "North-Central" region[21] of the U.S. at the age of 14. The two dotted lines are demonstrated in this figure to represent a direct comparison group of the wider population with similar background characteristics. In other words, to give context.

Before the age of 27, treatment group participation does not create a noticeable effect on labour market participation. The effect seems to be even negative, but this can be expected due to the slight increase in the educational attainment level of treatment group participants. However, after the age of 27, the Perry treatment groups labour market participation level

[21] This region is selected to parse out a more comprehensive comparison group to represent the areas near the state of Michigan, where the Perry program took place.

converges to the trajectory of the NLSY comparison group, whereas the Perry control group participants remain to participate in the labour market at a significantly decreased rate. The labour market participation per cent for the Perry program control group is stagnated among the 50 to 60 per cent range, whereas the treatment group and the NLSY comparison groups do attain an employment ratio of 70 per cent. Hence, the effect, on labour market participation is truly noticeable.

This significant difference in labour market participation rates does seem to stand the test of time. However, the 50-year follow-up (Heckman 2019), revealed unexpected results. From the age of 41 to 55 among the males, the control and treatment group employment participation levels are identical, at 43 per cent (Heckman 2019 Web Appendix Table 12). Whereas the female treatment group participants do participate in the labour markets with slightly higher rates, 55 per cent, than the control group, 47 per cent does (Heckman 2019 Web Appendix Table 13). Thus, the great differences in employment levels between the control and treatment groups, present in earlier follow-ups, have disappeared between the ages of 41 and 55.

This finding is in stark contrast with the lifelong positive treatment effect argument conveyed by Schweinhart (2005), Belfield (2006) and Heckman (2007,2010) since the increase in employment levels seems to be temporary rather than permanent. In fact, before the 50-year follow-up, the lifelong positive treatment effect was shared as a wide consensus within the literature, shared also in the future estimations as well. In fact, the future estimations predicted a widening difference between the control and treatment groups (Heckman 2010). Hence, the most recent follow-up findings were surprising, to say the least. Despite this, one should note that the differences in employment levels between the ages of 25 to 40, managed to create a vast contribution difference between the control and treatment groups, which have not been wiped away by this recent trend.

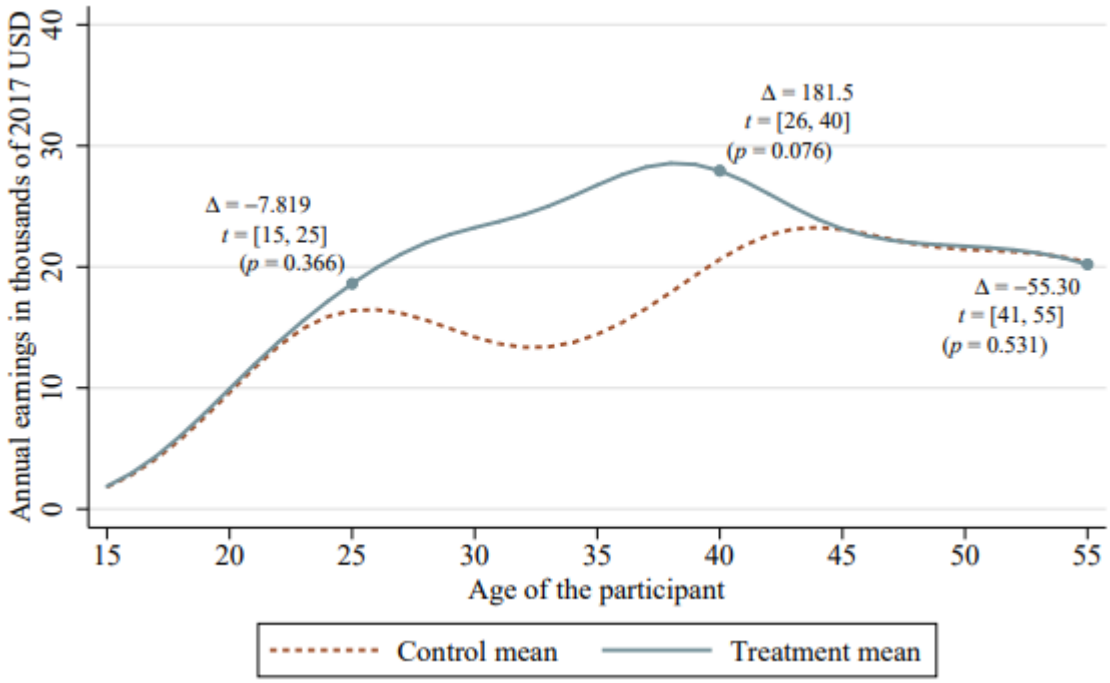
4.5.3 Earnings

Earnings serve as a great proxy for labour market outcomes in both quantitative and qualitative realms, wisdom widely shared among economists. Before the 50-year follow-up, the treatment group's earnings were noticeably and statistically significantly greater than the control group's earnings and they were estimated to grow even greater with age. (Heckman 2010) However, with the newest follow-up (Heckman 2019), the aggregated treatment effect

on the cumulative earnings between ages 15 and 55 (about 120 000$ in 2017 USD) was no longer statistically significant.

Figure 8 below illustrates the progression of male mean annual earnings in thousands as a function of age. The vertical axis represents the mean annual earnings in thousands (in 2017 USD), whereas the horizontal axis represents the function of age. The age metric starts from 15 because no significant earnings are expected from children. The blue line represents the treatment group's mean annual earnings as a function of age. The dotted red line represents the control group respectively.

Figure 8 Mean Annual Earnings over the Life Course for Males from Figure 8 Heckman (2019)



Notes: Δ = augmented inverse probability weighting estimate (AIPW) of the treatment effect on the cumulative annual earnings in the time period t = [a, b], where a and b are starting and ending ages; p = worst-case maximum p-value based on approximate randomization test using studentized AIPW.

The control and treatment groups' average male annual earnings do not significantly differ before the age of 25 or after the age of 40, which may stem from the incarceration of many of the untreated men. Due to the lack of complete prison term data on the exact dates of incarceration, this cannot be firmly concluded, as Heckman (2019) notes. However, the male earnings profiles do greatly differ between the ages of 25 and 40, creating a cumulative

earnings difference of about 180 000$ for the treated males. Although, it cannot be firmly concluded this difference seems to stem from the higher incarceration levels among the control group of males. As the incarcerated males return to the workforce in their late thirties, their annual earnings profiles convert to similar profiles possessed by the treated males, Heckman hypothesises. Interestingly, the male peak annual earnings are higher for the treated than for the control group, and they take place in the late thirties, as can be seen from figure 8 above.

For females, the earnings differences do lean towards the control group from the age of 15 to the age of 40, after which the earnings levels of the control group converge with the treatment group. In this sense, the mean earnings evolution is similar compared to the male subsample. However, the reasons why this difference exists differ, which will be discussed in length below.

Although most of the previous research indicates both significant and positive treatment effects on the earnings outcomes, also concerns have been raised. Anderson (2008) raises concerns, over the consensus regarding the heterogeneity of the treatment effect between men and women. According to Anderson, no consensus exists on whether the Perry Program offers a greater treatment effect to men or women. In this critique, Anderson refers to the finding that differing characteristics contribute to the increased lifetime earnings between men and women. On top of this, Anderson asks whether this ambiguity may stem from the large variety of outcomes tested in each of the studies. In other words, may multiple inferences create over rejection of the null hypothesis?

The first part of Anderson's critique can be seen as adequate since the main avenue through which male positive treatment effect, in the form of increased earnings, seem to occur, is the decreased level of criminal activity (Schweinhard 2005 and Heckman 2007, 2010, 2012 2019). Pinto et al. (2013) do suggest that these differences in criminal activity rates between the control and treatment groups can be explained by a change in the personality of the men subsample. Pinto et al. continue to argue that Perry Program had a substantial positive effect on the rates of aggressive, antisocial, and rule-breaking behaviours, which are all associated with criminal activity. In addition, when pushed further, these changes in measured behaviour are tied to the Agreeableness and Conscientiousness facets within the Big Five personality

framework (Heckman 2012, Pinto 2013). For males, the experienced personality changes seem to be a potential avenue through which these outcome improvements stem from.

For females, as discussed above, the criminal activity rates, and the differences between the control and treatment groups, are trivial in part because females do commit noticeable fewer crimes in general. (Schweinhart 2005) For females, the main driver behind increased earnings seems to be increased educational attainment. Moreover, whereas the increase in educational attainment is evident among the female participants, males receive little or no increase. (Heckman 2010b) The positive effect of educational attainment on earnings is well-established and persistent over time, which was largely assumed to explain the differences seen in the earnings levels in the Perry female subsample. However, the 50-year follow-up, greatly questions the validity of this assumption since the persistence of earnings advantage came to a halt.

On the level of personality, participation in the treatment group increased educational attainment and enhanced academic motivation, both of which do proxy the Openness to Experience facet in the Big Five framework (Heckman et al. 2012, Pinto et al. 2013). Further, a female treatment group participant was one-third as likely to be a high school drop-out, compared to the control group. A similar difference existed within educational attainment levels at the age of 40 as well. A similar effect in increased educational attainment or enhanced academic motivation was not apparent for the men subsample. Thus, Anderson's critique does seem to be accurate, treatment effect heterogeneity does seem to exist between men and women in the Perry Program.

In an attempt to confront the statistical insufficiencies apparent within the previous literature, and to answer the concerns raised by Anderson (2008), Heckman et al. (2010) took a systematic approach to address these issues. These methods, assumptions and weaknesses will be rigorously addressed in the following chapters.

Before Heckman et al (2010) paper, a common practice within the literature was not to consider the compromised randomization in the initial study or to dismiss the problems caused by missing earnings data points. In addition, even though, positive treatment effects, similar to Schweinhart (2005) and Heckman (2007, 2010, 2012, 2019), were found by Rolnick & Grunewald (2003) and Belfield et al. (2006), the statistical insufficiencies were not

taken into account. Also, the standard errors were not reported, which leaves the regressions, and the reader, unable to differentiate whether their reported findings are statistically different from zero.

Therefore, in this thesis, I will dive more in-depth into the analysis, findings and methodology used by Heckman et al (2010) because the statistical problems are clearly stated, and, at least to some extent, addressed. On top of this, the Heckman et al. (2010) analysis combined a diverse set of well-thought-out econometric tools, which opens the door for a more convincing causal interpretation. However, in the following segment of this thesis, I attempt to showcase the full array of assumptions one must accept if these results are to be taken into account.

5. Perry Program validity concerns and result interpolation

The Perry program faces major statistical challenges. These challenges can be by and large divided into two categories: internal challenges and external validity concerns. It should be noted that the internal validity challenges, and statistical issues, were, to a large extent, ignored in existing literature before Heckman et al. (2010), who systematically addressed these issues. This lack of systematic consideration of these statistical challenges casts a shadow over the findings that were reported before the release of the Heckman et al. (2010) paper.

In addition, one of the fundamental challenges this line of research encounters is the inability to distinguish the role each trait or ability plays in certain behaviour. In other words, performance on a single task is dependent on multiple parts of a person's non-cognitive functions, which is why singling out one part of a person's non-cognitive realm has proven to be difficult. This, in turn, creates a fundamental identification problem, that a majority of the papers simply ignore (Heckman 2012). In the following, I will first address the internal validity challenges, after which the external validity concerns will be discussed.

5.1 Internal validity

The first internal validity challenge is directed at the small sample size of 123 Perry participants. In the modern academic tradition, the program would have probably addressed

the complications that arise from a small sample size, even before the study took place. As Berrueta-Clement already in 1984 points out: “This experimental design was created during a simpler time. It was not yet learned how difficult experimental design in field research is”. When the study has been conducted the sample size cannot be altered. Fortunately, the low attrition rate together with comprehensive imputation method implementation addresses the data quality issues, which would otherwise, especially together with a small sample size, oppose a great threat to internal validity.

Furthermore, additional statistical procedures can be utilized to address these small sample size concerns. When tested with the small Perry sample, the process of randomly assigning treatment labels for treatment and control groups, referred to as permutation inference, produces an inference about the null hypothesis of no treatment effect. Similarly, the application of test statistics, only justified in large samples, yields an inference about the null hypothesis of no treatment effect as well. Hence, Heckman et al. (2013, 2009b) argue, that concerns over the small sample size are unfounded.

The second major concern for internal validity stems from the original study randomization, which was partly compromised. Initially, this took place in the following way: Firstly, in an entering cohort, younger siblings were assigned to the same treatment status as their older siblings. The study designers justified the exclusion of younger siblings from the randomization process due to concerns over possible within-family spillover effects that could arise if siblings would be represented in both the treatment and the control groups. Secondly, the remaining participants were ranked by their entry IQ scores. The balancing of IQ, on the other hand, produced imbalances in family background observable characteristics, which were taken into account in the second ‘balancing’ stage. Some participants from the initial assignment were changed to a participant with differing SES scores, but similar IQ scores.

On top of this, a few participants originally assigned to a treatment group, whose mothers were employed, were changed to control group participants whose mothers were not employed during the assignment period. The supervised home visits were regarded to be an essential part of the treatment and enabling this part of the treatment was regarded to be crucial for the study, while at the same time serving as the rationale behind this procedure.

Yet, even after these actions were taken the pre-program observable characteristics regarding family background were partly imbalanced. This potentially opens the avenue for problems because if the baseline characteristics and treatment assignment do correlate with each other, then the assumption of independence between treatment assignment and outcomes might be violated (Heckman et al. 2009b). The assumption of independence, a core assumption in any randomized experiment, could be violated even in the absence of treatment effects. On top of this, Heckman et al., emphasize that even if there exists statistical independence between treatment assignment and baseline variables, this compromised randomization can bring about a biased analysation of the results.

To tackle the imperfect randomization, Heckman et al. (2011) develop and apply a unique framework for inference, in which, the used assumptions are specifically crafted with the existing knowledge regarding how the randomization was conducted and is imperfect. In other words, Heckman et al. (2011), do create a novel situationally customized framework to suit this specific statistical challenge. By using this approach, "a procedure for testing the family of the null hypothesis in which each null hypothesis specifies that the program did not affect one of several outcomes of interest that controls the familywise error rate in finite samples", is established. Statistically significant treatment effects are observed even after the implementation of this approach. Thus, the inclusion of this method contributes to tackling the criticism indicated towards the reliability of the evidence gathered from the Perry program.

These concerns are especially crucial in the context of Perry study since if the analysis does not account for the compromises in the randomization process, the probability of treatment assignment will be miss specified leading to misleading estimates, results and mistake rejection of the null hypothesis also referred as type-1-error.

Lastly, the Perry Preschool curriculum was not fully stable across time. Hence, the actual treatment might have differed between the waves. The curriculum itself was consistent throughout the study years and centred around the framework, where the focus was placed on enhancing the ability of participants to plan, execute their plans, and reflect on their activities in social groups, as described in the background chapter. However, the details and applications of these frameworks did slightly vary from year to year due to their experimental,

and hence, developing nature (Schweinhart et al. 1993). The daily structure of the first years was freer-flowing, whereas the later years were constructed around structured daily routines. This variation in the treatment itself raises potential challenges but, within the literature, they are mostly dismissed due to the trivial magnitude of the differences in the treatment itself.

5.2 Generalizability

Even if the internal validity challenges could be overcome sufficiently, and the randomisation to control and treatment groups would be perfect, the participants would not necessarily represent a randomized sample of the whole population. This intrinsic embeddedness of an overwhelming level of pre-existing background burden has raised concerns over external validity and the generalizability of the results. One should note that the Perry project started as a local attempt to answer the problem of school failure and delinquency, which were disproportionately largely present in the disadvantaged segment of the school population consisting of mainly African Americans (Berrueta-Clement 1984). To ensure this, the selection for the study was limited to only African Americans with subaverage IQs.

As discussed previously, Anderson (2008) raised concerns about the heterogeneity of treatment effects between genders. More recently Xie et al. (2020) have raised concerns over the heterogeneity in the treatment effect between differing SES-level children within the Perry Program. Xie et al. (2020) argue that a reassessment of the Perry study data indicated that the most disadvantaged families experience the most significant and long-lasting positive causal effects for receiving treatment. Further, they argue, that even though experimental intervention programs like Perry Program are widely regarded as an answer to the selection bias problem, this is not necessarily the case. The challenge of the population-wide unrepresentative nature of the Perry Program discussed above, may even be strengthened if the observed treatment effects are heterogeneous. At least, Xie et al. (2020) continue, the generalizability of the results would require strong and untestable assumptions.

In addition to Xie et al. (2020) and Anderson (2008), Schweinhart (2006) raises questions regarding the external validity of the Perry Program, even though, they propose strong treatment effects for the Perry participants. Schweinhart (2006), points out that it is legitimate to ask whether one could assume similar treatment effects from a similar treatment under the

current economic conditions, and could the same results be derived from children with non-disadvantaged backgrounds.

Furthermore, Heckman et al. (2010, 2013), does not talk about the topic of external validity, while at the same time, heavily advocating on the behalf of how Perry Program results should be taken into wider consideration when policy interventions are considered. However, in Heckman et al. (2009b), it is noted that when it comes to assessing external validity, the Perry Program overrepresents the most disadvantaged segment of the African American population of children. Whereas Schweinhart (2006) argues that the effects on more advantaged families cannot be addressed with the data at hand, and thus the results cannot fully be generalized. Moreover, in Heckman et al. (2013), it is noted that based on the results examined, one is unable to determine the external validity of the presented evidence. Hence, external validity and generalizability remain to be a scarcely addressed, yet challenging, issue in the Perry Program.

On top of that, Schweinhart (2006) argues that one cannot distinguish the single component of the Perry Program that produces these significant treatment effects. Thus, if one pursues similar treatment effects, demonstrated in the Perry study in another setting, the full implementation of the program can be seen as necessary.

In addition to these more profound generalizability concerns, a short description of the Ypsilanti community is needed to understand the happenstance in which the Perry Program has operated. The importance of these notions can be seen as important since the missing information of Perry participants is imputed from, and partly compared to, the NLSY79 participants who might not share similar happenstance due to the geographical location of their dwellings. Hence, this could explain some variation in the key observable variables utilized in the later analysis of the study.

Firstly, one should note that as late as the 1950s, African Americans were openly discriminated against in local hiring. According to Berrueta-Clement (1984), it was not until the Civil Rights Act of 1965 that open discrimination in hiring stopped. Up until this point, the Perry Program had run for three years. African Americans have experienced open discrimination during this time all over the U.S., yet the magnitude of the discrimination might differ between the states for example. Hence, it would not be a surprise if the

discrimination would have affected the socio-economic status, employment rates, earnings, and housing opportunities of the African American population differently in Ypsilanti compared to the rest of the U.S. Nevertheless, these concerns have not been, to the best of my knowledge, addressed in the economic analysis literature of the Perry Program, which might indicate that this concern has not been seen as significant enough to be addressed. Yet, it could be addressed in future research due to its potential effect on the generalizability of results.

5.3 The economic returns of the Perry program

In an economic analysis of the Perry program, one must evaluate the economic returns the program yields. Several cost-benefit analyses have been conducted and while the results differ, they do heavily lean in the favour of the program. The earliest cost-benefit analysis was conducted by Barnett (1985) when the study participants were 20-year-olds. The analysis considered the initial two-year preschool program cost[22] (26 200$ in 2017 USD) and the costs associated with increased college attendance against savings in childcare, school cost savings, crime reduction, welfare reduction and earnings increase. The present values were discounted at a 5% rate and were inflation-adjusted.

Already at the age of 19, the program's economic benefits accounted for the full cost of the first year of the program, even though the most important benefit, treatment group participants' higher earnings, had not taken place. These benefits stemmed mainly from savings in elementary and secondary education driven by the more attentive and committed treatment group participation. (Bernett 1985).

More recent cost-benefit analyses have reported high returns on investment. Rolnick & Grunewald (2003) report a 16 per cent rate of return to the Perry program, whereas Belfield et al. (2006) report a 17 per cent rate of return. However, these analyses have faced similar statistical challenges described above and below. Yet, even after these statistical challenges have been to some extent addressed by Heckman (2010), their cost-benefit analysis estimates the overall annual social rate of return to be 7-10 per cent with a 3 per cent discount rate while

[22] The initial cost for two-year program was reported to be 9708$ in 1981 USD per child (Barnett 1985). The costs included both the operating and capital costs.

being statistically significant. Hence, while accounting for the statistical challenges and evaluating the social costs of crime differently does significantly decrease the estimated returns, the returns do remain to be significant and positive. Interestingly, Heckman (2010) does note that the annual rates of return derived from the Perry program do exceed the historical U.S. equity returns.

Further, as the program's benefits to the health and well-being of future generations are not accounted for, the cost-benefit analysis is likely to provide a lower bound on the true rate of return, Heckman (2010) continues. In a recent paper Heckman (2019b) implicates that the Perry program does indeed provide intergenerational externalities. Even though the implications of intergenerational spillover effects do suggest that the earlier cost-benefit analysis might have a positive force yet to be extracted, these findings should be accounted for with caution.

5.4 Result interpolation

Traditionally the mechanism through which the positive personality alterations have been created has been seen to be caused by the initial enhancement of IQ (figure 5). (Almlund Chapter 8, 2011) This initial enhancement of IQ has been theorized to have caused the alteration of personality traits permanently, which has in turn translated into more favourable life outcomes. The role of IQ in this line of thinking has been seen as the moderator enhancing one's understanding of their environment and the promotion of other traits, similarly, leading to reinforcement of skill development. In other words, the initial cognitive gain has translated into non-cognitive gain, leading to a later cognitive gain. (Xie et al. 2020)

Furthermore, when the analysis is limited to the most disadvantaged children among the treated, those who received the greatest IQ scores gained a confidence boost, leading to an increased interest in academic studies, which in turn improved their cognitive gains, Xie et al. (2020) argue. In this sense, the treatment effect heterogeneity worries, earlier presented by Anderson (2008) regarding the treatment effect differences between sexes, do seem to be highly relevant also in the context of the most versus least disadvantaged treatment group participants.

In contrast to this notion portrayed by the traditional school of thought and most recently by Xie, Heckman et al. (2013) propose an alternative explanation for the sources of the program treatment effects. In essence, Heckman et al. (2013) propose that personality alterations (enhancements), discussed in chapter 2.2.3, account for the persistent changes in personality skills that play a substantial part in the creation of the Perry program's success. Further, they estimate the role of enhancements in Externalizing Behaviour and Academic Motivation in producing the Perry treatment effects. The association between Externalizing Behaviour (rates of aggressive, antisocial, and rule-breaking behaviours), and criminal activity is well documented in both psychological and criminological literature.

A reduction in Externalizing Behaviour among the Perry treatment group is especially strong, which is in line with the existing literature. In addition, the Perry program demonstrates a statistically significant treatment effect on Externalizing Behaviour at the 5 per cent level. Hence, Heckman et al. (2013) conclude, that the bulk of the treatment effect of the Perry program on criminal and labour market outcomes is derived from the reduction in Externalizing Behaviour. Although this may route of interpolation may be compelling, Heckman et al. (2013) can not rule out the traditional explanation about the enhanced IQ being the source of the treatment effect. In this sense, this alternative explanation is controversial at best. However, the all-encompassing initially enhanced IQ explanation remains in dispute as well.

5.5 Perry Program Conclusions

Despite the great predictive power of IQ, a large proportion of labour market outcomes is left unexplained even after the role of pure cognitional ability has been taken into account. Further, the previous literature has placed a substantial emphasis on cognitive ability compared to other traits, which has left a huge part of the equation unexplained. Hence, this thesis has focused on the literature on non-cognitive traits and their association with labour market outcomes. As mentioned above, the fundamental reason why the Perry Program has functioned as one of the most influential studies in this branch of literature is that it causally demonstrated how an early education intervention significantly altered life outcomes, while the mean levels of IQ remained unchanged. As Heckman et al. (2010) articulate, the Perry

Program changed something other than IQ and that something created strong treatment effects.

The treatment group participants lifetime earnings estimates experienced an increase of 10 or even 35 per cent. Similarly, the employment rate increased by 20 per cent among the treatment group participants compared to the control group counterparts. It is worth noting that, these positive treatment effects do take different forms for each gender. The criminal activity levels among the male treatment group members decreased significantly, whereas for women this was not seen. The positive treatment effect for women seems to stem from an increased level of educational attainment and from a decreased rate of teenage pregnancies.

Even though the treatment effects have been significant, the Perry Program has received its fair share of criticism regarding both the internal and external validity. The critics have raised concerns over the small sample size, partially compromised randomization process and non-stable treatment, in the form of a changing curriculum within the program. Despite these concerns, a comprehensive and systematic econometric fine-tooling performed by Heckman et al. (2010, 2011, 2013), has largely alleviated concerns over the internal validity.

However, the concerns over external validity persist. The program participant was selected to be low IQ African American individuals with as disadvantaged family backgrounds as possible, who were similarly, at the time of the research, openly discriminated against. Thus, extrapolation of the results into the whole population can be seen as challenging. Further, this population-wide unrepresentative nature of the Perry Program can even be strengthened if the observed treatment effects are heterogeneous as Xie et al. (2020) have recently argued. In addition, the extraction of a single avenue through which the treatment effect takes place is impossible, and hence, to pursue similar treatment effects, the full implementation of the program can be seen as necessary (Schweinhart 2006).

Despite these limitations, the Perry Program has stood the test of time for decades, while simultaneously serving as one of the first causal studies presenting a causal link between non-cognitive traits and labour market outcomes.

6. Discussion

In this thesis, I have tried to examine the relationship between psychometric traits and labour market outcomes from multiple points of view. Intuitively it is obvious that psychometrical traits, do both greatly affect the opportunities open to a person, and hence, influence the choices they make, which in turn, generate experienced life outcomes. The role cognitive ability plays in this relationship is already well-established (Becker 1978). Yet, non-cognitive abilities have not received similar attention within the economic literature, which is why my thesis greatly emphasizes this line of literature over the role of pure cognitive ability.

This inherent imbalance within the literature stems partly from not having sufficient metrics to accurately account for non-cognitive abilities, and when these metrics have existed, their implementation has proven to be especially challenging. However, the challenge set by insufficient metrics has somewhat been lifted, mainly because during the past two decades the Big Five has stabilized its status as the leading personality framework. For this reason, modern individual-level datasets do usually include some metrics for personality, which has made a large sample size of personality available. This has coexisted with the rise of great econometrical efforts made by, e.g., Heckman et al. (2010) and Almlund et al. (2011), that allow the implementation of non-cognitive traits (personality) into economic models. Hence, today, examining the role of non-cognitive abilities at the heart of labour market outcomes is more plausible than before.

The main findings from this literature review are the following: first, non-cognitive traits are associated with a variety of labour market outcomes. Especially the Big Five personality factor Consciousness is constantly associated with significantly better labour outcomes (Almlund Chapter 7, 2011; Cubel 2016), even when cognitive ability and educational level are controlled for (Alderotti 2021), whereas Neuroticism and Agreeableness, on the other hand, are constantly associated with more unfavourable labour market outcomes (Almlund Chapter 7, 2011; Cubel 2016)

Secondly, there seem to exist significant gender differences in both the distribution of non-cognitive, e.g., personality traits, and the effects and associations between certain non-cognitive traits and labour market outcomes. (Cubel 2016, Heckman 2010, 2014) More specifically, as illustrated in figure 4, even within a randomly selected laboratory experiment sample of university students, there exist significant gender differences in the distribution of

Neuroticism and Agreeableness. (Cubel 2016) This finding is consistent with the current consensus within the psychological literature on gender differences in personality. (Weisberg 2011)

Thirdly, Heckman et al. (2010) do present a *causal* relationship between non-cognitive traits and labour market outcomes in their Perry Program analysis. The causal evidence presented in Heckman et al. (2010), shows that participation in the treatment group increased lifetime[23] earnings estimate by 10 or even 35 per cent, while also increasing the employment rate approximately by 20 per cent among the treatment group participants compared to the control group counterparts. Interestingly, there existed no difference between the level of cognitive ability between the control and treatment group participants, whereas the non-cognitive abilities seem to have changed.

On top of the inherent measurement error in personality measurements, as described in chapter 2.3, an additional challenge can arise from the implementation of bold underlying assumptions in the analysis of the current datasets, especially, when predicting trends based on them. As I have attempted to demonstrate, e.g., the various differing imputation methods utilized by Heckman et al. (2010), estimated significantly different earnings in the 40-year follow-up, Table 2, even though, those earnings had already taken place. Moreover, the unanimous consensus among researchers was that as the study participants would age, the perceived differences between the control and treatment groups would not only continue to exist, and most likely increase in accelerating speed. At the time, this was a truly reasonable assumption since all indicators pointed in that direction.

However, at the 50-year follow-up (Heckman et al. (2019), the annual earnings between control and treatment groups had converged (figure 8), and no significant differences existed. Hence, earlier predictions proved to be misinforming, to say the least. This does not make the earlier findings obsolete, but it does highlight the crucial role of inherent measurement error and underlying assumptions in this line of research – especially when it comes to predictions.

Even though the Perry Program has been central to my thesis, it most certainly is not the only longitudinal study studying the relationship between early education and later life outcomes.

[23] Lifetime earnings estimates do include information regarding the age 40 follow-up.

Another well-established longitudinal study, called Project STAR (The Student/Teacher Achievement Ratio), randomly assigned one cohort of 11,571 (6,025 joined from kindergarten and 5,546 between grades one to three) children and their teachers to classes within their schools, 79 schools in total in Tennessee U.S. from 1985 to 1989, from kindergarten to third grade. Some students were assigned to small classes (15 students on average) and others were assigned to large classes (22 students on average). After the third grade, the participating children were returned to regular classes for fourth grade and subsequent years. Literature around the STAR experiment has analysed the role class size, teacher quality, and peers have had on standardized test scores. In a later analysis, similarly to Heckman et al. (2010) for the Perry Program, Chetty et al. (2011) evaluate the long-term impacts of this classroom allocation by linking the experimental data to admirative records of tax returns.

What is interesting is that Chetty et al. (2011) find that the classroom quality[24] a student was assigned to has a significant effect on the end-of-class test scores in the beginning, yet these positive test score effects fade away as students age. A one (within-school) standard deviation increase in kindergarten class quality increased end-of-kindergarten test scores by almost 6.5 per cent. Yet, by fourth grade, these students no longer scored significantly higher on tests. Here, the similarity to the Perry Programs' initial boost on children's IQ levels offers an intriguing parallel. Moreover, Chetty et al. (2011) show that the differences in initial classroom quality re-emerge for plural outcomes 20 years later. Students' random assignment to one standard deviation higher quality (within school) classroom is associated with a three per cent increase in earnings at age 27. In addition, students assigned to higher-quality classes are also significantly more likely to attend college, enrol in higher quality collages and exhibit improvements in the summary index of other outcomes, e.g., greater levels of marriage and lower levels of criminal activity. The found associations resemble, to a great degree, Perry Programs effects. What both of these studies share, however, is the inability to shed light on which factors should be manipulated if one would like to improve adult outcomes.

[24] Class quality is proxied by the average test scores of classmates at the end of kindergarten. Chetty et al. (2011) argue that "end-of-class peer test scores are an omnibus measure of class quality because they capture peer effects, teacher effect, and all other classroom characteristics [common to all within the classroom] that affect test scores".

Nonetheless, the STAR program differs from Perry Program on multiple fronts, e.g., the STAR experiment sample size is significantly greater compared to the Perry one. Further, in STAR, due to the nature of a natural experiment, some students leave and enter the classes during the experiment because of natural migration, whereas no later entries took place in the Perry Program. However, in the STAR experiment, these latecomers were randomly assigned to classrooms. The class quality impacts are similar for students who entered the experiment in grades 1-3, and hence, the STAR program should be viewed as evidence of long-term impacts on early childhood education than mere kindergarten, which is contrary to the Perry Program. Also, in the Perry Program, the participating children were tested yearly on both cognitive and non-cognitive realms, whereas the STAR program only collected noncognitive measures for a subset of STAR students in grades 4 and 8.

These observed differences do call forth a potential explanation, and Chetty et al. (2011) offer noncognitive skills as such since they appear to be correlated with earnings through channels not picked up by subsequent standardized tests, and hence, could explain the fade-out and re-emergence dynamic expressed above. They estimate, by regressing earnings on non-cognitive measures[25], while conditioning with demographic characteristics, that a one percentile increase in noncognitive abilities increases earnings by 101$ (approximately 0,7 per cent of total earnings by age 27). On a broader level of analysis Chetty et al. (2011), estimate that a one standard deviation increases in within school class quality raise earnings by 1,520$ (9.6 per cent) at age 27. Yet, as they emphasize: "this figure includes all potential benefits from an improved classroom environment, including better peers, teachers, and random shocks, and hence is useful primarily for understanding the stakes at play in early childhood education".

I should highlight that the inability to explain the avenue through which non-cognitive traits affect labour market outcomes unites both the association and causal studies. Even though, Heckman et al. (2010) *causally* demonstrated that an early education intervention

[25] However, the noncognitive ability measures are far from perfect. They consist of two data points, the other one being fourth grade a random subset of teacher evaluation of their students on a scale of 1-5 on several behaviour measures. These responses were consolidated into four standardized scales measuring each student's effort, initiative, nonparticipatory behaviour and how the student is seen to "value" the class. In grade 8, math and English teachers were asked to rate of a subset of their students on a similar set of questions, which were again consolidated into the same four standardized scales. Among the 6,025 students who entered Project START in kindergarten and of whom IRS and noncognitive skill data is available (1,671, 28% in fourth grade and 1,780, 30% in grade eight). Later, justifiable concerns over selective attrition have been largely alleviated by a detailed investigation completed by Dee and West (2011).

significantly altered life outcomes, while the mean levels of IQ remained unchanged, the study is unable to pinpoint the mechanism through which these outcome differences arise. The same was the case for Chetty et al. (2011) STAR analysis. Hence, the fundamental challenge this line of research encounters is the inability to distinguish the role each trait or ability plays in certain behaviour. In other words, performance on a single task is dependent on multiple parts of a person's non-cognitive and cognitive functions which is why singling out one part of a person's non-cognitive realm has proven to be difficult. This, in turn, creates a fundamental identification problem, that most of the papers have simply ignored.

This inherent inability to express the mechanism through which non-cognitive abilities are tied to outcomes causes constant debate over the external validity, and especially, the generalizability of a specific study. What has in the past increased the concerns is the fact that the majority of the published papers, especially before the last decade, suffer from relatively low levels of statistical power mainly due to small sample sizes. More recently, as discussed above, the wider inclusion of personality metrics into individual-level datasets has, to a certain extent, slightly alleviated this concern.

On a more qualitative level, the observed gender differences seem to follow the following sequence: *for men*, the increased labour market outcomes in the Perry Program seem to stem from a decreased level of criminal activity, and consequently, a lower rate of incarceration, whereas, *for women*, the increased labour market outcomes seem to stem from increased educational attainment and the decreased rate of teenage pregnancies. (Heckman 2010) The decreased level of criminal activity observed, is also apparent in the STAR program. Moreover, even though, GED receiving women's hourly wage does not differ from high school dropouts, they are more likely to participate in labour markets, whereas GED receiving men's labour market outcomes do not significantly differ from those high school dropouts. (Heckman 2014). To conclude, there exists no conclusive evidence of the mechanism through which personality affects labour market outcomes, but the mechanism seems to differ between men and women.

It should not go unnoticed that since the effect personality has on achievement tests and other proxies for cognitive ability has not been accounted for in the majority of studies, they are likely to overemphasize the actual role cognitive ability plays in the dynamics of the labour

market outcomes. It would be reasonable to suggest that the true cognitive ability of individuals with the lowest levels of measured cognitive ability has been systematically undervalued within the literature, and the cognitive ability proxies have simultaneously accounted for economically valuable personality traits, i.e., higher levels of Conscientiousness and Emotional stability (opposite of Neuroticism) as well as intelligence.

Possible future studies on the association between personality and labour market outcomes could investigate extreme levels of certain Big Five factors to account for the possibility of non-linearity in the effects, in a similar manner to Heinec and Anger (2010). Additionally, future studies should attempt to parse out the knowledge about the role non-cognitive traits play in the labour market status, occupational choice and even the compensation scheme selected within an occupation. Also, the literature on the gender wage gap could benefit from the study of the role of detected gender differences in the average levels of certain personality factors.

Finally, what goes largely unsaid within the economic analysis of the intervention programs, is that one can argue over the ethical nature of an intentional personality altering program since within the psychological literature personality is perceived as value neutral. No level of a certain Big Five factor is better or worse than another level of it. Yet, as a soon-to-graduate economist, when studying the findings made in early education interventions literature, it quickly becomes clear why it might be tempting to design an intervention program with a specific purpose to encourage the manifestation of economically valuable personality traits (e.g., Conscientiousness). This initial allure should, however, be treated with caution.

7. Conclusions

In this thesis, I have examined the relationship between psychometrical traits and labour market outcomes from various perspectives by highlighting methodologically different studies, both descriptive and causal, from the fields of psychology and economics. I show that this association between personality and labour market outcomes is widely accepted, and especially Consciousness is associated with higher earnings, whereas Neuroticism and Agreeableness, are, on the other hand, constantly associated with more unfavourable labour market outcomes. Moreover, the Perry Program analysis demonstrates a significant causal

effect between a change in the personality traits of the treatment group and the later manifested labour market outcomes. The treatment group's lifetime earnings increased by 10 or even 35 per cent, while the employment rate also increased by approximately 20 per cent compared to the control group. However, the Perry results are likely not, at least to the full extent, generalizable. Further, it is evident that the mechanism through which personality traits affect labour market outcomes remains unknown. Interestingly, despite this, this mechanism seems to differ between men and women.

In addition, I have emphasized the crucial role of econometrical tools, and consequent assumptions required to prove an associational or causal relationship between psychometric traits and labour market outcomes, while also highlighting the possible pitfalls on both practical and ethical levels.

8. References

Alderotti, G., Rapallini, C., & Traverso, S. (2021). *The Big Five Personality Traits and Earnings: A Meta-Analysis* (No. 902 [rev.]). GLO Discussion Paper.

Almlund, M., Duckworth, A. L., Heckman, J., & Kautz, T. (2011). Personality psychology and economics. In *Handbook of the Economics of Education* (Vol. 4, pp. 1-181). Elsevier

Ashton, M. C., Lee, K., Perugini, M., Szarota, P., De Vries, R. E., Di Blas, L., ... & De Raad, B. (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.

Barnett, W. S. (1985). Benefit-cost analysis of the Perry Preschool Program and its policy implications. *Educational evaluation and policy analysis*, *7*(4), 333-342.

Belfield, C. R., Nores, M., Barnett, S., & Schweinhart, L. (2006). The high/scope perry preschool program cost–benefit analysis using data from the age-40 followup. *Journal of Human resources*, *41*(1), 162-190.

Berrueta-Clement, J. R. (1984). *Changed Lives: The Effects of the Perry Preschool Program on Youths through Age 19. Monographs of the High/Scope Educational Research Foundation, Number Eight*. Monograph Series, High/Scope Foundation, 600 North River Street, Ypsilanti, MI 48197.

Bietenbeck, J. (2020). The long-term impacts of low-achieving childhood peers: evidence from Project STAR. *Journal of the European Economic Association*, *18*(1), 392-426.

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

Carneiro, P., Heckman, J. J., & Masterov, D. V. (2005). Labor market discrimination and racial differences in premarket factors. *The Journal of Law and Economics*, *48*(1), 1-39.

Chetty, R., Friedman, J. N., Hilger, N., Saez, E., Schanzenbach, D. W., & Yagan, D. (2011). How does your kindergarten classroom affect your earnings? Evidence from Project STAR. *The Quarterly journal of economics*, *126*(4), 1593-1660.

Costa Jr, P. T., & McCrae, R. R. (1992). Four ways five factors are basic. *Personality and individual differences*, *13*(6), 653-665.

Costa, P. T., & McCrae, R. R. (1992). Normal personality assessment in clinical practice: The NEO Personality Inventory. *Psychological assessment*, *4*(1), 5.

Cubel, M., Nuevo-Chiquero, A., Sanchez-Pages, S., & Vidal-Fernandez, M. (2016). Do personality traits affect productivity? Evidence from the laboratory. *The Economic Journal*, *126*(592), 654-681.

Cunha, F., Heckman, J. J., Lochner, L., & Masterov, D. V. (2006). Interpreting the evidence on life cycle skill formation. *Handbook of the Economics of Education*, *1*, 697-812.

Dee, T. S., & West, M. R. (2011). The non-cognitive returns to class size. *Educational Evaluation and Policy Analysis*, *33*(1), 23-46.

DeYoung, C. G., Hirsh, J. B., Shane, M. S., Papademetris, X., Rajeevan, N., & Gray, J. R. (2010). Testing predictions from personality neuroscience: Brain structure and the big five. *Psychological science*, *21*(6), 820-828.

DeYoung, C. G., Quilty, L. C., & Peterson, J. B. (2007). Between facets and domains: 10 aspects of the Big Five. *Journal of personality and social psychology*, *93*(5), 880.

Feher, A., & Vernon, P. A. (2021). Looking beyond the Big Five: A selective review of alternatives to the Big Five model of personality. *Personality and Individual Differences*, *169*, 110002.

Gottfredson, L., & Saklofske, D. H. (2009). Intelligence: Foundations and issues in assessment. *Canadian Psychology/Psychologie canadienne*, *50*(3), 183.

Gottschalk, P. (2005). Can work alter welfare recipients' beliefs?. *Journal of Policy Analysis and Management: The Journal of the Association for Public Policy Analysis and Management*, *24*(3), 485-498.

Heckman, J. J., Humphries, J. E., & Kautz, T. (2014). The economic and social benefits of GED certification.

Heckman, J. J., Humphries, J. E., LaFontaine, P. A., & Rodriguez, P. L. (2012). Taking the easy way out: How the GED testing program induces students to drop out. *Journal of labor economics*, *30*(3), 495-520.

Heckman, J. J., Moon, S. H., Pinto, R., Savelyev, P. A., & Yavitz, A. (2010). The rate of return to the HighScope Perry Preschool Program. *Journal of public Economics*, *94*(1-2), 114-128.

Heckman, J. J., Pinto, R., Shaikh, A. M., & Yavitz, A. (2011). *Inference with imperfect randomization: The case of the Perry Preschool Program* (No. w16935). National Bureau of Economic Research.

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

Heckman, J. J., & Karapakula, G. (2019). *The Perry Preschoolers at late midlife: A study in design-specific inference* (No. w25888). National Bureau of Economic Research.

Heckman, J. J., & Karapakula, G. (2019). *Intergenerational and intragenerational externalities of the Perry Preschool Project* (No. w25889). National Bureau of Economic Research.

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

Kautz, T., Heckman, J. J., Diris, R., Ter Weel, B., & Borghans, L. (2014). Fostering and measuring skills: Improving cognitive and non-cognitive skills to promote lifetime success.

Kuncel, N. R., Ones, D. S., & Sackett, P. R. (2010). Individual differences as predictors of work, educational, and broad life outcomes. *Personality and individual differences*, *49*(4), 331-336.

Lee, K., & Ashton, M. C. (2007). Factor analysis in personality research. *Handbook of research methods in personality psychology*, 424-443.

McCabe, L. A., Cunnington, M., & Brooks-Gunn, J. (2004). The development of self-regulation in young children: Individual characteristics and environmental contexts.

McClelland, D. C., Koestner, R., & Weinberger, J. (1989). How do self-attributed and implicit motives differ?. *Psychological review*, *96*(4), 690.

Murray, C. (2002). IQ and income inequality in a sample of sibling pairs from advantaged family backgrounds. *American Economic Review*, *92*(2), 339-343.

Nores, M., Belfield, C. R., Barnett, W. S., & Schweinhart, L. (2005). Updating the economic impacts of the High/Scope Perry Preschool program. *Educational Evaluation and Policy Analysis*, *27*(3), 245-261.

Paulhus, D.L., 1984. Two-Component Models of Socially Desirable Responding. J. Personal. Soc. Psychol. 46 (3), 598–609

Paunonen, S. V., & Ashton, M. C. (2001). Big five factors and facets and the prediction of behavior. *Journal of personality and social psychology*, *81*(3), 524.

Preuss, M., & Hennecke, J. (2018). Biased by success and failure: How unemployment shapes locus of control. *Labour Economics*, *53*, 63-74.

Roberts, B. W., Walton, K. E., & Viechtbauer, W. (2006). Patterns of mean-level change in personality traits across the life course: a meta-analysis of longitudinal studies. *Psychological bulletin*, *132*(1), 1.

Roberts, B. W. (2009). Back to the future: Personality and assessment and personality development. *Journal of research in personality*, *43*(2), 137-145.

Rolnick, A., & Grunewald, R. (2003). Early childhood development: Economic development with a high public return. *The Region*, *17*(4), 6-12.

Schweinhart, L. J., & Weikart, D. P. (1993). Success by Empowerment: The High/Scope Perry Preschool Study through Age 27. *Young children*, *49*(1), 54-58.

Social and Character Development Research Consortium (2010). *Efficacy of Schoolwide Programs to Promote Social and Character Development and Reduce Problem Behavior in Elementary School Children* (NCER 2011–2001). Washington, DC: National Center for Education Research, Institute of Education Sciences, U.S. Department of Education.

Taylor, R. D., Oberle, E., Durlak, J. A., & Weissberg, R. P. (2017). Promoting positive youth development through school-based social and emotional learning interventions: A meta-analysis of follow-up effects. *Child development*, *88*(4), 1156-1171.

Weisberg, Y. J., DeYoung, C. G., & Hirsh, J. B. (2011). Gender differences in personality across the ten aspects of the Big Five. *Frontiers in psychology*, *2*, 178.

Xie, Y., Near, C., Xu, H., & Song, X. (2020). Heterogeneous treatment effects on Children's cognitive/non-cognitive skills: A reevaluation of an influential early childhood intervention. *Social science research*, *86*, 102389.