THREE ESSAYS IN CHILD DEVELOPMENT

A DISSERTATION
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

JIUCHEN DENG

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Paul Glewwe, Advisor

May, 2025

Acknowledgments

I discovered my passion for economics when I was twelve years old, reading Dr. Wu Jinglian's book, *Calling for Market Economy under the Rule of Law*. I still wish that I could meet in person and speak with Dr. Wu once in my life, even though it is unlikely to happen. Regardless, I have become an economist myself, and that passion and knowledge have been carried forward by our generation.

I am deeply thankful to my parents, wife, advisor, peers, dissertation committee members, teaching mentor, and many other people who provided me with support. I am an extremely fortunate person in at least two ways.

First, I was fortunate enough to discover my lifelong passion early on in my "human capital developmental period". I was also lucky that my family provided me with sufficient space and resources for me to grow in the way that I wished for. Now as I recollected, I had a clear goal for my myself, plentiful "parental investments", and an open-minded "household environment" to support the development of my human capital. This makes me an extremely fortunate individual.

Second, I have been an extremely lucky student and researcher. I have met and grown up with people of very different backgrounds and perspectives. The teachers, professors, and advisors have provided me with all the support and knowledge that they can offer. I am an extremely fortunate person that I have been taught, advised, and mentored by some of the kindest human beings in the world. I would like to quote Bertrand Russell's message to future generations in an interview in 1959, where he said "Love is wise. Hatred is foolish... If we are to live together and not die together, we must learn the kind of charity and the kind of tolerance, which is absolutely vital to the continuation of human life on this planet." I am deeply fortunate to have received both intelligence and love along the journey.

People often tell me that I am a diligent and resilient individual. However, I also acknowledge that many of my peers are exceptionally industrious and intelligent researchers. I always remind myself that I am a fortunate person and sometimes I am just luckier than

many other people. I am an empirical economist, and I want to say that randomness, timing, and pure luck matter, especially at critical junctures in life. It is arrogant and irresponsible to attribute one's "success" solely to one self's cleverness and hard work, but we must make every effort before we can even hope to seize a moment of luck. We do our best and leave the rest to the law of large numbers.

To my parents and my wife, for their love and support throughout this journey.

Abstract

This dissertation consists of three independent essays on educational and health issues in child development. The first and second chapters are independent empirical studies authored solely by me, while the third chapter is co-authored with Yue Bao. The first essay investigates the development of personality traits and cognitive skills during adolescence. I find evidence of cross-effects and self-productivity of personality traits and cognitive skills, using Australian data. Both parental investments and child rearing are critical for the development of personality traits and cognitive skills during adolescence. I also find that personality traits and cognitive skills are important for adult earnings. The second essay examines the effect of living in the vicinity of an electronic-waste dumpsite on infant health outcomes in Ghana. I find that exposure from the e-waste dumping site significantly increases the probability of diarrhea and respiratory illness. I find two mechanisms, tetanus infection and water pollution. I find that premature birth is more likely to occur when mothers have not received a tetanus vaccine during or before pregnancy in the exposed area. I also find that diarrhea can be mitigated by consuming safer sources of drinking water. In the third essay, we develop a two-period discrete choice model to examine how higher education decisions are shaped when student loans play an increasingly crucial role in the US. We still find disparities in higher education attainment under a constructed scenario in which all students, regardless of family income, rely on student loans. The model highlights and helps us better understand whether disparities in higher education attainment are due to unequal access to financing or to deeper structural inequalities (for example, inequalities in K-12 preparation).

Table of contents

Chapter 1

Introduction

My main research lies at the intersection of labor economics, development economics, health economics, and child development. My dissertation centers on three key questions in child development. First, I use Australian data to investigate how personality traits and cognitive skills are developed, and their implications for labor market outcomes, with a particular focus on personality development. The development of personality traits is an area often understudied in the existing literature (Brunello and Schlotter, 2011). Second, I explore the health impacts of adverse early childhood environments, emphasizing that human capital development should encompass not only cognitive skills and personality traits but also physical and mental well-being. In this chapter, I examine the effect of living in the vicinity of an electronic-waste dumpsite on infant health outcomes in Ghana. Lastly, my third essay examines how students make independent decisions about college attendance, under a constructed scenario in which all students, regardless of family income, rely on student loans. The three chapters of my dissertation are connected under the bigger umbrella of child development, where I study three crucial aspects of child development, the formation of human capital, health outcomes, and higher education. The three chapters also differ, as they investigate different research topics under a diversity of settings.

In the first essay, I investigate how these traits and skills evolve, their responsiveness to investments and the individual's environment, and the implications for labor market outcomes. Noncognitive skills are essential to child development, educational attainment, labor market outcomes, and intra-household resource allocation (Carneiro et al., 2007; Deming, 2017; Flinn et al., 2018). However, the formation process of personality traits remains understudied (Brunello and Schlotter, 2011).

To investigate the development of adolescents' Big Five personality traits and cognitive

skills, I apply first a dynamic factor model, and then a control function approach, both of which are connected through an overarching data generating process. I find evidence of cross-effects and self-productivity in adolescents' personality traits and cognitive skills. Specifically, I discover a positive impact of mathematical education on the development of emotional stability, which is consistent with findings from recent neuroscience research. I find evidence that personality traits are malleable to investments, and to the individual's environment, in late adolescence, which is a contribution to the literature, as no consensus exists on the malleability of noncognitive skills in late adolescence and early adulthood. I also find that both personality traits and cognitive skills are important for adult earnings. Based on the control function approach, I find positive impacts on the development of adolescents' personality traits and cognitive skills from family income, which includes income from the mother's employment, that fully offsets the negative effect of maternal hours worked on those traits and skills. I also find that the results from the dynamic factor model and the control function approach are internally consistent.

The second essay investigates the effect of living in the vicinity of an e-waste dumpsite, named Agbogbloshie, on infant health outcomes in Ghana. Economists have examined a wide range of fetal shocks and found that early shocks can have substantial negative impacts not only on health but also on economic outcomes later in life (Currie and Almond, 2011). I study the impact of adverse early childhood environments, emphasizing that human capital development should encompass not only the development of cognitive and noncognitive skills, but also physical and mental well-being.

More than 18 million children and adolescents, and as many as 12.9 million women, were actively engaged in the informal electronic waste (e-waste) sector work around the globe (World Health Organization, 2021). I use a difference-in-differences strategy and find that exposure from the e-waste dumpsite in Ghana increases children's diarrhea and respiratory illness. I utilize triple differences to study the mechanisms of the effects, and I uncover that premature birth is more likely to occur when mothers in the exposed area have not received a tetanus vaccine. I also find that children's diarrhea can be mitigated by the household consuming safer sources of drinking water. This research provides important policy implications to effectively address the adverse impact of living close to an e-waste dumpsite on children.

In the third essay, my coauthor and I investigate the decision-making process of college attendance in the US, when student loans play an increasingly essential role. Education is essential to the development of skills, children's long-term success, as well as many

dimensions of economic life. We observe that a substantial and growing share of students now rely on loans to cover tuition costs, regardless of household income (Fry, 2014). In this essay, we examine whether disparities in higher education attainment would still exist, when education is not perceived as a tool for intergenerational transfers but rather as a decision that is independently made by children to maximize their life-cycle utility, with children fully funding their higher education through borrowing.

In the setting where all students borrow student loans for higher education, we still find significant disparities in college access and school quality between students from the lower and the upper income levels. This inequality is largely attributed to the differences in the quality of pre-college education. We run policy experiments based on this simple setting and find that the key approach to close the gap between students from low- and high-income families is to provide education of higher quality to disadvantaged children during earlier years.

Chapter 2

The Development of Personality Traits and Cognitive Skills in Adolescence: Evidence from a Skill Formation Model and a Control Function Approach

2.1 Introduction

In the economics literature, many researchers have found that noncognitive skills are essential to child development, educational attainment, labor market outcomes, and intra-household resource allocation (Carneiro et al., 2007; Deming, 2017; Flinn et al., 2018). The formation, and implications, of cognitive skills have been widely examined. Yet, noncognitive skills and personality traits are still understudied in economics, even though noncognitive skills are at least as important as cognitive skills for individuals' development, education, and labor market success (Brunello and Schlotter, 2011). For instance, noncognitive skills have a stronger predictive power than cognitive skills at the lower end of the earnings distribution, where men with low noncognitive skills are significantly more likely to be unemployed than those with low cognitive skills, and the former tend to experience longer spells of unemployment (Lindqvist and Vestman, 2011). With respect to education, Carneiro et al. (2007) found that the marginal effect of cognitive skills on the likelihood of educational

attainment beyond age 16 was small at low values of noncognitive skills, but the marginal effect was much larger at high values of noncognitive skills.

Despite its importance, we have limited knowledge about the development process regarding each element within the package called "noncognitive skills" that researchers typically refer to as whole. Economists have developed several innovative models to estimate the formation of cognitive and noncognitive skills (Cunha and Heckman, 2008; Cunha et al., 2010; Agostinelli and Wiswall, 2023). Todd and Wolpin (2003, 2007) provided insightful models for the production function of cognitive skills. Attanasio et al. (2020) contributed to the study of the production of cognitive skills and health for children in India. Aucejo and James (2021) estimated the of formation of math and verbal skills and explained the role of cognitive skills played in determining gender gaps in college enrollment in England.

However, unlike cognitive skills, which have often been disaggregated into math and verbal skills, noncognitive skills, in the literature, are typically collapsed into a single factor using multiple related measurements. In reality, noncognitive skill is a multifaceted concept which embodies many nuances, especially the development of personality traits in the adolescent period (De Fruyt and Karevold, 2021; Van den Akker et al., 2021; Arnett, 2023). Therefore, it is important to investigate and understand the development of each individual component within the "noncognitive skills" package, and the five-factor model (the Big Five personality traits) is a long-pursued and well-established approach that depicts different aspects of an individual's noncognitive skill or personality (Goldberg, 1992; Saucier, 1994).[1] The development of children's personality traits during the span of adolescence is particularly intriguing, due to the drastic changes in pubertal hormone concentrations (Van den Akker et al., 2021). Thus, in this paper, I investigate the formation process of each element of children's Big Five personality traits, as well as cognitive skills, during the adolescent period.[2] I implement two different methods, a skill formation model and a control function approach, that are linked by an overarching data generating process.

This paper makes three contributions. First, I contribute to both the literature that studies children's skill formation process and the literature that examines the impact of maternal labor supply and family income on children's outcomes. In contrast to the current

[1]The five-factor model of personality is a model that organizes personality traits into five basic dimensions: extroversion, agreeableness, emotional stability, conscientiousness, and openness to experience.

[2]In this paper, the terms "noncognitive skills" and "personality traits" refer to essentially the same concept. I use "noncognitive skills" when describing findings from other studies, as this is the term commonly used in the literature. The term "personality traits" refers to the disaggregated components of noncognitive skills that I examine in this paper.

literature on the skill formation process, which perceives noncognitive skill as a single factor, I disaggregate noncognitive skill into the Big Five personality traits (extroversion, agreeableness, openness, conscientiousness, and emotional stability) and estimate the formation process for each of the personality traits during adolescence, using a skill formation model following Cunha and Heckman (2008). I empirically estimate the cross-effects (the existing stock of one skill affects the level of a different skill in the next period) and the self-productivity (the existing stock of one skill increases the level of the same skill in the next period) for adolescents' Big Five personality traits, math skills, and reading skills. While the current literature that examines the impact of maternal hours worked and family income on children's outcomes tends to focus on cognitive skills or treat noncognitive skills as a single index, I investigate the impact of maternal hours worked and family income on the development of each component of adolescents' Big Five personality traits, and of cognitive skills, using a bunching with control function approach based on (Caetano et al., 2021).

Second, I put recent pioneering discoveries in cognitive neuroscience to the test, applying two different approaches from economics. Zacharopoulos et al. (2021) found that adolescent students who lacked mathematical education in school displayed reduced γ-aminobutyric acid (GABA) concentration levels in a key brain area that was involved with the ability of inhibitory control. Inhibitory control (ability to control one's impulses and behavioral responses to stimuli), by definition, is reflected in traits such as emotional stability and conscientiousness from the five-factor model. I cross-validate this neuroscience discovery based on the evidence of the positive cross-effect between mathematical skill and emotional stability. I find that the stock of mathematical skill in the current period has a positive cross-effect on the formation of adolescents' emotional stability in the next period. This finding is consistent with what neuroscience research suggests (Zacharopoulos et al., 2021), which provides complementary evidence of the interplay between education and neurobiology.

Third, I study the malleability of personality traits to investments and to the environment during the "emerging adulthood" period.[3] Understanding whether these traits can significantly change during early adulthood is crucial, as personality traits are essential for labor market success. Employers, for instance, value personality as a form of soft skills, raising the question of whether training and learning in the workplace can contribute to their development. However, no consensus exists on the malleability of noncognitive skills or personality traits in later life, with some researchers arguing that these traits are largely

[3] Arnett (2023) introduced the concept of "emerging adulthood" that occurred from the ages of 18 to 25.

fixed by the end of the teenage years, while others suggest that these traits can be changed at any age (Brunello and Schlotter, 2011). My research addresses this gap by rigorously examining the development of the Big Five personality traits, and I find that these traits are malleable to investments and to the environment during the emerging adulthood period.

I utilize the Household, Income and Labour Dynamics in Australia (HILDA) data set for both the skill formation model and the control function approach to study adolescents' skill formation. To investigate the formation of personality traits and cognitive skills, I employ a linear specification of the skill formation model of Cunha and Heckman (2008).[4] I also use a bunching with control function approach to explore the development of these traits and skills, as well as to examine the impact of maternal labor supply and household income on adolescents' traits and skills, following (Caetano et al., 2021). The results from the skill formation model and from the control function approach are internally consistent, and the results from both methods are largely consistent with relevant studies.

Estimating the the skill formation model, I identify the parameters of the law of motion for traits and skills using an instrumental variable estimator, where a measurement of a trait or skill is used to instrument for the second measurement of the same trait or skill. Further exposition and the validity of this instrumental variable estimator are discussed in Subsection 2.3.2. The results yield important findings that are key to understand the formation of personality traits, their self-productivity, and the cross-effects of personality traits and cognitive skills.

First, I discover statistically significant evidence of a positive cross-effect of mathematical skill on emotional stability and conscientiousness. Based on the estimates of the skill formation model, a one standard deviation increase in the math skill in late adolescence leads to an increase in emotional stability in early adulthood by 0.14 standard deviations, a result that is significant at the 1% level. The result is consistent with findings in neuroscience research, which uncovers that mathematical education exerts a positive impact on adolescents' inhibitory control (a concept closely related to to emotional stability) (Zacharopoulos et al., 2021).[5] I also find significant self-productivity of the five personality traits, the math skill,

[4] I also apply a non-linear transcendental logarithmic (translog) skill production function based on Agostinelli and Wiswall (2023) as a robustness check. The results from the linear specification and the results from the translog specification are largely consistent. The estimation of the translog specification is in Subsection 2.7.1.

[5] Studies report adverse impacts of a lack of mathematical education on the development of key brain areas. A lack of math education is associated with a decreased gamma-aminobutyric acid (GABA) concentration within the middle frontal gyrus (MFG) and is also negatively associated with frontoparietal connectivity. Lower brain GABA levels are associated with inhibitory control defcits. Greater connectivity between the

and the reading skill. Furthermore, parental investments in the previous period exhibit a statistically significant positive marginal effect on adolescents' agreeableness, extroversion, and reading skill.

In addition to the skill formation model, I estimate a bunching with control function method based on Caetano et al. (2021). While the control function method has been well-established and widely-discussed (Florens et al., 2008; Glewwe and Todd, 2022), the bunching with control function approach is a relatively recent identification strategy (Caetano, 2015; Caetano et al., 2021). The key idea of this method is to address selection on unobservables by exploiting the zero hours of work constraint (the bunching), constructing a control function accordingly, and netting out the confounding relationship by including the control function in the regression framework.[6]

The findings based on the bunching with control function approach are consistent with the results from the skill formation model. I find statistically significant evidence of self-productivity and cross-effects of personality traits and cognitive skills. The cross-effect between adolescents' math skill in the current period and emotional stability in the next period is positive and statistically significant at the 5% level. The effect of family labor income in the previous period is positive and statistically significant on adolescents' agreeableness, extroversion, openness, and the reading skill. The impact of maternal hours worked in the previous period on adolescents' emotional stability is negative and statistically significant at the 10% level. However, I find that the positive income effect due to maternal labor supply fully offsets the direct negative impact of maternal labor supply on adolescents' outcomes. I also find that both personality traits and cognitive skills are essential for adult earnings. The magnitudes of the impacts of the Big Five personality traits on adult earnings are generally as large as the magnitudes of the impacts of cognitive skills.

The malleability of personality traits and cognitive skills in the "emerging adulthood" period provides important policy implications. For instance, the earlier view of training programs for young adults suggested that the effects were small and not very cost-effective (Heckman, 1994; Hirshleifer et al., 2016). These findings regarding training programs might have led some economists to the a priori assumption that traits and skills are not malleable in early adulthood, explaining why skill formation in later periods is an understudied area.

frontoparietal network has a significant positive correlation with higher intelligence scores and superior cognitive functioning (Marek and Dosenbach, 2018; Zacharopoulos et al., 2021).

[6]The zero hours of work constraint means that the desired hours of work for mothers who stay home is unobserved, which can be exactly zero and also can be negative hours of work but constrained at zero. Please refer to Section 2.5 for a more detailed exposition.

In contrast, my findings suggest that investments and the environment during emerging adulthood do matter. Furthermore, my paper sheds light on the direction that research should pursue to design more effective training programs, since the ineffectiveness of some of these programs may be due to the program design itself rather than the rigidity of traits and skills in early adulthood.

In the next section, I introduce the data set and describe the measurements for personality traits, cognitive skills, and household characteristics. In Section 2.3, I briefly present the skill formation model and discuss the identification of the law of motion for skills. In Section 2.4, I report the empirical results from estimating the model described in Section 2.3. In Section 2.5, I introduce the bunching with control function strategy. In Section 2.6, I present the empirical results based on the control function approach. In Section 2.7, I conduct various robustness checks for the skill formation model as well as for the control function approach. Section 2.8 concludes.

2.2 Data

I use the Household, Income and Labour Dynamics in Australia (HILDA) longitudinal survey to estimate both the skill formation model and the reduced-form method (the bunching with control function approach). The HILDA data set is a nationally representative household panel survey that was first implemented in 2001 with an initial sample of 19,914 individuals from 7,682 households in Australia. I mainly utilize three types of variables from the HILDA data set: measurements for the Big Five personality traits and cognitive skills; measurements for parental investments and household environment; and variables for parental characteristics.

In wave 5 (year 2005), the HILDA started to collect the Big Five personality traits assessments, once every four years. The Big Five personality traits are extroversion, agreeableness, openness, conscientiousness, and emotional stability. The personality traits assessments are based on the 36 Text Dependent Analysis items (TDA-36) of the trait descriptive adjectives approach (Summerfield et al., 2023). The TDA-36 approach employed by the HILDA was based on the original TDA-40 items scale developed by Saucier (1994). In Figure A.1, I present a table from Losoncz (2009) that shows the TDA-36 items that the HILDA uses to assess the Big Five personality traits. The personality traits measurements from the HILDA data are believed by some economists to be of the highest quality among all nationwide surveys (Todd and Zhang, 2020).

Every item of the TDA-36 items is measured on a categorical scale from one to seven, with one being "this descriptive adjective does not describe me at all" and seven being "this descriptive adjective describes me very well". Not all 36 items are used to summarize and derive the five personality factors. HILDA first examines the reliability of items and omits any item with a correlation of less than 0.3. Second, HILDA performs a principal component analysis to retain items based on their factor loadings, and the retained items are used to generate the five personality traits (Summerfield et al., 2023). The HILDA data set provides the five personality traits derived from the set of items, which I use directly for my analysis. The five composite personality traits are on a continuous scale from one to seven. The scale of the Big Five personality traits is continuous because they are composite scores derived from a set of categorical items.

The cognitive skills that I investigate are comprised of two parts: mathematical skill and reading skill. Mathematical and reading skills are based on self-rated math and reading skills "compared to the average Australian" that are collected every four years. The math and reading skills are on a scale from zero to ten, with an increment of one, where zero is "very poor", five is "average", and ten is "very good".

HILDA provides another measure of reading skill, which is a 25-item version of the 50-item National Adult Reading Test. Mathematical skill also has a second related measure, which is a Symbol Digit Modalities Test that measures the time a participant needed to pair abstract symbols with specific numbers.[7] However, these two measures are available in waves 12 and 16 only, and thus are not adequate for the analysis. To examine the reliability of the self-rated math and reading skills, I compute the correlation coefficient and also administer an OLS regression between the self-rated reading and math skills and the National Adult Reading Test and the Symbol Digit Modalities Test scores, respectively. The correlation coefficients are 0.438 and 0.318, respectively, for reading and math, which are considered as medium-sized correlation based on the criteria by Cohen (2013). The regression coefficients are presented in Table A.1, where self-rated reading skill is predicted with the National Adult Reading Test score and self-rated math skill is predicted with the Symbol Digit Modalities Test score. The results are statistically significant at the 0.1% level.

Parental investments in and household environment of child development are measured using five items: family education-related expenditure, expenditure on internet and phone, expenditure on child clothing, family income, and the frequency that a child watches

[7] A Symbol Digit Modalities Test example is shown in Figure A.2

television or movies. Parental characteristics consist of mother's and father's education, mother's self-reported math and reading skills, and mother's hours worked on a weekly basis.

Table 2.1: Summary Statistics of Measurements and Parental Characteristics

	Mean	Std. Dev.		Mean/Proportions (%)	Std. Dev.
Trait and Skill Variables			***Parental Characteristics***		
Emotional Stability	4.88	1.09	Mother's Annual Hours Worked	1318.93	897.26
Conscientiousness	4.75	1.02	Mother's Math Skill	6.67	1.97
Agreeableness	5.20	0.98	Mother's Reading Skill	8.18	1.87
Extroversion	4.53	1.06			
Openness	4.23	1.06	Mother's Education:		
Self-assessed Math Skill	6.73	1.96	None	1.7%	
Self-assessed Reading Skill	7.79	1.83	Primary School Only	13.4%	
			Some Secondary School (but no more)	42.1%	
Investment Variables			Year 11 or Equivalent	10.6%	
Family Income	125738.90	63884.83	Year 12 or Equivalent	32.2%	
Family Education Expenditure	3115.72	4582.24			
Internet and Phone Expenditure	1917.24	1588.76	Father's Education:		
Television or Movies Expenditure	1082.91	1338.83	None	1.1%	
Child Clothing Expenditure	1040.18	754.15	Primary School Only	16.0%	
			Some Secondary School (but no more)	43.1%	
			Year 11 or Equivalent	7.9%	
			Year 12 or Equivalent	31.8%	
Total Individuals			6,713		

Note. This table shows the main descriptive statistics of the HILDA sample used for analysis. Personality traits measures are on a scale from one to seven, where seven indicates the highest score. Math and reading skills are measured on a scale from zero to ten, where ten indicates the highest level. All investment variables are on a yearly basis. The currency is in 2005 Australian dollars.

Summary statistics of the variables are reported in Table 2.1. The estimations focus on youths from the ages of 15 to 29. The final sample is comprised of 20,512 total observations from 6,713 individuals, of which about 51.1% are male and 48.9% are female.

2.3 A Skill Formation Model

2.3.1 Skill Production Function

In this section, I implement a linear skill formation model based on Cunha and Heckman (2008) to analyze the formation of personality traits and cognitive skills. Leveraging this linear specification for simplicity and tractability, I identify the cross-effects and self-productivity of the stocks of personality traits and cognitive skills. I keep the presentation brief and refer the reader to Cunha and Heckman (2008) for proofs and a more detailed exposition.

Personality traits are represented by the Big Five personality traits, and cognitive skills consist of math skill and reading skill. I examine the cross-effects of every pair of skills, estimate the self-productivity of each skill, and investigate the impact of parental investments on traits and skills. Note that I use a dynamic factor model to retrieve parental investments as a latent factor from a set of observable investment variables, when I estimate the skill formation model. I will present the dynamic factor model in Subsection 2.3.2.

In this section, I keep the technology of skill formation in a linear specification, because a non-linear model will unlikely overrule the conclusions if I find little evidence in a linear setting. However, in the robustness checks subsection (Subsection 2.7.1), I estimate a non-linear skill production model to examine the stability of the results from the linear model.

Assume that each agent is born with initial condition $\theta_1 = \left(\theta_1^C, \theta_1^P\right)$, where θ_1^C represents a vector of two elements of cognitive skills: mathematical skill and reading skill. Similarly, θ_1^P is a vector of the five elements of the Big Five personality traits: extroversion, agreeableness, openness, conscientiousness, and neuroticism. From now on, I will refer to neuroticism as its converse, emotional stability, in order to cause less confusion in interpretation and keep the "direction" to be the same for the five personality traits. The production technology for skill k in period t is

$$\theta_{t+1}^k = f_t^k\left(\theta_t^C, \theta_t^P, \theta_{k,t}^I\right) \tag{2.1}$$

for $k \in \{C, P\}$, $t \in \{1, 2, 3, ..., T\}$, and I represents family investments in child's human capital. This production technology for skills underlines the idea that the stocks of the two cognitive skills and the five personality traits, and family investments, together produce traits and skills in the next period.

An agent's adult human capital h is a function of period $t+1$ traits and skills accumulated by adulthood (age 25),[8] specified as

$$h = g\left(\theta_{t+1}^C, \theta_{t+1}^P\right). \tag{2.2}$$

[8] Arnett (2023) introduced the concept of "emerging adulthood" that occurred from the ages of 18 to 25. Arnett (2023) also stated, on page 7, that 18 to 29 was a reasonable range internationally, since median ages of entering marriage were higher in other developed countries than in the United States. Therefore, I set age 25 as the beginning of adulthood and use traits and skills at age 29 as second measures. The reason that a second measure is needed for identification will be discussed in Subsection 2.3.3.

2.3.2 Dynamic Factor Model

I apply a dynamic factor model for the key latent variables following Cunha and Heckman (2008). The key latent variables are the agent's traits and skills represented by the set of his or her mathematical skill, reading skill, extroversion, agreeableness, openness, conscientiousness, and emotional stability. The key latent variables also include parental investments in the agent's abilities. Formally speaking, I employ a confirmatory factor model where the latent variables cause the measurements (the reflective indicators). The factor model is important for the analysis because it describes the covariance relationships among many observables, the measurements of abilities, in terms of several underlying unobservables or latent variables, which in this case are personality traits and cognitive skills.

The value of a latent variable is not directly observable; instead, I observe a vector of measurements that serve as reflective indicators of the latent factor. The measurement system of the factor model is formulated as:

$$Y_{j,t}^k = \mu_{j,t}^k + \alpha_{j,t}^k \theta_t^k + \varepsilon_{j,t}^k, \text{ for } j \in \left\{1, \ldots, m_t^k\right\}, k \in \{C, P, I\}, \tag{2.3}$$

where $Y_{j,t}^k$ is an $m_t^k \times 1$ vector of the observed measurements, $\mu_{j,t}^k$ is a vector of the mean values associated with these observed measurements, and m_t^k is the number of measurements on cognitive skills, personality traits, and investments in period t. Finally, $\varepsilon_{j,t}^k$ is the measurement error term and $\mathbb{E}[\varepsilon_{j,t}^k] = 0$. For brevity, C represents a vector of the two cognitive skills, P is a vector of the Big Five personality traits, and I represents parental investments. A specific cognitive skill, personality trait, or family investment at stage t is represented by a latent variable θ_t^k. The parameter $\alpha_{j,t}^k$ is a vector of the factor loadings for the measurements of latent skills, traits, and investments at time period t, which shows the correlation between the vector of latent variables and the vector of the measurements. The stocks of traits and skills at each stage t is not directly observable, but rather reflected by a vector of measurements. Using the factor model, I can estimate the variance covariance matrix between two latent variables as well as the variance of the error terms $\varepsilon_{j,t}^k$ associated with the measurements.

To perform the analysis, each latent variable requires at least two measurements that are associated with it. Then I normalize the factor loading of the first measurement to one, which is to set $\alpha_{1,t}^k = 1$.[9] This makes the item the reference measurement, and other factor loadings are interpreted relative to the reference item, which yields a standardized solution.

[9] In factor analysis, it is a common practice to constrain one factor loading of a measurement to be one.

Given two measurements $Y_{j,t}^k, j \in \{1,2\}$ for a latent variable k, I can calculate the covariance of the two measurements, $Cov(Y_{1,t}^k, Y_{2,\tau}^l)$, from the data for all pairs of different time periods t and τ and different measurements k and l. The covariance of any two latent variables in any two time periods can be found by computing the covariance of their corresponding measurements from those two time periods, given the general formula:

$$\text{Cov}\left(Y_{j,t}^k, Y_{j,\tau}^l\right) = \alpha_{j,t}^k \alpha_{j,\tau}^l \,\text{Cov}\left(\theta_t^k, \theta_\tau^l\right), \tag{2.4}$$

for $k, l \in \{C, P, I\}$, $j \in \{1,2\}$, and $t, \tau \in \{1, ..., T\}$. The factor loadings $\alpha_{j,t}^k$ and $\alpha_{j,\tau}^l$ can be computed as follows. Given that $\alpha_{1,t}^k$ is normalized to one, the covariances can be computed based on the general formulas:

$$\text{Cov}\left(Y_{1,t}^k, Y_{1,t+1}^k\right) = \text{Cov}\left(\theta_t^k, \theta_{t+1}^k\right) \tag{2.5}$$

and

$$\text{Cov}\left(Y_{j,t}^k, Y_{1,t+1}^k\right) = \alpha_{j,t}^k \,\text{Cov}\left(\theta_t^k, \theta_{t+1}^k\right). \tag{2.6}$$

Then $\alpha_{j,t}^k$ can be computed by having Equation (2.6) divided by Equation (2.5). The parameter $\alpha_{k,\tau}^l$ can also be identified in a similar fashion.

The variance of any latent variables θ_t^k can be identified utilizing the following formula:

$$\text{Cov}\left(Y_{1,t}^k, Y_{2,t}^k\right) = \alpha_{2,t}^k \,\text{Var}\left(\theta_t^k\right), \tag{2.7}$$

where the factor loading $\alpha_{2,t}^k$ can computed by taking the ratio of Equations (2.5) and (2.6), $\text{Cov}\left(Y_{1,t}^k, Y_{2,t}^k\right)$ can be computed from the data, and thus $\text{Var}\left(\theta_t^k\right)$ is identified. Assume that the error term in the measurement system $\varepsilon_{j,t}^k$ is mean zero, independent across agents and over time, independent of the latent factors, and serially independent. Then I can identify the variance covariance matrix of the error term $\varepsilon_{j,t}^k \sim \text{N}\left(0, \sigma_{k,j,t}^2\right)$ associated with any two measurements $j \in \{1,2\}$ for any pair of factors $k \in \{C, P, I\}$ from this formula:

$$\text{Var}\left(Y_{j,t}^k\right) - \left(\alpha_{j,t}^k\right)^2 \text{Var}\left(\theta_{j,t}^k\right) = \sigma_{k,j,t}^2, \tag{2.8}$$

for $k \in \{C, P, I\}$ and $t \in \{1, ..., T\}$. The parameter $\sigma_{k,j,t}^2$ can be identified, since elements $Var\left(Y_{j,t}^k\right)$, $\alpha_{j,t}^k$ and $Var\left(\theta_{j,t}^k\right)$ can be calculated by the previous arguments.

2.3.3 Identification of the Law of Motion for Skills

I formulate the evolution of the five personality traits (extroversion, agreeableness, openness, conscientiousness, and emotional stability) and the two cognitive skills (math and reading) in the following equation system:

$$\begin{pmatrix} \theta_{t+1}^P \\ \theta_{t+1}^C \end{pmatrix} = \begin{pmatrix} \gamma_1^P & \gamma_2^P & ... & \gamma_7^P \\ \gamma_1^C & \gamma_2^C & ... & \gamma_7^C \end{pmatrix} \begin{pmatrix} \theta_t^P \\ \theta_t^C \end{pmatrix} + \begin{pmatrix} \gamma_8^P \\ \gamma_8^C \end{pmatrix} \theta_t^I + \begin{pmatrix} \eta_t^P \\ \eta_t^C \end{pmatrix}, \tag{2.9}$$

where the latent personality trait factor $\theta_t^P = (\theta_t^{p1}\, \theta_t^{p2}\, \theta_t^{p3}\, \theta_t^{p4}\, \theta_t^{p5})^T$ represents a column vector of the five personality traits, and the latent cognitive skill factor $\theta_t^C = (\theta_t^{c1}\, \theta_t^{c2})^T$ represents a column vector of math and reading skills for brevity of presentation of the equation system, and θ_t^I is a scalar that represents parental investments.

For a single skill $k \in \{C, P\}$, I estimate a linear law of motion at period $t \in \{1, ..., T\}$ as:

$$\theta_{t+1}^k = \gamma_0^k + \gamma_p^k (\theta_t^P)^T + \gamma_c^k (\theta_t^C)^T + \gamma_8^k \theta_t^I + \eta_t^k, \tag{2.10}$$

where $(\theta_t^P)^T$ is a row vector of the stocks of the Big Five personality traits from the previous period , $(\theta_t^C)^T$ is a row vector of the stocks of the two cognitive skills from the previous period, θ_t^I is a scalar of parental investments, and the error term η_t^k is assumed to be independent of the latent factors $\{\theta_t^P, \theta_t^C, \theta_t^I\}$. Note that γ_p^k is a vector of parameters for the vector of five personality traits, and γ_c^k is a vector of parameters for the vector of two cognitive skills. The law of motion for a single skill can be interpreted as an autonomous equation that in itself, has economic meaning in the sense that if the parameters of the law of motion are known, for any child, I can find his or her traits and skills in the next period, given the values of the traits and skills and parental investments in the current period (Wooldridge, 2010).[10] The law of motion for skills demonstrates that the stock of any skill θ_{t+1}^k in a period $t+1$ is produced by the stocks of the two cognitive skills, the stocks of the five personality traits, and family investments in the previous period t.

Next, I define

$$\begin{aligned} \tilde{Y}_{1,t+1}^k &= Y_{1,t+1}^k - \mu_{1,t+1}^k \\ \tilde{Y}_{1,t}^k &= Y_{1,t}^k - \mu_{1,t}^k \end{aligned} \tag{2.11}$$

based on Equation (2.3) of the measurement system. I can substitute the measurement

[10] Page 239.

equations $\tilde{Y}^k_{1,t+1}$ and $\tilde{Y}^k_{1,t}$ as proxies for θ^k_{t+1} and θ^k_t for $k \in \{C, P, I\}$, which yields:

$$\tilde{Y}^k_{1,t+1} = \gamma^k_0 + \gamma^k_p \tilde{Y}^P_{1,t} + \gamma^k_c \tilde{Y}^C_{1,t} + \gamma^k_8 \tilde{Y}^I_{1,t} + \left(\varepsilon^k_{1,t+1} - \gamma^k_{p,t}\varepsilon^P_{1,t} - \gamma^k_{c,t}\varepsilon^C_{1,t} - \gamma^k_{8,t}\varepsilon^I_{1,t} + \eta^k_t\right). \quad (2.12)$$

The regressors $\tilde{Y}^P_{1,t}$, $\tilde{Y}^C_{1,t}$ and $\tilde{Y}^I_{1,t}$ are correlated with the composite error term in the parentheses; therefore, an ordinary least squares estimator will produce inconsistent estimates.

To address this issue, I apply an instrumental variables estimator. I use $Y^P_{2,t}$, $Y^C_{2,t}$ and $Y^I_{2,t}$ as instrumental variables for $\tilde{Y}^P_{1,t}$, $\tilde{Y}^C_{1,t}$ and $\tilde{Y}^I_{1,t}$, respectively, and implement a two-stage least squares regression to estimate the parameters in Equation (2.10). This process of estimation can be applied to identify the parameters in other time periods.

The key assumption for this instrumental variable method to be valid is that the error terms of the measurement equations for the first and second measurements are uncorrelated; otherwise, the exclusive restriction condition for the instrumental variable framework is violated. However, within the setup of the dynamic factor model, this becomes a testable assumption. Utilizing Equation (2.8), I can explicitly compute the degrees of correlation of the the error terms of the measurement equations between any two measurements. In Section 2.4 of the paper, I will present a table of the correlation matrix of the measurement errors between every pair of measurements of personality traits, and test whether the correlations are significantly different from zero.

In theory, the independence assumption between the error term η^k_t and the latent factors $\{\theta^P_t, \theta^C_t, \theta^I_t\}$ can be relaxed, allowing unobservable inputs to be correlated with the latent factors. The key is to decompose the error term η^k_t into a time-invariant omitted input and a time-varying omitted input. Then the time-invariant term can be eliminated by applying a first-differencing method between the law of motion for skills in stage $t+1$ and the law of motion in stage t. Finally, an IV estimator is applied to ensure consistent estimation. That is, the identification utilizes different measurements from an earlier period, $\left\{\left(Y^k_{j,t-1} - Y^k_{j,t-2}\right)\right\}_{j=2}^{m^k_t}$, to instrument for the regressor in the first-differencing equation of skill k, $(\tilde{Y}^k_{1,t} - \tilde{Y}^k_{1,t-1})$.

However, since participants' personality traits and cognitive skills are not surveyed by the HILDA until age 15, I am not allowed to relax the independence assumption because this procedure requires the data set to have at least two measurements for personality traits and cognitive skills from two prior periods, $t-1$ and $t-2$. In the HILDA data set, personality traits are not surveyed until in wave 5, and 15-year-old individuals are the youngest ones whose information on personality traits and cognitive skills are available. For

instance, if I am interested in studying the law of motion for skills from the late adolescent period (age 17-21) to early adulthood (age 25), to relax the independence assumption, two measurements from age 9 to 13 and two measurements from age 13 to 17 will be required, which are not available in the data.

In Section 2.4, I present the estimated parameters of the law of motion for skills from the late adolescent period (age 17 to 21) to the early adulthood under the independence assumption.[11] Despite a stronger assumption, it provides a comprehensive view of the cross-effects and self-productivity of personality traits and cognitive skills. In Section 2.5, I address selection on unobservables by employing a bunching with control function approach, and delve deeper into evaluating the impact of mother's time (maternal hours worked) and investment (family labor income) on the formation of youths' personality traits and cognitive skills. I present the causal estimates from the control function approach in Section 2.6. The results based on the skill formation model and the control function approach are largely consistent.

2.4 Empirical Results: Skill Formation Model

I estimate the law of motion for skills expressed in Equation (2.12) by applying a two-stage least squares regression as discussed in Subsection 2.3.3. For the late adolescent period, I use youths' personality traits and cognitive skills at age 21 to instrument for those at age 17. For the early adulthood period, I use youths' personality traits and cognitive skills at age 29 to instrument for those at age 25. In the estimations, I include mother's cognitive skills, mother's education, and father's education as covariates, which are considered as time-invariant parental characteristics. The empirical results are presented in Table 2.2, where the inputs from the previous period are located in the "Period t" column, and outputs in adulthood are located in the "Period t+1" row.

[11] For the late adolescent period, I use youths' personality traits and cognitive skills at age 21 to instrument these traits and skills at age 17. For the early adulthood, traits and skills at age 29 are used to instrument those at age 25

Table 2.2: Marginal Effects (Elasticities) of Skills: Estimates of Linear Technology of Skill Formation

Period t \ Period t+1	Emotional Stability	Conscientiousness	Agreeableness	Extroversion	Openness	Math	Reading
Emotional Stability	0.686***	-0.002	0.002	-0.014***	0.001	0.010	0.004
	(0.011)	(0.021)	(0.009)	(0.002)	(0.020)	(0.010)	(0.010)
Conscientiousness	0.076***	0.764***	-0.009	0.014	-0.048**	-0.003	0.000
	(0.015)	(0.022)	(0.007)	(0.013)	(0.020)	(0.010)	(0.013)
Agreeableness	0.071***	0.030	0.739***	0.022***	-0.005	0.004	0.021*
	(0.017)	(0.021)	(0.011)	(0.007)	(0.020)	(0.010)	(0.013)
Extroversion	-0.030***	-0.021	0.010**	0.835***	0.016	0.002	-0.009
	(0.005)	(0.016)	(0.005)	(0.020)	(0.017)	(0.009)	(0.008)
Openness	-0.086***	-0.052***	-0.031***	-0.053***	0.773***	0.012	0.033
	(0.011)	(0.019)	(0.005)	(0.007)	(0.020)	(0.010)	(0.021)
Math Skill	0.139***	0.105*	0.059***	0.054***	0.020	0.299***	0.092
	(0.008)	(0.061)	(0.015)	(0.006)	(0.059)	(0.032)	(0.063)
Reading Skill	-0.004	0.086*	0.023**	-0.049***	-0.018	0.079***	0.384***
	(0.004)	(0.045)	(0.010)	(0.012)	(0.043)	(0.023)	(0.109)
Parental Investments	-0.020	0.003	0.048***	0.056***	-0.010	0.001	0.043***
	(0.029)	(0.047)	(0.005)	(0.019)	(0.048)	(0.022)	(0.010)

Note. Period t is the late adolescent period and period t+1 is the early adulthood. The sample size is 1,833. Clustered standard errors at household level in parentheses * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

I find strong and statistically significant self-productivity of the five personality traits and the two cognitive skills. For example, a one standard deviation increase in emotional stability in the late adolescent period (17-21) leads to a 0.686 standard deviation increase in emotional stability in early adulthood (age 25-29). A one standard deviation increase in math skill in the late adolescence increases emotional stability in early adulthood by 0.139 standard deviations. The self-productivity of mathematical skill and reading skill are statistically significant at the 1% level, while reading skill displays a slightly greater magnitude of self-productivity than mathematical skill does. Parental investment also exhibits a statistically significant positive marginal effect on the formation of adolescents' agreeableness, extroversion, and reading skill in the next stage.

I also observe evidence for cross-effects between personality traits and cognitive skills, as well as among personality traits and among cognitive skills. Since the measurements of skills and investments are standardized, the estimated parameters of cross-effects are interpreted as elasticities or marginal effects. For instance, a one standard deviation increase in mathematical skill in the previous period raises emotional stability by 0.139 standard deviation in the next stage at the 1% significance level. The marginal effect of mathematical skill on the production of conscientiousness in the next stage is 0.105. This finding is consistent with discoveries from studies in neuroscience and psychology, such as Marek and Dosenbach (2018) and Zacharopoulos et al. (2021), that mathematical education exerts a positive impact on persons' inhibitory control. Inhibitory control, one of a person's core executive functions, is defined as the ability to control one's attention, behavior, thoughts,

or emotions to override a strong internal predisposition or external enticement, and instead behave in a more appropriate way (Diamond, 2013). The concept of inhibitory control is closely associated with emotional stability and conscientiousness.[12] One of the motivations of this paper is to examine this recent discovery in neuroscience using alternative methods in economics.

Openness (open to different experiences) displays a negative marginal effect on other personality traits. In many other studies, openness is found to strengthen the link between negative situations and life satisfaction, because openness is associated with maladaptive emotion regulation strategies, risk-taking, unusual beliefs and experiences, and perceptual dysregulation, which are aspects of a more disordered personality (Joshanloo, 2022). Studies on openness' strengthening effect of negative experience and proneness to disordered personality explain the negative cross-effects of openness on emotional stability and conscientiousness. Openness is also shown to exhibit a greater inclination of disagreeableness regarding common beliefs.

Openness is reported to be strongly involved with one's preference for artistic imagination, novelty, intellectual curiosity, and a tendency toward non-traditional values (Schretlen et al., 2010), while in the HILDA survey, openness is also related to being creative, complex, and imaginative. Nonetheless, I do not find compelling explanations in the psychology literature that support the negative cross-effect of openness on extroversion. My reading of the literature indicates that the cross-effects between openness and other personality traits are still understudied in psychology. Future research on this topic, and more compelling evidence or exposition, would be very useful.

The key assumption for the instrumental variable method to be valid is that the error terms of the first and second measurements of a given trait or skill must be uncorrelated. To test this assumption, I apply Equation (2.8) to explicitly compute the covariances and degrees of correlation between the the error terms of the measurement equations of any two measurements.[13] The correlation matrix of the measurement errors between every pair of two measurements of personality traits is presented in Table 2.3.

[12] In the HILDA data set, emotional stability is defined by the ability to control one's own moods, such as envy, irritability, distress, and jealousy, and maintain a stable temperament. Conscientiousness is defined by the ability to be orderly, systematic, efficient, and organized, as opposed to being sloppy (Summerfield et al., 2023).

[13] I refer the reader to Cunha and Heckman (2008) for more detailed derivations.

Table 2.3: Contemporaneous Correlation Matrices in Measurement Error: Measurements for Personality Traits

Period t	Emotional stability 1	Emotional stability 2	Conscientiousness 1	Conscientiousness 2	Agreeableness 1	Agreeableness 2	Extroversion 1	Extroversion 2	Openness 1	Openness 2
Emotional Stability 1	1.0000	0.0299	0.1234	-0.0100	0.3242	-0.0602	0.1831	-0.0608	-0.1156	0.0034
Emotional Stability 2		1.0000	-0.0026	-0.0019	0.0009	0.0024	0.0292	-0.0063	0.0196	-0.0016
Conscientiousness 1			1.0000	0.0002	0.0771	-0.0041	-0.1424	0.0075	0.0098	-0.0141
Conscientiousness 2				1.0000	-0.0210	-0.0028	0.0065	0.0075	0.0175	0.0016
Agreeableness 1					1.0000	0.0008	-0.1797	0.0127	0.1840	-0.0217
Agreeableness 2						1.0000	0.0004	0.0035	0.0016	-0.0005
Extroversion 1							1.0000	0.2091	0.0718	0.0295
Extroversion 2								1.0000	-0.0011	-0.0103
Openness 1									1.0000	0.0000
Openness 2										1.0000
Period t+1	Emotional stability 1	Emotional stability 2	Conscientiousness 1	Conscientiousness 2	Agreeableness 1	Agreeableness 2	Extroversion 1	Extroversion 2	Openness 1	Openness 2
Emotional Stability 1	1.0000	0.0962	0.3088	0.1442	0.0923	-0.1718	-0.3127	0.1065	0.0186	-0.1664
Emotional Stability 2		1.0000	0.1108	0.3213	-0.0685	0.1115	-0.0586	-0.0261	-0.1093	0.2492
Conscientiousness 1			1.0000	0.0884	0.1591	-0.0998	-0.0372	0.1467	0.0868	-0.1016
Conscientiousness 2				1.0000	-0.0625	0.0493	0.0778	-0.0453	-0.0976	0.1049
Agreeableness 1					1.0000	0.1072	-0.1286	0.1242	0.1864	-0.0635
Agreeableness 2						1.0000	-0.1425	-0.1207	-0.1134	0.0526
Extroversion 1							1.0000	0.2530	-0.1130	0.0171
Extroversion 2								1.0000	0.0609	-0.1253
Openness 1									1.0000	0.0560
Openness 2										1.0000

Note. Period t represents late adolescence. Period t+1 represents early adulthood. The sample size is 2,698.

Table 2.3 displays the estimated contemporaneous correlation matrix across the measurement errors in the indicators of personality traits. In the late adolescence (age 17 to 21), most of the correlations are lower than 0.2, except for one out of forty-five correlations (not including the diagonal values) equal to 0.3242. In the early adulthood (age 25 to 29), most of the correlations are lower than 0.15, except for three out of forty-five correlations (not including the diagonal values) between 0.3 to 0.4. Overall, in both periods, all of the correlations between the measurement errors are below 0.4 in absolute value and most of them are well below 0.15 in absolute value. I also perform Wald tests for joint significance of the covariances between the measurement errors for these two periods. The p-values of the two Wald tests are 0.999 and 0.896 for late adolescent and early adulthood periods, respectively, indicating that there is no evidence to reject the null hypothesis that the covariances between the measurement errors are jointly equal to zero. Both the correlation matrix and results from the Wald tests suggest that the assumption for the instrumental variable method is justified, and that the exclusion restriction condition for valid instrumental variable estimation holds.

Using Equations (2.4) and (2.7) of the dynamic factor model, I compute the contemporaneous correlation matrices of the latent factors and report them in Table 2.4.

Table 2.4: Contemporaneous Correlation Matrices: Cognitive Skills and Personality Traits

Period t	Emotional Stability	Conscientiousness	Agreeableness	Extroversion	Openness	Math	Reading	Investments
Emotional Stability	1.0000	0.2549	0.0342	0.2127	-0.3708	0.1816	0.0522	0.1081
Conscientiousness		1.0000	0.3098	0.0911	0.0553	0.3166	0.1365	0.0344
Agreeableness			1.0000	0.1549	0.3186	0.0423	0.1561	0.0627
Extroversion				1.0000	-0.0022	0.0217	0.0403	0.0676
Openness					1.0000	0.1842	0.3251	0.0332
Math						1.0000	0.7109	0.3448
Reading							1.0000	0.2581
Investments								1.0000

Period t+1	Emotional Stability	Conscientiousness	Agreeableness	Extroversion	Openness	Math	Reading	Investments
Emotional Stability	1.0000	0.2900	0.1013	0.2310	-0.2258	0.1866	0.0260	0.0298
Conscientiousness		1.0000	0.3160	0.1818	0.0952	0.3062	0.1661	0.0369
Agreeableness			1.0000	0.2383	0.3657	0.0995	0.2132	0.0653
Extroversion				1.0000	0.0757	0.0835	0.0741	0.1060
Openness					1.0000	0.2310	0.3370	0.0162
Math						1.0000	0.8178	0.2488
Reading							1.0000	0.1494
Investments								1.0000

Note. Period t represents late adolescence. Period t+1 represents early adulthood. The sample size is 2,698.

There is a strong correlation between math and reading skills, 0.711, at the late adolescent stage (period t), where skills measures from age 21 are used. The correlation between math and reading skills increases to 0.818 at early adulthood stage (period t+1), where skills

measures from age 25 are used. I also find evidence of a medium-sized correlation between conscientiousness and mathematical skill, where the correlation is 0.317 in late adolescent period and 0.306 in early adulthood. This provides further evidence for the finding that mathematical education brings a positive impact on persons' inhibitory control as measured by emotional stability and conscientiousness.

The negative contemporaneous correlation between openness and emotional stability is also worth noting. The rationale stems from the constructs that openness and emotional stability reflect. It was found that openness as a trait was related to maladaptive emotion regulation strategies, risk-taking, and perceptual dysregulation (Joshanloo, 2022). These characteristics associated with openness are the opposite of emotional stability, which characterizes the ability to deal with stress, regulate emotional impulses, and stay calm.

The correlation between family investment and mathematical skill falls from 0.345 to 0.249 between period t and period t+1, indicating that an earlier stage in child development is a relatively more sensitive period (a period or age when investments in children are more productive in producing future traits and skills) to improve adolescents' mathematical skill. My findings are consistent with notions from Cohen (2013) that a correlation coefficient of 0.3 is considered a medium effect size, as Cohen (2013) documented that correlations among representative tests of creativity averaged to almost exactly 0.3.

In the next section of the paper, I delve deeper into investigating the development of adolescents' personality traits and cognitive skills, using a bunching with control function approach. The strength of the dynamic factor model is that it provides a comprehensive view of the human capital formation process and it addresses measurement errors associated with the latent factors. An important motivation for the analysis using the control function approach is to cross-validate the estimates from the dynamic factor model, where an independence assumption is applied to account for potential selection on unobservables. Employing the control function approach, I can explicitly address selection on unobservables by estimating the control functions and controlling for the effect of the selection-on-unobservable. Another motivation for applying the the control function approach is that I take investments as given and estimate the outcomes using the dynamic factor model. The control function approach is built on a maternal labor supply model, where maternal labor supply is another crucial input in the linear production function of skills, in addition to parental investments. To make the estimates from both methods comparable, I construct the control function framework in a way that the two approaches (the dynamic factor model and control function method) are connected through an overarching data generating process.

2.5 Empirical Strategy: A Control Function Approach

2.5.1 Introduction to the Idea of Bunching

The amount of time and care received by children is determined by maternal hours worked, and a household's investments in child development are directly determined by family labor income. Estimating their effects on children's traits and skills is challenging because maternal labor supply and family labor income may be correlated with unobservables that are themselves affecting children's skill production. Therefore, I implement a bunching with control function strategy to address this selection on unobservables issue.

The idea of this approach is to leverage the zero hours to work constraint, where the desired hours to work for mothers who stay home is unobserved. At this bunching point of zero hours to work, there are two types of mothers: the first is mothers whose desired work hours are exactly zero, and the second is mothers whose desired work hours are negative but constrained at zero. Then I utilize a distributional assumption to estimate the desired hours of work and construct a control function that models the source of the endogeneity. By including the control function in the regression framework, I can control for the confounding relationship and obtain unbiased parameter estimates of the impact of maternal hours worked and household labor income on children's traits and skills.

To demonstrate the idea of bunching, I start by establishing a simple constrained labor supply model following the approach of Caetano et al. (2021). A child's skill S is determined by mother's labor supply, L, observable attributes, X, and unobservables, ϵ, as follows:

$$S = f(L, X; \boldsymbol{\beta}) + g(X) + \epsilon \tag{2.13}$$

where L is maternal work hours in the previous period, X is a vector of covariates, and ϵ is the error term that accounts for unobservable factors. The key problem to overcome is that L is endogenous since L and ϵ are correlated conditional on X, and this problem must be addressed to obtain unbiased estimates of the vector of parameters $\boldsymbol{\beta}$.

The central idea of bunching is to employ the fact that a mother's chosen number of hours to work is constrained to be non-negative when deciding the optimal hours to work, because people cannot work for negative hours, even though the desired optimal solution may be in the negative domain. The desired number of hours to work, L^*, can be written as:

$$L^* = h(X) + \eta, \tag{2.14}$$

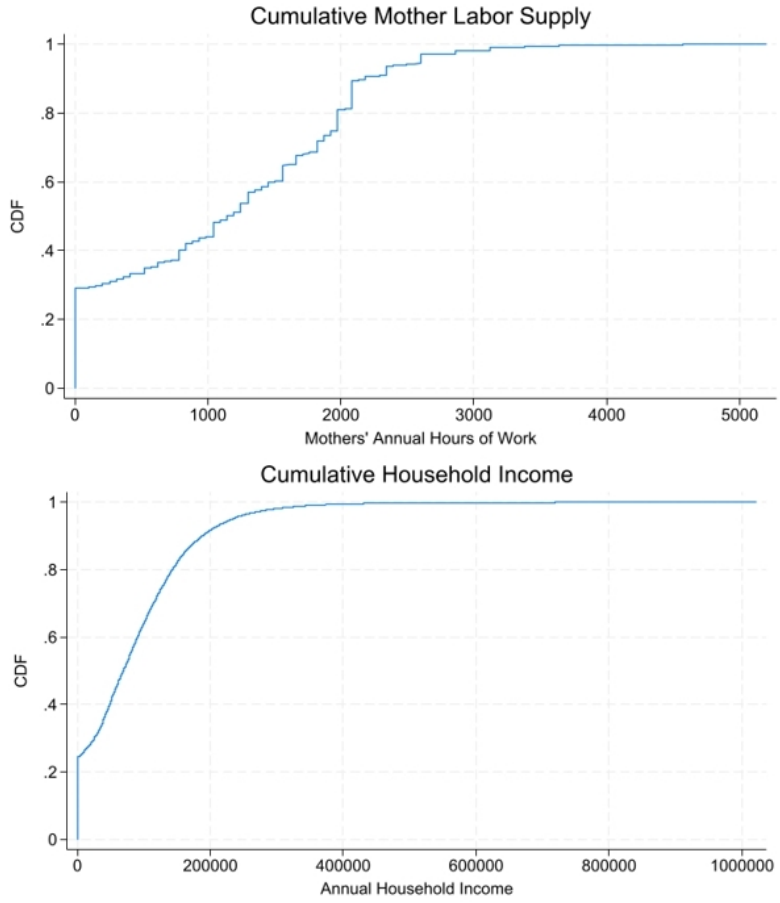
where $L = \max\{0, L^*\}$, reflecting the fact that the desired hours to work is bounded by the constraint of zero. The term η is the collection of unobservables that determine L^*, conditional on observables X. The idea of bunching with respect to family labor income from working is parallel to the constrained labor supply model, where labor supply L is replaced with income I. Due to the fact that the amount of earnings is also bounded to be non-negative, the desired amount of earnings is written as $I = \max\{0, I^*\}$. The type of mothers who desire negative hours worked reasonably overlaps with the type that have a negative desired amount of family labor income (from wages and salary), because when a person works for negative hours (if allowed), his or her expected salary will be negative as well. In this paper, family labor income is referred to income only from wages and salary.

Note that when I estimate the skill formation model, I use parental investments, while I have mother's time and household income when I apply the control function approach. This is because the two identification methods are fundamentally different. The dynamic factor model allows me to obtain parental investments as a latent factor from a set of observable variables related to parental investment and household environment. The control function approach is built upon the constrained labor supply model. Next, I show the evidence of bunching and selection from the HILDA data.

2.5.2 Evidence of Bunching and Selection

Figure 2.1 shows evidence of bunching of maternal hours worked and family labor income at the bunching point zero. The curve in the top panel is the cumulative distribution of mother's annual hours of working, and the curve in the lower panel is the cumulative distribution of annual family labor income.

Figure 2.1: Evidence of Bunching in Maternal Hours Worked and Household Income



Note. There is 29.0% of individuals bunched at zero hours of work. The figure also exhibits number heaping in maternal hours worked at multiples of 5 and 10. For example, about 8% of individuals bunched at 40 hours per week (2080 hours per year). There is 27.6% of households bunched at zero annual household income from working.

Approximately 29% of mothers are bunched at zero hours of working, and about 28% of households are bunched at zero annual family labor income. The proportion of bunching in the data that I use is close to the level (25%) that mothers are bunched at zero hours worked found in Caetano et al. (2021). The cumulative distribution of mother's annual hours worked exhibits number heaping at multiples of 5 and 10, because individuals' working hours are surveyed on a weekly basis.[14] For example, about 8% of individuals reported that they worked 40 hours per week (2080 hours per year).

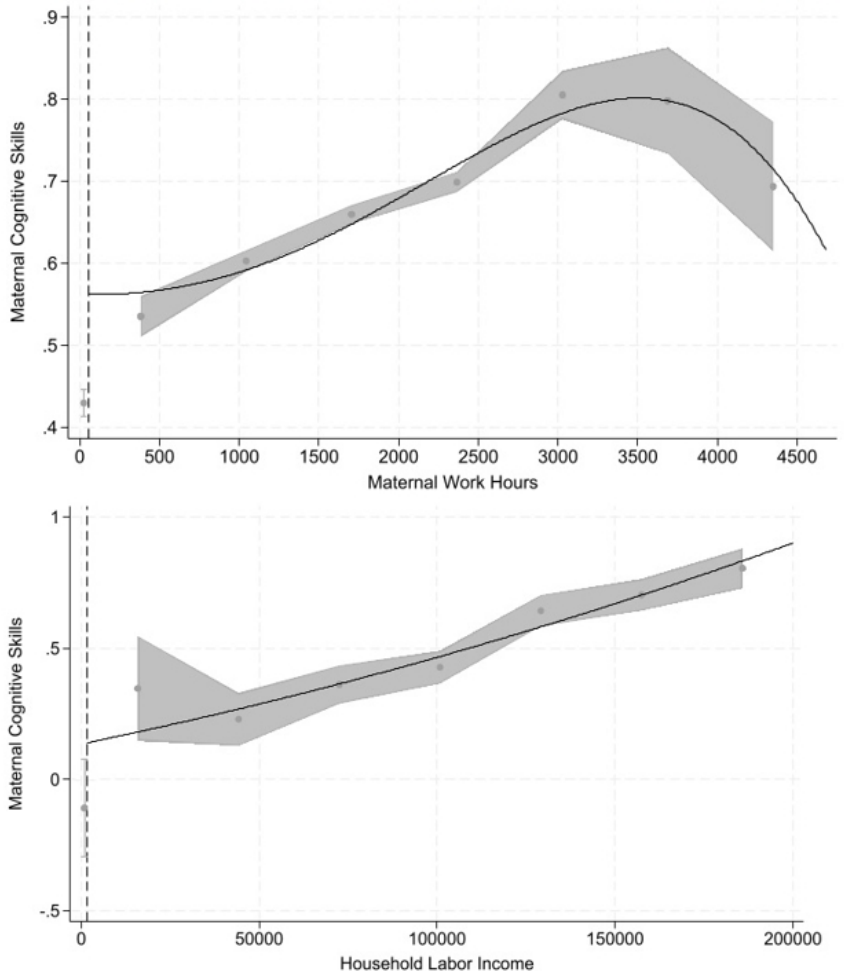
The evidence of bunching in maternal hours worked and family labor income shown in the previous figure may be attributed to individuals whose true type is exactly at $L^* = 0$. Nonetheless, this hypothesis makes sense only if the observable maternal characteristics of

[14]In addition to number heaping, measurement error could also be one of the reasons.

those who bunch at $L^* = 0$ are similar to the characteristics of those who work for a small amount of time where $L^* = l$ for reasonably small $l > 0$. The same reasoning also goes for income.

This hypothesis is rejected, as I observe that mother's cognitive skills at $L^* = 0$ and $I^* = 0$ are distinctly different then $L^* = l$ and $I^* = i$, respectively, as shown in Figure 2.2.

Figure 2.2: Mothers Work Zero Hours are Discontinuously Different than Other Mothers



Note. The top panel uses a regression discontinuity method to check the limit as maternal hours to work approaches zero with work hours at zero. The bottom panel uses a regression discontinuity method to check the limit as household labor income approaches zero with income at zero. Confidence intervals are at the 95%.

The top panel is created from a regression discontinuity method to compare the maternal cognitive skills as maternal hours to work approaches zero with the maternal cognitive skills of mothers with zero work hours. The bottom panel also uses a regression discontinuity method to compare the maternal cognitive skills as household labor income approaches zero

with the maternal cognitive skills of mothers with zero labor income. Confidence intervals at $L = 0$, $I = 0$, and the shaded confidence bands are at the 95% significance level.

Figure 2.2 shows that the confidence intervals of mother's cognitive skills at $L = 0$ and $I = 0$ do not overlap with the corresponding confidence bands or the global polynomial fits. This provides evidence that mothers' cognitive skills of those who bunch at $L^* = 0$ and $I^* = 0$ are dissimilar to the skills of those who work at $L^* = l$ for a reasonably small $l > 0$, and of those who have $I^* = i$ for a reasonably small $i > 0$. This suggests that mothers who work for exactly zero hours and whose family labor incomes are exactly zero tend to have discontinuously lower cognitive skills and thus are distinctly different from those who are in the positive domain. Thus the data reject the hypothesis that the bunching at $L^* = 0$ and $I^* = 0$ simply reflects a group of mothers whose true preference is exactly at $L^* = 0$ and $I^* = 0$ because, if so, their characteristics should be comparable to those with small and positive $L^* = l$ and $I^* = i$. This finding suggests that there is likely a notable share of mothers who have desired L^* and I^* in the negative domain. The finding also suggests positive selection of mother's labor supply and family labor income close to $L^* = 0$, meaning that mothers who work for a small and positive amount of hours and mothers in households with a small and positive family labor income tend to have sharply larger magnitudes of covariates that are positively associated with the development of adolescents' cognitive and noncognitive skills.

2.5.3 Constructing the Control Function

To address the endogeneity in maternal labor supply and family labor income, I formulate control functions that explicitly model the source of endogeneity to obtain unbiased parameter estimates. Building upon the constrained labor supply model in Equation (2.13), I decompose the error term into a selection-on-unobservable part, $\delta(X)\eta$, and an uncorrelated part, ε, such that $\epsilon = \delta(X)\eta + \varepsilon$. Thus, the skill development function is re-written as

$$S = f(L, X; \beta) + g(X) + \delta(X)\eta + \varepsilon, \text{ with } \mathbb{E}[\varepsilon \mid L, X, \eta] = 0 \tag{2.15}$$

The children whose mothers bunch at $L = 0$ receive the same "treatment" of $L = 0$, since the lower bound of labor supply is constrained at zero. After controlling for the vector of observables X, the remaining structural differences in S for children who have the same value of $L = 0$ and the same observables X are attributed to differences in η. Then I can build a control function to identify the unbiased β by controlling for the effect of the

selection-on-unobservable part, η, on skill S.

To construct the control function, I combine Equations (2.14) and (2.15), which yields the following expression for skill S:[15]

$$\mathbb{E}[S \mid L, X] = f(L, X; \beta) + g(X) - \delta(X)h(X) + \delta(X)\left[L + \mathbb{E}\left[L^* \mid L = 0, X\right]\mathbf{1}(L = 0)\right] \tag{2.16}$$

This updated expression shows that I can use $(L + \mathbb{E}\left[L^* \mid L = 0, X\right]\mathbf{1}(L = 0))$ as a control function to address selection on unobservables after $\mathbb{E}\left[L^* \mid L = 0, X\right]$ is estimated. The bunching with control function method requires a key distributional assumption in order to identify $\mathbb{E}\left[L^* \mid L = 0, X\right]$. According to Caetano et al. (2021), the distributional assumption is formally established as: for all censored quantiles q_0, $\eta \mid X$ has symmetric tails below q_0 and above $1 - q_0$. Intuitively, the statement assumes symmetry in the distribution of unobservables η conditional on the vector of observables X.

To estimate $\mathbb{E}\left[L^* \mid L = 0, X\right]$, the first step is to use clustering to form a finite partition of the support of X into K clusters, $\hat{\mathcal{C}}_K$.[16] Two observations are clustered in the same set if they have similar observables X, which is similar to the idea of matching. As the number of clusters increase, the observations in the same cluster have increasingly closer values of X. In the estimation of $\mathbb{E}\left[L^* \mid L = 0, X\right]$, X is replaced with clusters $\hat{\mathcal{C}}_K$. Then the expectation term $\mathbb{E}\left[L^* \mid L = 0, X\right]$ can be estimated by

$$\hat{\mathbb{E}}\left[L^* \mid L = 0, X \in \hat{\mathcal{C}}_k\right] = \hat{F}^{-1}_{L|\hat{\mathcal{C}}_k}\left(1 - \hat{F}_{L|\hat{\mathcal{C}}_k}(0)\right) - \hat{\mathbb{E}}\left[L \mid L \geq \hat{F}^{-1}_{L|\hat{\mathcal{C}}_k}\left(1 - \hat{F}_{L|\hat{\mathcal{C}}_k}(0)\right), X \in \hat{\mathcal{C}}_k\right] \tag{2.17}$$

where $\hat{F}_{L|\hat{\mathcal{C}}_k}(0)$ is the proportion of bunched-at-zero observations in cluster k shown in the data, $\hat{F}^{-1}_{L|\hat{\mathcal{C}}_k}\left(1 - \hat{F}_{L|\hat{\mathcal{C}}_k}(0)\right)$ is the labor supply estimated by plugging the proportion calculated by $\left(1 - \hat{F}_{L|\hat{\mathcal{C}}_k}(0)\right)$ into the inverse of the empirical cumulative distribution of L for observations in cluster k, and the last expectation term is computed as the sample average of L in cluster k of those observations with L greater than labor supply $\hat{F}^{-1}_{L|\hat{\mathcal{C}}_k}\left(1 - \hat{F}_{L|\hat{\mathcal{C}}_k}(0)\right)$.

The control function for family income $(L + \mathbb{E}\left[L^* \mid L = 0, X\right]\mathbf{1}(L = 0))$ can be computed in a similar fashion, under the distributional assumption that assumes symmetry in the

[15]The derivation of Equation (2.16) is as follows: $\mathbb{E}[S \mid L, X] = \mathbb{E}[f(L, X; \beta) \mid L, X] + \mathbb{E}[g(X) \mid L, X] + \mathbb{E}[\delta(X)(L^* - h(X)) \mid L, X] + \mathbb{E}[\varepsilon \mid L, X] = f(L, X; \beta) + g(X) + \delta(X)[(L + \mathbb{E}\left[L^* \mid L = 0, X\right]\mathbf{1}(L = 0)) - h(X)] = f(L, X; \beta) + g(X) - \delta(X)h(X) + \delta(X)\left[L + \mathbb{E}\left[L^* \mid L = 0, X\right]\mathbf{1}(L = 0)\right]$.

[16]I partitioned the support of X into $K = 50$ clusters. In one of the robustness checks, I present the coefficients of the main estimation with varying numbers of clusters from 10 to 100, also using updated cluster indicators and re-computed control functions. The estimates are stable across varying numbers of clusters.

distribution of unobservables conditional on the vector of observables X.

2.5.4 Empirical Framework

Given the estimate of the expectation term $\mathbb{E}[L^* \mid L = 0, X]$, Equation (2.16) can be estimated using this general regression framework,

$$S = \beta_L L + X'\tau + \sum_{k=1}^{K} \alpha_k 1\left(X \in \hat{\mathcal{C}}_k\right) + \delta_L \left[L + \hat{\mathbb{E}}\left[L^* \mid L = 0, \hat{C}_K\right] \mathbf{1}(L = 0)\right] \tag{2.18}$$

where β_L is the coefficient of interest, X is a vector of controls consisting of mother's math skill, mother's reading skill, mother's personality traits, mother's education, father's education, and children's sex and age, and $\mathbf{1}\left(X \in \hat{\mathcal{C}}_K\right)$ is the cluster indicator for cluster $\hat{\mathcal{C}}_K$.

The first term $f(L, X; \beta)$ in Equation (2.16) is estimated by a linear specification of $\beta_L L + X'\tau$ in the regression framework. The second and third terms $(g(X) - \delta(X)h(X))$ in Equation (2.16) are estimated via $X'\tau + \sum_{k=1}^{K} \alpha_k 1\left(X \in \hat{\mathcal{C}}_k\right)$, where $X'\tau$ linearly controls for the effect of observables $g(X)$, and the differences between clusters, $\delta(X)h(X)$, are nonparametrically accounted for through the cluster indicators. The final term in Equation (2.16), $\delta_L \left[L + \hat{\mathbb{E}}\left[L^* \mid L = 0, \hat{C}_K\right] \mathbf{1}(L = 0)\right]$, is estimated by Equation (2.17), and it is the control function term included in the regression to address selection on unobservables.

However, Equation (2.18) is not the final regression framework of the analysis, since household income and adolescents' traits and skills in the previous period are yet to be included. Building on Equation (2.18), I add the term of household labor income $\beta_I I$ and the control function for household labor income to obtain unbiased estimates of the impact of maternal hours worked, β_L, and of the impact of family labor income, β_I on children's skills outcomes.

Moreover, an important motivation for the analysis using the control function approach is to cross-validate the estimates from the dynamic factor model, where an independence assumption is applied to address potential selection on unobservables. Therefore, I construct the control function framework in a way that the two approaches (the dynamic factor model and control function method) are connected through an overarching data generating process. Thus, I include adolescents' traits and skills from the previous period (age 21), $\sum_{j=1}^{7} S'_{j,t-1}\xi_j$, in the regression framework, so that the estimates from the control function approach are directly comparable to those from the dynamic factor model. The final regression framework

is written as follows:

$$
\begin{aligned}
S_{i,t} = \beta_L L + \beta_I I + \sum_{j=1}^{7} S'_{j,t-1}\xi_j + X'\tau + \delta_L \left[L + \hat{\mathbb{E}} \left[L^* \mid L = 0, \hat{C}_K \right] \mathbf{1}(L = 0) \right] \\
+ \delta_I \left[I + \hat{\mathbb{E}} \left[I^* \mid I = 0, \hat{C}_K \right] \mathbf{1}(I = 0) \right] + \sum_{k=1}^{K} \alpha_k 1 \left(X \in \hat{\mathcal{C}}_k \right),
\end{aligned}
\tag{2.19}
$$

where β_L, β_I, and ξ_j are the coefficients of interest. Maternal hours worked, L, and family income, I, are standardized values and are from the previous period (21-24) of individuals' early adulthood. Term X is a vector of controls including mother's math skill, mother's reading skill, mother's personality traits, mother's education, father's education, and children's sex and age. Two control functions are included (one for maternal labor supply and one for family labor income). The clusters, $\hat{\mathcal{C}}_K$, and corresponding cluster indicators are the same as in the prior regression.

2.6 Empirical Results: Control Function Approach

2.6.1 Main Results

I estimate the regression in Equation (2.19) and present the estimated coefficients β_L, β_I, and ξ_j, in Table 2.5. The coefficient β_L is the impact of maternal hours worked in the previous period (age 21-25) on personality traits and cognitive skills in early adulthood (age 25). The coefficient β_I is the impact of family labor income in the same previous period on traits and skills in early adulthood. The coefficient ξ_j is the impact of the traits and skills in late adolescence (age 21) on traits and skills in early adulthood, which measures cross-effects and self-productivity. Table 2.5 is constructed in a similar way as Table 2.2, where the inputs from the previous period are located in the column of "Period t", and outputs in adulthood are located in the row of "Period t+1".

Table 2.5: Estimates from the Control Function Approach

Period t \ Period t+1	Emotional Stability	Conscientiousness	Agreeableness	Extroversion	Openness	Math	Reading
Emotional Stability	0.684***	0.022*	0.008	-0.045***	-0.024*	0.021	-0.004
	(0.016)	(0.013)	(0.013)	(0.012)	(0.012)	(0.017)	(0.016)
Conscientiousness	0.112***	0.788***	0.035***	0.010	-0.036***	0.057***	0.044**
	(0.015)	(0.014)	(0.012)	(0.013)	(0.012)	(0.017)	(0.018)
Agreeableness	0.091***	-0.003	0.801***	0.074***	-0.007	-0.027	0.021
	(0.015)	(0.015)	(0.016)	(0.014)	(0.014)	(0.017)	(0.018)
Extroversion	-0.019	-0.025**	0.018*	0.834***	0.010	-0.003	-0.004
	(0.013)	(0.012)	(0.011)	(0.011)	(0.011)	(0.017)	(0.014)
Openness	-0.039**	-0.077***	-0.039***	-0.055***	0.771***	-0.019	0.064***
	(0.015)	(0.014)	(0.014)	(0.013)	(0.014)	(0.015)	(0.016)
Math Skill	0.036**	0.012	-0.018	0.002	-0.015	0.522***	-0.005
	(0.014)	(0.014)	(0.012)	(0.013)	(0.012)	(0.016)	(0.017)
Reading Skill	-0.027	0.058***	0.063***	0.012	0.052***	0.343***	0.482***
	(0.017)	(0.014)	(0.016)	(0.015)	(0.015)	(0.021)	(0.021)
Maternal Work Hours	-0.023*	-0.016	0.003	0.015	0.001	-0.021	0.002
	(0.014)	(0.013)	(0.012)	(0.012)	(0.012)	(0.016)	(0.017)
Family Income	0.002	0.012	0.017*	0.025***	0.015*	-0.010	0.045***
	(0.010)	(0.011)	(0.009)	(0.008)	(0.009)	(0.012)	(0.013)

Note. Period t is the late adolescent period and period t+1 is the early adulthood. The sample size is 3,799. Clustered standard errors at household level in parentheses * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

The estimates from the control function approach are largely consistent with the estimates based on the dynamic factor model. I find statistically significant evidence of self-productivity and cross-effects for both personality traits and cognitive skills. The values of the self-productivity estimates in Table 2.5 are close to the values in Table 2.2, and all of them are statistically significant at the 1% level. Since the personality traits, cognitive skills, maternal hours worked, and family labor income are standardized, the estimates reflect elasticities or marginal effects. For example, a one standard deviation increase in emotional stability in late adolescence, holding all other factors constant, leads to an increase in adult emotional stability of 0.684 standard deviations.

The cross-effect between adolescents' math skill and emotional stability is positive and statistically significant at the 5% level. This provide complementary evidence to the neuroscience research finding that math education exhibits a positive impact on students' inhibitory control, which is a trait closely related to emotional stability (Zacharopoulos et al., 2021).

The effect of a higher level of family labor income in the previous period is positive and statistically significant for the reading skill and for three of the five personality traits. For example, a one standard deviation increase in family labor income during late adolescence causes adolescents' level of extroversion to increase by 0.025 standard deviations in early adulthood. The income effect that I find in this period, especially on reading skill, may be

attributed to parental payments for adolescents' college tuition. Unfortunately, the HILDA data set does not provide the information on college tuition payments, which prevents me from examining the impact through this channel.

The impact of maternal hours worked in the previous period on adolescents' traits and skills in early adulthood is generally statistically insignificant, except that the effect on adolescents' emotional stability is significant at the 10% level. Specifically, a one standard deviation increase in maternal hours worked in late adolescence, holding all other factors constant, reduces adolescents' emotional stability by 0.023 standard deviations in adulthood.[17] This result is consistent with developmental psychology studies finding that maternal caregiving affects adolescents' emotional regulation in two key ways: stress response dysregulation and psychological security. Regarding stress response dysregulation, Ding et al. (2022) find that maternal care buffers adolescents' stress responses, and adolescents may have an overactive stress response system lacking maternal care.[18] Chronic stress can impair the ability to cope with emotional challenges, increasing the risk of emotional instability and impulsive reactions. Regarding psychological security, Yoder et al. (2019) discover that adolescents may feel insecure without consistent maternal support and secure attachment. This sense of instability can manifest in emotional instability, as adolescents become more prone to feelings of rejection, anxiety, or anger when navigating social relationships. These two connections between maternal caregiving and adolescents' emotional regulation explain the negative impact of maternal work hours on adolescents' emotional stability.

Based on both Table 2.5 and Table 2.2, I find that personality traits are malleable to investments and to the environment in late adolescence. The evidence is robust to I addressing selection on unobservables, using the control function approach. This discovery confirms the findings of the earlier literature, which suggest that while cognitive skills mature at earlier stages, noncognitive skills remain more adaptable and can be modified until later years (Cherniss et al., 1998; Carneiro and Heckman, 2003; Boyatzis, 2008).

Furthermore, when the estimates using the control function method are compared to those from the dynamic factor model in Section 2.4, the results are similar. For example, the effects of parental investments and family labor income on the formation of personality traits and cognitive skills are largely consistent, especially for agreeableness, extroversion,

[17] One standard deviation of yearly maternal hours worked is 897.26 hours.

[18] When a person's stress response system is overactive due to continuous or excessive stress, it triggers the release of cortisol more frequently or keeps it elevated for longer periods. Maternal engagement is associated with reduced overall cortisol levels in children during their recovery from emotional arousal (Blair et al., 2008).

and reading skill. The self-productivity of the traits and skills, found using both approaches, have similar values and statistical significance. The fact that my findings from the skill formation model and from the control function approach are similar points to the emerging view that structural models and reduced-form analyses can be complementary to each other in useful ways (Todd and Wolpin, 2023). In this paper, the dynamic factor model and the control function approach are connected through an overarching data generating process. The dynamic factor model explicitly addresses measurement errors associated with the latent factors, and the control function approach accounts for maternal labor supply as another input, in addition to parental investments, and it addresses potential selection on unobservables.

In the next subsection, I relate personality traits and cognitive skills in early adulthood to adult earnings. In Subsection 2.6.3, I compute whether the income effect offsets the impact of mothers' labor supply on adolescents' outcomes.

2.6.2 The Effect of Early Adulthood Skills on Earnings

In order to provide greater economic insights into the development of adolescents' personality traits and cognitive skills, I estimate the impact of the stocks of traits and skills in the final period of child development (early adulthood at age 25) on early adult earnings (average income from age 25 to 29). This analysis is related to the "anchoring" concept from Cunha and Heckman (2008). Moreover, these estimates will be used to calculate whether the positive income effect offsets the negative effect of maternal work hours on adolescents' outcomes.

I estimate the following OLS regression:

$$Earnings = \mu_k + \alpha_k \theta_T^k + \eta_k, \tag{2.20}$$

where the adult earnings, $Earnings$, are measured by the standardized average income from age 25 to 29, and θ_T^k is the individual's standardized trait or skill in the final period of the child development process (early adulthood at age 25). I employ the same instrumental variable method as before to address the measurement error issue, where I use a person's skill at age 21 to instrument for the same skill at age 25, and then use the predicted $\hat{\theta}_T^k$ to estimate α_k.[19] The estimation results are presented in Table 2.6.

[19] I do not use a person's skill at age 25 to instrument for the same skill at age 29, due to the possibility of reverse causality, which is the effects of earnings on cognitive and noncognitive skills. Using skills at age

Table 2.6: Adult Earnings and the Stocks of Early Adulthood Traits and Skills

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Emotional Stability	0.129***							0.059**
	(0.022)							(0.023)
Conscientiousness		0.176***						0.154***
		(0.022)						(0.025)
Agreeableness			0.015					-0.062**
			(0.023)					(0.025)
Extroversion				0.122***				0.094***
				(0.020)				(0.020)
Openness					-0.009			-0.037
					(0.020)			(0.022)
Math Skill						0.192***		0.087*
						(0.051)		(0.051)
Reading Skill							0.354***	0.337***
							(0.055)	(0.056)
Observations	4628	4628	4628	4628	4628	4628	4628	4628

Note: Columns (1) to (7) present the parameter of interest from regressions where adult earnings are regressed on each trait or skill separately. Column (8) is the regression that includes all the traits and skills on the right hand side. Standard errors (clustered at the individual level) in parentheses * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

In the first seven columns, I regress adult earnings on each early adulthood trait or skill separately. In the last column, I regress adult earnings on all of the early adulthood traits and skills together. Most of the early adulthood traits and skills are positively related to adulthood earnings, with high statistical significance, except for agreeableness and openness. Since the regressors and earnings are standardized, the coefficients are interpreted as elasticities or marginal effects. For instance, a one standard deviation increase in adulthood conscientiousness increases adulthood earnings by 0.154 standard deviations, if I refer to the results from the combined regression. Both math and reading skills in early adulthood predict adulthood earnings, and the effect from the reading skill has a larger magnitude. Overall, the magnitudes of the effects of the Big Five personality traits on adult earnings are also economically large, even when they are compared to the magnitudes of the impacts of cognitive skills.

2.6.3 Computation of the Net Effect

In most societies, the burden of child rearing responsibilities still falls primarily on women, and this unequal burden not only perpetuates gender disparities in the labor market but also limits mothers' earning potential and career advancement (Lundborg et al., 2017). Therefore, in this subsection, I examine whether mother's labor supply has a net positive

25, instrumented using skills at age 21, for earnings from age 25 to 29 can reduce the possibility that I am measuring the reverse causality.

or negative effect on adolescents' outcomes, in order to address working mothers' concerns related to the choice of work.

First, I run an OLS regression to predict how maternal hours worked affect family income, where I regress family income (in Australian dollars) on the number of maternal hours worked. Then I compute the impacts of a hypothetical one standard deviation increase in mothers' working hours on adolescents' outcomes through two channels: the indirect family income channel and the direct maternal labor supply channel. Adding up the impacts from both channels provides the net effect. The estimates from the OLS regression that predicts the effect of maternal hours worked on family income are presented in Table 2.7.

Table 2.7: Predicting Household Income Using Maternal Hours Worked

	(1)	(2)	(3)	(4)
Maternal Hours Worked	24.336***	24.336***	20.964***	20.964***
	(1.376)	(2.597)	(1.493)	(2.679)
Controls	No	No	Yes	Yes
Cluster-robust standard errors	No	Yes	No	Yes
Observations	6,625	6,625	5,672	5,672

Note: Standard errors (clustered at household level) in parentheses * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

The effect of maternal hours worked on family income is statistically significant at the 1% level. The estimates from the first and the second columns in Table 2.7 indicate that for every one hour increase in maternal hours worked, family income is increased by 24.35 Australian dollars. After controlling for mothers' cognitive skills, mothers' education, fathers' education, and mothers' age, for a one hour increase in maternal hours worked, family income increases by 20.96 Australian dollars (as shown in the third and the fourth columns in Table 2.7).

To compute the the impact of a one standard deviation increase in mothers' working hours on adolescents' outcomes through the maternal hours worked channel, I first calculate how a one standard deviation increase in maternal hours worked affects adolescents' personality and skill outcomes, based on Table 2.5. In this case, a one standard deviation increase in maternal hours worked reduces adolescents' emotional stability by 0.023 standard deviations. Through the maternal hours worked channel, adolescents' emotional stability is the only statistically significant outcome. Next, I calculate how a 0.023 standard deviation decrease in adolescents' emotional stability reduces adult earnings, using Table 2.6. Using the estimate from the combined regression, a 0.023 standard deviation decrease in adolescents' emotional stability reduces adult earnings by approximately $0.059 * 0.023 \approx 0.0014$ standard deviations.

To calculate the the impact of a one standard deviation increase in mothers' working hours on adolescents' outcomes through the family income channel, I first compute how much a one standard deviation increase in maternal hours worked increases family income, using the summary statistics in Table 2.1. One standard deviation of maternal hours worked is 897.26 hours, and one standard deviation of family income is 63,884.83 Australian dollars. Thus, a one standard deviation increase in maternal hours worked leads to approximately a $24.34 * 897.26/63884.83 \approx 0.34$ standard deviation increase in family income, where the estimate of 24.34 is from Table 2.7. Next, a 0.34 standard deviation increase in family income increases, for example, adolescents' extroversion by $0.34 * 0.025 = 0.0085$ standard deviations, based on Table 2.5. Then, a 0.0085 standard deviation increase in adolescents' extroversion leads to an increase in adult earnings of $0.0085 * 0.094 \approx 0.0008$ standard deviations, based on Table 2.6. I repeat the same computation process for agreeableness and reading skill, as family income exhibits statistically significant impact on them (as shown in Table 2.5). I do not repeat the computation for openness, since openness does not have a statistically significant impact on adult earnings (as shown in Table 2.6). Finally, given the effects through both the family income channel and the maternal hours worked channel, I sum up all the effects, which yields the total net effect. I present the effects from both channels and the total net effect in Table 2.8.

Table 2.8: T-Accounts of One S.D. Increase in Maternal Work Hours

Panel A.	Maternal Labor Effect	Household Income Effect
Via Emotional Stability	-0.0014	
Via Agreeableness		-0.0004
Via Extroversion		0.0008
Via Reading		0.0052
Total Effect		0.0043

Panel B.	Maternal Labor Effect	Household Income Effect
Via Emotional Stability	-0.0014	
Via Agreeableness		-0.0003
Via Extroversion		0.0007
Via Reading		0.0044
Total Effect		0.0034

Note. This table presents the calculation of the total effect of one standard deviation increase in maternal work hours on youths' adult earnings (in standard deviation)

In Panel A of Table 2.8, the computation of the effects via the household income channel is based on the estimate of 24.34 in Table 2.7. I also repeat the same calculation using the estimate of 20.96 in Table 2.7, and show the results in Panel B of Table 2.8. I find that

the overall positive income effect outweighs the (marginally significant) negative impact of maternal hours worked, resulting in a net positive effect, meaning that maternal labor supply during children's late adolescence provides a net positive impact on children's adult earnings. Despite small magnitudes, the net total effects are positive for both panels.

This finding provides meaningful implications for working mothers' choice of work in the U.S. labor market, where women's labor force participation rate is approximately 10% lower than men's, as in Australia (where the HILDA data are collected), women's labor force participation rate is also approximately 10% lower than men's. In terms of child rearing responsibilities, these two labor markets are comparable, given the similar differences in labor force participation rates between women and men.[20]

2.7 Robustness Checks

To examine the robustness of the findings, I implement three robustness checks on the estimations. The first robustness check (the translog skill production function) is applied to the estimation using the skill formation model. The second and third robustness checks are applied to the control function approach.

2.7.1 Transcendental Logarithmic Skill Production Function

In the estimation of the skill formation model, I implemented a linear specification of skill production function for simplicity and stability. It is critical to note that if little evidence of self-productivity or cross-effects is found using a more parsimonious linear model, it is unlikely that findings from a more elaborate non-linear setting will overturn the results (Cunha and Heckman, 2008).

It is, on the other hand, a good practice to employ a non-linear skill production to check the robustness of the findings from a more parsimonious specification. The most recent and rigorous studies in the human capital formation literature leverage non-linear models to investigate skill formation, often times utilizing the constant elasticity of substitution (CES) skill production function (Cunha et al., 2010; Attanasio et al., 2020; Aucejo and James, 2021).

[20] In the U.S., the labor force participation rate is 57.3% among women and is 68.1% among men in 2023, according to FRED. In Australia, the labor force participation rate is 61.5% among women and is 71.4% among men in 2023, according to the World Bank.

For this robustness check, I apply a transcendental logarithmic (translog) skill production function, based on the framework of Agostinelli and Wiswall (2023) as an approximation to the CES model. The translog estimator is more simple, tractable, and robust than estimating a formal CES model (Agostinelli and Wiswall, 2023). The translog model is also a well-known approach to approximate a CES function using a second order Taylor expansion (Christensen et al., 1973; Hoff, 2002).

I employ the following translog skill production function as an approximation to the formal CES function and use the same IV estimator that I used in the linear specification.

$$\ln \theta_{i,t+1} = \sum_{j=1}^{7} \alpha_{j,t} \ln \theta_{j,t} + \alpha_{k,t} \ln I_t + \sum_{j=1}^{7} \alpha_{m,t} \left(\ln \theta_{j,t} \ln I_t \right) \tag{2.21}$$

I analyze self-productivity and cross-effects of the seven traits and skills (math, reading, and the the Big Five personality traits) in total, and each skill in period t is represented by $\theta_{i,t}$. Parental investments in period t is written as I_t. In Table 2.10, I present the estimated self-productivity and cross-effects of the seven traits and skills and parental investments, where period t is the late adolescence and period $t+1$ is the early adulthood.

Table 2.9: Estimates of the Translog Technology of Skill Formation

Period t / Period t+1	Log Emotional Stability	Log Conscientiousness	Log Agreeableness	Log Extroversion	Log Openness	Log Math	Log Reading
Log Emotional Stabilit	0.479***	-0.007	0.006	0.021	-0.019	-0.021	0.054**
	(0.041)	(0.023)	(0.022)	(0.028)	(0.026)	(0.024)	(0.027)
Log Conscientiousness	0.164***	0.672***	0.079***	0.177***	-0.010	-0.001	0.019
	(0.029)	(0.034)	(0.024)	(0.026)	(0.030)	(0.025)	(0.030)
Log Agreeableness	0.109***	0.098***	0.733***	0.127***	0.086**	-0.027	0.048
	(0.025)	(0.026)	(0.031)	(0.029)	(0.036)	(0.027)	(0.035)
Log Extroversion	0.014	0.022	-0.012	0.446***	0.058**	0.067***	-0.027
	(0.024)	(0.024)	(0.025)	(0.034)	(0.029)	(0.024)	(0.028)
Log Openness	-0.097***	-0.103***	-0.051***	-0.032*	0.564***	-0.019	-0.014
	(0.021)	(0.019)	(0.017)	(0.019)	(0.035)	(0.025)	(0.028)
Log Math Skill	0.201***	0.156***	0.145***	0.094***	0.164***	0.235***	-0.036
	(0.045)	(0.037)	(0.030)	(0.035)	(0.041)	(0.054)	(0.057)
Log Reading Skill	0.053*	0.047*	-0.002	-0.020	-0.065**	0.078**	0.358***
	(0.028)	(0.025)	(0.020)	(0.027)	(0.030)	(0.033)	(0.035)
Log Investment	-0.003	0.003	0.007	0.006	0.013*	-0.008	0.006
	(0.006)	(0.007)	(0.006)	(0.006)	(0.007)	(0.007)	(0.008)
Log Emotional Stabilit	0.063***	-0.012***	-0.008*	-0.017***	0.004	0.004	-0.007
x Log Investments	(0.007)	(0.004)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)
Log Conscientiousness	-0.024***	0.044***	-0.016***	-0.031***	0.001	-0.000	0.001
x Log Investments	(0.005)	(0.007)	(0.004)	(0.005)	(0.005)	(0.005)	(0.006)
Log Agreeableness	-0.017***	-0.018***	0.029***	-0.030***	-0.030***	0.004	-0.009
x Log Investments	(0.005)	(0.005)	(0.006)	(0.006)	(0.007)	(0.006)	(0.007)
Log Extroversion	-0.008**	-0.001	0.004	0.090***	-0.010**	-0.009**	0.004
x Log Investments	(0.004)	(0.004)	(0.004)	(0.006)	(0.005)	(0.004)	(0.005)
Log Openness	0.004	-0.001	0.001	-0.001	0.043***	0.004	0.006
x Log Investments	(0.003)	(0.003)	(0.003)	(0.003)	(0.005)	(0.004)	(0.004)
Log Math Skill	0.001	-0.001	-0.004	-0.010**	-0.010**	0.007	0.007
x Log Investments	(0.005)	(0.006)	(0.003)	(0.004)	(0.005)	(0.005)	(0.006)
Log Reading Skill	-0.011***	-0.005	-0.000	-0.002	-0.000	-0.004	0.004
x Log Investments	(0.003)	(0.003)	(0.002)	(0.003)	(0.003)	(0.003)	(0.004)

Note. Period t stands for the late adolescence and period t+1 is the early adulthood. The sample size is 1,833. Clustered standard errors at the household level in parentheses * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

When the self-productivity and cross-effects estimated from the translog specification are compared to the parameters estimated using the linear model, the estimates are largely consistent despite applying a different skill production function. None of the major conclusions is overturned. The key discoveries remain stable as well, such as the self-productivity of each skill, the positive impact of mathematical skill on the formation of emotional stability, and cross-effects among other personality traits.

The estimated parameters of the interaction terms between each skill and investments also introduce important insights on the heterogeneous elasticities of skill production with respect to investment. The heterogeneous investment elasticities provide useful policy implications about what policy interventions would have a larger effect on skill-disadvantaged adolescents and what interventions would have a larger effect on skill-advantaged adolescents. For

instance, the off-diagonal negative coefficients on the interaction terms between skills and investments indicate that investments have a larger cross-effect on skill-disadvantaged adolescents in terms of developing a related skill. On the other hand, the on-diagonal positive coefficients on the interaction terms between skills and investments suggest that investments have a larger effect on skill-advantaged adolescents in terms of self-productivity of the same skill.

Due to limited length of a paper and a departure from the intended research question, I have to leave many other analyses to my future work. First, I can utilize the estimated coefficients from the translog specification to calibrate the skill multiplier and the degree of complementarity parameters of the approximated CES function. Second, given the estimates, I am allowed to conduct policy experiments and examine their impact on adolescents' skill development and adult outcomes, for policies such as but not limited to: child allowances and tax credits, income transfers, subsidized child care, and education subsidies. Third, I am able to perform cost-benefit analyses for whether policy interventions mentioned above are justified given the cost and gains. It is economically meaningful to identify what policy interventions are more cost-effective.

2.7.2 OLS without Control Functions

To obtain a better understanding of selection on unobservables and to analyze the direction of biasness with regard to the estimation of the control function approach, I present the estimates of β_L, β_I, and ξ_j in a simple OLS regression, where the cluster indicators and the control functions for maternal labor supply and family labor income are not included. The estimates are shown in Table 2.10.

Table 2.10: OLS Estimates without Control Functions

Period t \ Period t+1	Emotional Stability	Conscientiousness	Agreeableness	Extroversion	Openness	Math	Reading
Emotional Stability	0.682***	0.025*	0.024*	-0.048***	-0.019	0.017	-0.003
	(0.016)	(0.013)	(0.014)	(0.012)	(0.013)	(0.017)	(0.016)
Conscientiousness	0.122***	0.785***	0.015	0.013	-0.031**	0.056***	0.045***
	(0.015)	(0.014)	(0.011)	(0.012)	(0.012)	(0.016)	(0.017)
Agreeableness	0.092***	0.002	0.786***	0.072***	-0.006	-0.025	0.022
	(0.015)	(0.015)	(0.016)	(0.014)	(0.014)	(0.016)	(0.018)
Extroversion	-0.018	-0.020*	0.014	0.834***	0.006	0.006	-0.003
	(0.012)	(0.011)	(0.011)	(0.011)	(0.011)	(0.016)	(0.014)
Openness	-0.038**	-0.077***	-0.028**	-0.051***	0.776***	-0.023	0.061***
	(0.015)	(0.014)	(0.014)	(0.013)	(0.014)	(0.015)	(0.016)
Math Skill	0.031**	0.014	-0.012	0.002	-0.018	0.529***	-0.009
	(0.014)	(0.014)	(0.012)	(0.012)	(0.012)	(0.016)	(0.017)
Reading Skill	-0.026	0.054***	0.043***	0.009	0.047***	0.337***	0.485***
	(0.017)	(0.014)	(0.016)	(0.015)	(0.016)	(0.020)	(0.021)
Maternal Work Hours	-0.023*	-0.009	0.012	0.014	0.006	-0.015	0.013
	(0.013)	(0.011)	(0.011)	(0.010)	(0.011)	(0.014)	(0.015)
Family Income	0.006	0.010	0.012	0.025***	0.010	-0.009	0.047***
	(0.010)	(0.010)	(0.009)	(0.008)	(0.009)	(0.012)	(0.012)

Note. Period t is the late adolescent period and period t+1 is the early adulthood. The sample size is 3,799. Clustered standard errors at household level in parentheses * p < 0.10, ** p < 0.05, *** p < 0.01

Overall, the estimates of the OLS specification, where the control functions are not present, are not materially different from the results using the specification that addresses selection on unobservables. As a result, none of the main conclusions are overturned, suggesting modest selection on unobservables in the setting of this research using the HILDA data set. To examine the direction of biasness from potential selection on unobservables, the effect of potential selection on unobservables is nonuniform, as some estimates from the OLS without control functions are slightly smaller and some are greater in values. For example, the magnitude of the effect of family labor income on adolescents' agreeableness is greater when using the control function estimation, while the magnitude of the impact of family labor income on adolescents' reading skill is greater when using the OLS specification without the control functions. In general, the comparison of results from these two specifications suggests that potential selection on unobservables does not significantly bias the estimates in this sample.

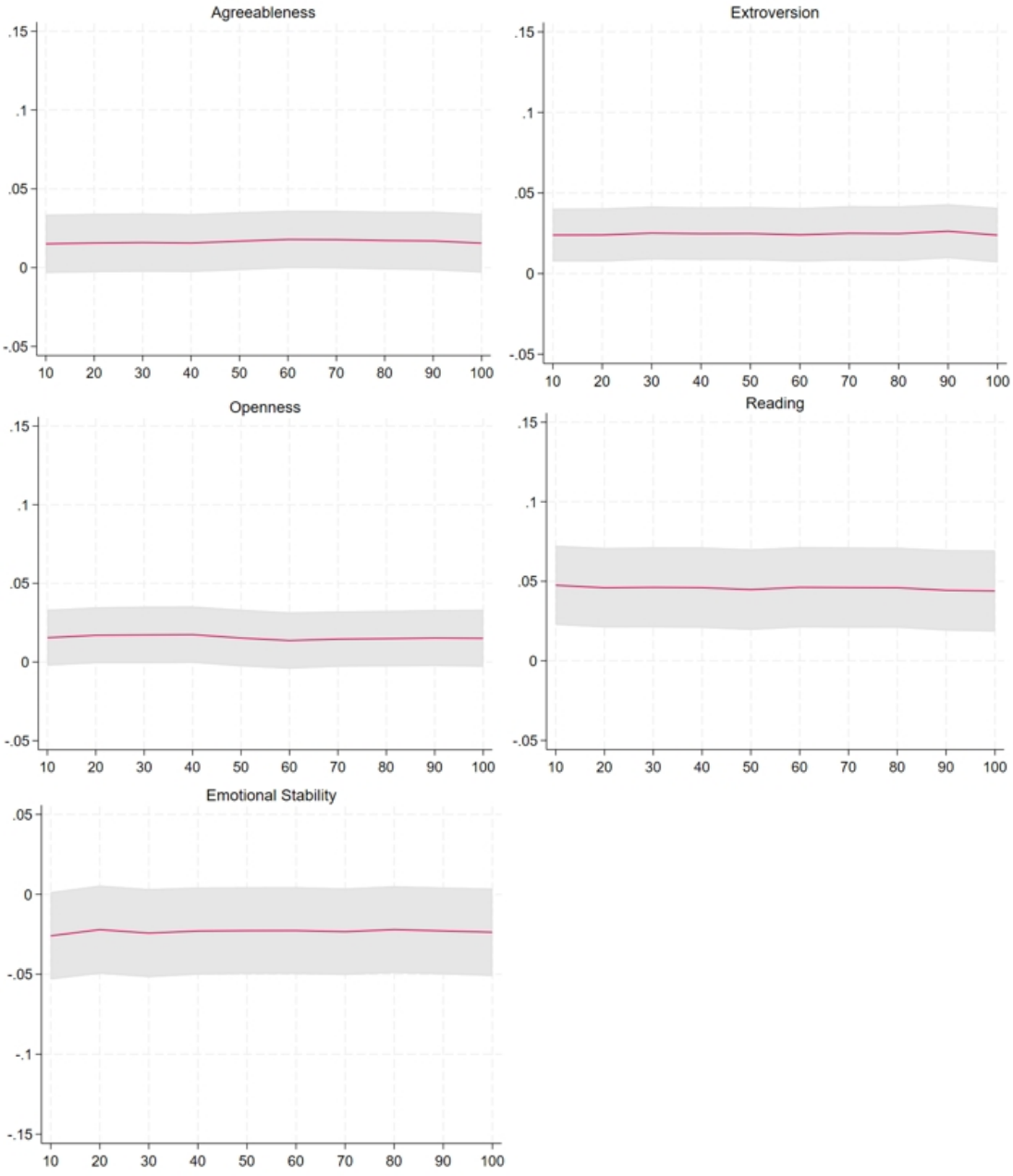
2.7.3 Applying Different Numbers of Clusters

When the expectation $\mathbb{E}\left[L^* \mid L = 0, X \in \hat{\mathcal{C}}_k\right]$ was estimated, I partitioned the support of X into $K = 50$ clusters for $\hat{\mathcal{C}}_K$. In this subsection, I re-estimate the regression of the control function approach in Equation (2.19), using varying numbers of clusters from 10 to 100, as a robustness check. The cluster indicators are updated, and the control functions based on

the corresponding number of cluster are also re-computed, when I re-estimate the control function regressions.

In Figure 2.3, I plot the point estimates of the effects of family income and maternal hours worked on adolescents' outcomes, with 95% confidence intervals, against varying numbers of clusters. The first four panels show the effects of family income on four different adolescents' outcomes. The bottom left panel plots the effect of maternal work hours on adolescents' emotional stability. I choose to show these five outcomes, because the family income or maternal hours worked display statisitcally significant impact on these five adolescent outcomes, as shown in Table 2.5. The estimates are generally stable across varying numbers of clusters for all five outcomes.

Figure 2.3: Robustness Check: Varying Cluster Numbers



Note. The first four panels are from running the main control function regression with varying cluster values, where the presented coefficients are on income. The bottom left panel is from the main control function regression with varying cluster values, where the coefficient is on maternal work hours.

2.8 Conclusion

I use a skill formation model to investigate the cross-effects and self-productivity of adolescents' Big Five personality traits, mathematical skill, and reading skill, as well as the impact of parental investments on these traits and skills. I also employ a bunching with control function approach to to further explore the effects of maternal labor supply and family labor income on the development adolescents' personality traits and cognitive skills, while also examining the the cross-effects and self-productivity of these traits and skills. I utilize the Household, Income and Labour Dynamics in Australia (HILDA) data set for both parts of the analyses.

Based on the skill formation model, I identify parameters of the law of motion for skills using an instrumental variable estimator. The results provide essential findings that are key to understanding the formation of personality traits and cognitive skills. For example, I find statistically significant evidence of cross-effect of mathematical skill on emotional stability and conscientiousness. These results align with findings from recent neuroscience research, which reveal that mathematical education has a positive impact on individuals' inhibitory control, often reflected in traits such as emotional stability and conscientiousness. (Zacharopoulos et al., 2021).

I addition to the skill formation model, I also apply a bunching with control function approach to further investigate the development of adolescents' personality traits and cognitive skills, when maternal labor supply is factored in and selection on unobservales are accounted for. The findings based on the control function approach are consistent with the results from the skill formation model. The cross-effect between adolescents' math skill and emotional stability is positive and statistically significant at the 5% level. The effect of family labor income in the previous period is positive and statistically significant on adolescents' agreeableness, extroversion, openness, and reading skill. The impact of maternal hours worked in the previous period on adolescents' emotional stability is negative and statistically significant at the 10% level. However, the positive income effect due to maternal labor supply fully offsets the direct negative impact of maternal labor supply. My findings from the skill formation model and from the control function approach are consistent with each other, suggesting that structural models and reduced-form analyses are able to meaningfully complement each other (Todd and Wolpin, 2023).

I also find that, compared to the effects of math and reading skills on earnings, the Big Five personality traits also exhibit economically large impacts on adult earnings. This is

consistent with the finding of Brunello and Schlotter (2011), suggesting that noncognitive skills are at least as important as cognitive skills for labor market success.

The malleability of personality traits and cognitive skills to investments and the environment in the emerging adulthood period offers valuable policy implications. This finding provides a new perspective on the earlier view that the effects of on-the-job training programs for young adults were minimal, which may have led some economists to assume that these traits and skills are not malleable in early adulthood. My findings suggest that investments and the environment during emerging adulthood play a significant role in the development of these traits and skills. Thus, this paper suggests a new direction for research aimed at designing more effective training programs, as the ineffectiveness of some training programs may stem from their design rather than the rigidity of traits and skills in early adulthood.

At last, I perform robustness checks for the control function approach, showing that the estimates remain generally stable. I also use a translog skill production function as a robustness check for the linear technology of skill production, and the estimates are largely consistent. Moreover, the estimation of the translog specification reveals intriguing results regarding the heterogeneous elasticities of skill production with respect to investments. Further analysis of these findings will be conducted in future work.

Chapter 3

The Impacts of E-Waste Dumping on Infant Health Outcomes in Ghana

3.1 Background and Introduction

Electronic waste (e-waste) is a serious environmental and public health issue that is often overlooked worldwide. There are numerous e-waste dumpsites in many developing countries, exposing workers to hazardous and carcinogenic substances such as mercury, lead, and cadmium (UN Environment Programme, 2019). According to a report by United Nations University 2020, a record of 53.6 million tons of e-waste was produced globally in 2019 (United Nations University, 2020). More than 18 million children and adolescents and as many as 12.9 million women are actively engaged in the informal e-waste sector (World Health Organization, 2021). This paper assesses the adverse effects of e-waste on infant health in Agbogbloshie, Ghana, which is one of the largest e-waste dumpsites on Earth in terms of the volume of wastes recycled per day and the number of people who work in the recycling industry there (Amoyaw-Osei et al., 2011). Agbogbloshie is located near the center of Accra, the capital and the largest city in Ghana by population size (Amoyaw-Osei et al., 2011).

Over the last 20 years, Agbogbloshie has become an informal recycling hub and a large scrapyard for e-waste (Owusu-Sekyere et al., 2022). For every 1,000 tons of e-waste shipped to the shore, it creates 30 jobs in landfills, 15 jobs in recycling, and 200 jobs in repairing (Sampson, 2015). Providing the opportunities for creating quick employment and profits, the resulting total e-waste flows at the Agbogbloshie range between 13,090 and 17,094 metric

tons per annum over the past 20 years (Owusu-Sekyere et al., 2022). However, short-run economic prosperity boosted by recycling parts from e-waste comes with lasting costs in health (Daum et al., 2017).

While a number of studies in public health and epidemiology have have found associations between pollution from e-waste and adverse health outcomes, little or no research has causally examined the impact, particularly in economics (Daum et al., 2017; Rai et al., 2019; Thanomsangad et al., 2020). Therefore, in this paper, I investigate the causal relationship between infants' health outcomes and exposure to e-waste pollution. This paper contributes to the literature in the following three ways.

First, I causally examine the impact on two sets of infant health outcomes directly related to exposure to e-waste pollution. The first set of infant health outcomes are physical symptoms: diarrhea and the respiratory illness (coughing). The second set of health outcomes consists of infant mortality, birth weight, height-for-age z-score, and weight-for-age z-score. Second, I contribute to the literature by rigorously examining two potential mechanisms (sources of drinking water and tetanus vaccine take-up) that link the exposure to e-waste and adverse health outcomes. To my knowledge, this paper is the first study in the economics literature, and one of the few papers in the field of epidemiology and public health that investigates the impact of women receiving tetanus vaccines during or before pregnancy on child health in the vicinity of an e-waste dumpsite. Lastly, I analyze the association between household characteristics and the consumption of safer alternatives of drinking water as well as mothers' take-up of the tetanus vaccine. This paper is also one of the first two papers in the economics literature that studies the impact of e-waste on children's health outcomes.[1]

Examining the impact of e-waste on children's health outcomes and understanding the mechanisms of transmission is critical, because research suggests that pollutants from e-waste are linked to both acute health symptoms and potentially fatal health risks. Studies have found that e-waste dumping sites are point sources of heavy metal pollutants, such as lead, arsenic, mercury, and chromium. These heavy metals are released into water and

[1]To give a sense of the timeline, Lovo and Rawlings (2021) posted their discussion paper on a similar topic in July 2021. We have been working on this topic completely independently. My research project of e-waste and child health was initiated in Spring 2020, and the DHS approved my data application on October 12, 2020 for my registered e-waste project. Lovo, Rawlings, authors of the other paper, and I got in touch in October 2021, and we agreed on not to combine our papers partly due to numerous points of departure in our works - most importantly in key identifications and major contributions. According to our email exchanges, Lovo and Rawlings started their project of pollution and health in 2017. We were unaware of each other's work before Lovo and Rawlings (2021) went public first.

soil, and they accumulate in surrounding water systems, food crops, and animals. Then these heavy metals disseminate into human bodies through water and food consumption (Daum et al., 2017; Rai et al., 2019; Thanomsangad et al., 2020). Ingesting heavy metals, such as cadmium and mercury, in higher amounts can lead to stomach irritation and result in vomiting and diarrhea (Jaishankar et al., 2014). Rehman et al. (2018) also discovered that heavy metal-contaminated water results in child morbidity and mortality.

In addition to heavy metals, e-waste dumping sites also produce airborne organic chemicals, most notably polybrominated diphenyl ethers (PBDEs) and hexabromocyclododecanes (HBCDs) (Daum et al., 2017). The organic chemicals are released into the air, soil, and water, entering the bodies of e-waste recyclers and residents of nearby communities. Airborne pollutants disseminate into the air, as many recyclers engage in open burning of e-waste byproducts to extract valuable metals, for example copper, from e-waste for resale. Recycling workers and residents in neighboring communities reported having respiratory illnesses (Daum et al., 2017). Therefore, in this paper, I examine the effects across multiple dimensions of adverse health outcomes in infants: acute health symptoms (diarrhea and coughing), potentially fatal health risks (infant mortality and low birth weight), and related health measures (height-for-age and weight-for-age z-scores). These risks in early childhood are particularly concerning, as Currie and Almond (2011) point out the importance of child development before age five, and some damages inflicted on children's health are irreversible in later life.

I also identify two potential mechanisms of how e-waste pollution leads to adverse health outcomes: the mother's tetanus vaccine take-up during or before pregnancy and the sources of drinking water. Tooher et al. (2005) have pointed out that, at present, few empirical studies have documented the interactive relationship between infant health outcomes and the increased risk of infectious disease, particularly tetanus infection, even though epidemiologists believe that it is a much more worrisome issue in communities near e-waste dumping sites. To the best of my knowledge, this paper is the first in the economics literature, and among the few in epidemiology and public health, to examine the impact of maternal tetanus vaccination during or before pregnancy on child health outcomes near e-waste dumpsites.

Environments with poor sanitation and frequent injuries, such as an e-waste recycling site, can pose a higher risk for tetanus infections (Parvez et al., 2021). Thus, maternal tetanus vaccination is a key measure to mitigate tetanus infection in children, in addition to children's uptake of the Tetanus, Diphtheria, and Pertussis (Tdap) vaccine. After getting

the vaccine, mothers' bodies starts to make antibodies against the bacteria. Some of these antibodies can cross the placenta and be passed to the fetus, and help the child obtain antibodies until the child can receive their own vaccines (Campbell et al., 2018). The source of drinking water is also a critical mechanism to explore, as heavy metals from e-waste can contaminate water supplies and enter the human body through drinking and food consumption (Daum et al., 2017; Rai et al., 2019; Thanomsangad et al., 2020).

I use the Demographic and Health Survey (DHS) data for Accra, Ghana, from 1993 to 2017. The DHS provides the latitude and longitude geocode for each survey cluster, which allows me to pinpoint the geographic location of a cluster in the Geographic Information System (GIS) so that I can determine its distance to Agbogbloshie. I employ a difference-in-differences (DID) framework to investigate the effects on infants' health outcomes of being born within the exposure area of Agbogbloshie, Ghana. I compare them to the health outcomes of infants born in the same city, Accra, but far away from this e-waste dumpsite.

Here are the main findings from this study. I discover that living in the exposure area (5 kilometers) of Agbogbloshie significantly increases the probability of a child suffering from diarrhea, by 52 percentage points.[2] When the household does not have access to safe sources of drinking water, children in the exposed area are 66 percentage points more likely to have diarrhea than children who do not have access to secure drinking water in the unexposed area. Having safer sources of drinking water significantly mitigates the risk of an infant having diarrhea by 47 percentage points (about 71% of the total adverse effect). Living at a close distance to the e-waste dumping site also increases the likelihood that an infant has a respiratory illness (cough) by 48 percentage points.

Infants are 0.95 kilograms lighter when mothers have not received a tetanus vaccine in the exposed area compared to infants' birth weight of whose mothers have not received a tetanus vaccine in the unexposed area. However, this adverse effect can be significantly mitigated by mothers' uptake of at least one dose of tetanus vaccine before or during pregnancy, which mitigates lower birth weight by 0.68 kilograms (about 72% of the total adverse effect).

I also investigate two channels that lead to the adverse health outcomes due to the e-waste dumpsite. First, detrimental health outcomes are caused by the consumption of contaminated water for families that live near an e-waste dumpsite site, as the surface water and pipelines near the dumpsite are more susceptible to heavy metal pollution (Monney et al., 2013). The second channel is the exposure to higher risks of tetanus infection when a

[2]The average probability of children having diarrhea in the control group is about 27%.

household lives within the exposure area of the e-waste dumpsite. I find that when a mother is vaccinated against tetanus, her newborn is less likely to experience premature birth. This is because antibodies for tetanus bacteria can cross the placenta and be passed to the fetus, providing the child early protection until they are old enough to receive vaccinations (Campbell et al., 2018).

After studying the association between household characteristics and the consumption of safer alternatives of drinking water, I find evidence that the less wealthy households are aware of the deleterious effects of drinking unsafe sources of water in the treated area. The less wealthy families are more willing to spend money on the more expensive alternatives of water if they live in the exposed area.

The structure of the rest of this paper is as follows. Section 3.2 discusses the data sets, determination of the year of intervention, and the selection method of the treatment and control groups. Section 3.3 introduces the empirical frameworks. Section 3.4 presents the main results. Section 3.5 presents the robustness checks. Section 3.6 concludes.

3.2 Data

This paper employs individual-level survey data for Ghana from the Demographic and Health Surveys (DHS) and the Multiple Indicator Cluster Surveys (MICS) from the United Nations Children's Fund (UNICEF). The DHS and the MICS data are structured as repeated cross-sectional data sets but include retrospective birth panels of mothers. Households are surveyed only once in the DHS and the MICS data, so the households to be surveyed in the next wave are different from those in the previous waves. However, a mother's retrospective birth panel embedded in each year's data contains information about a mother's entire birth history. I utilize the retrospective birth panel for the study.

The DHS data provide geographic information of clusters for the following waves: 1993, 1998, 2003, 2008, 2014, and 2017. The cluster-level GPS information enables connecting individuals' survey clusters to a pinpoint of longitudinal and latitudinal geocode. There are in total of 3,125 infants from the DHS data and 566 infants from the MICS data, and the combined sample size is 3,691. The control group contains 163 clusters, and the treatment group has 132 clusters in total. Among the 3,691 children in the sample, 1,636 were born in the treated group and 2,055 were born in the control group.

The MICS data are highly comparable to the DHS data. They have similar data structures and survey questions. Instead of providing anonymized geocode information,

the MICS offers the locality (neighborhood of Accra) information of the respondents in its 2010-2011 survey. This allows for a more precise calculation of the distance to the dumping site. The MICS provides locality information only in the 2010-2011 survey of Ghana, but not in other waves; therefore, only a small sample size from the MICS can be used.

Information on household characteristics, such as the wealth level and mother's education, is available in both the DHS and the MICS data sets. The outcomes of interest consist of children's physical symptoms and birth outcomes. Children's physical symptoms include diarrhea and the respiratory illness, which are directly impacted by living in the vicinity of the dumping site (Jaishankar et al., 2014; Daum et al., 2017). Birth outcomes consist of infant mortality, infant birth weight, child's weight-for-age z-score, and height-for-age z-score. Diarrhea and the respiratory illness are recorded in surveys as whether the child had the symptom in the past two weeks. They are binary variables coded as 1 if yes and 0 otherwise. Note that infant birth weight is measured in kilograms, and it is topcoded at the 99th percentile to avoid extreme weights. Infant mortality is a binary outcome coded as 1 if a child dies before reaching the age of 12 months; it is coded as 0 otherwise. The weight-for-age and height-for-age (age in months) z-scores are calculated based on the 2007 WHO reference (Onis et al., 2007).

3.2.1 Determining the Year of Intervention

A key challenge that this paper must address is to identify an exact year of intervention. So far, I have not found any organizational or administrative document that convincingly pinpoints a precise year when this e-waste dumpsite started to become a hazard to the surrounding residents. Here is a brief review of the history of Agbogbloshie based on Amoyaw-Osei et al. (2011).

In 1994, the National Youth Council of Ghana, the custodians of the land that later became Agbogbloshie, leased the land to the Scrap Dealers' Association of Ghana. It then became the hub of an informal recycling industry in Ghana, but the place was not yet a hazardous e-waste dumping site. Second-hand electrical and electronic devices such as mobile phones, televisions, computers, and radios were frequently shipped from Europe and North America, and this flow of imports increased by nearly a factor of three between 2003 and 2008 (Amoyaw-Osei et al., 2011). Through the accumulation of used electronic devices shipped from foreign countries, this piece of land gradually became the infamous dumpsite Agbogbloshie.

To pinpoint an exact year of intervention, I refer to the research literature in marine biology and environmental studies that first documented the adverse effect of Agbogbloshie on the surrounding environment and residential community. Alarming marine consequences have been first observed by environmental scientists and marine biologists as early as 2002. Consequences included harmful effects on aquatic plant and animal species, for example, smaller, sicker, and sparser fish stocks (Owusu Boadi and Kuitunen, 2002; Bandowe et al., 2014; Daum et al., 2017). Note that Agbogbloshie is close to the center of the coastal city Accra, Ghana. This channel is consistent with the research that shows that e-waste dumping sites release heavy metals into water and soil and cause damage to surrounding water systems and the environment (Jaishankar et al., 2014; Daum et al., 2017; Thanomsangad et al., 2020). I utilize the findings from these studies to pinpoint that the year of treatment is the year 2002.

3.2.2 Selecting the Treatment and Control Groups

To ensure that respondents' confidentiality is maintained, the DHS randomly displace the GPS latitude and longitude positions for all survey clusters or communes. For clusters in the urban area, the DHS first selects a random direction (angle) between 0 and 360 degrees. Then it selects a random distance from a minimum of 0 and a maximum of 2 kilometers for the urban cluster. Finally, combining the results of the previous two steps, the DHS assigns the new GPS coordinate based on the true GPS information to the cluster. Therefore, the clusters used in this paper, which are all drawn from the urban area, contain a minimum of 0 and a maximum of 2 kilometers of error in the precision of geocode. Households in the same survey cluster or commune are assigned with the same geocode, and the noise in geocode is implemented at the cluster level. In my sample, each cluster has 10.42 households on average. To account for the noise implemented in the geographic information, I assign the survey clusters to the treated and control groups as follows.

First, the sample only includes survey clusters in the urban areas of the Greater Accra Region of Ghana. There are two reasons. Firstly, Agbogbloshie is located close to the urban center of Accra. Thus, people in the treated clusters (individuals living in the vicinity of the dumpsite) are all urban residents. It is unreasonable to include rural households in the control group, because of potential structural differences in the baseline characteristics between urban and rural households, hence different outcomes in children's health outcomes. Secondly, the DHS imposes a 2 km noise to the geocode of the urban clusters, while it

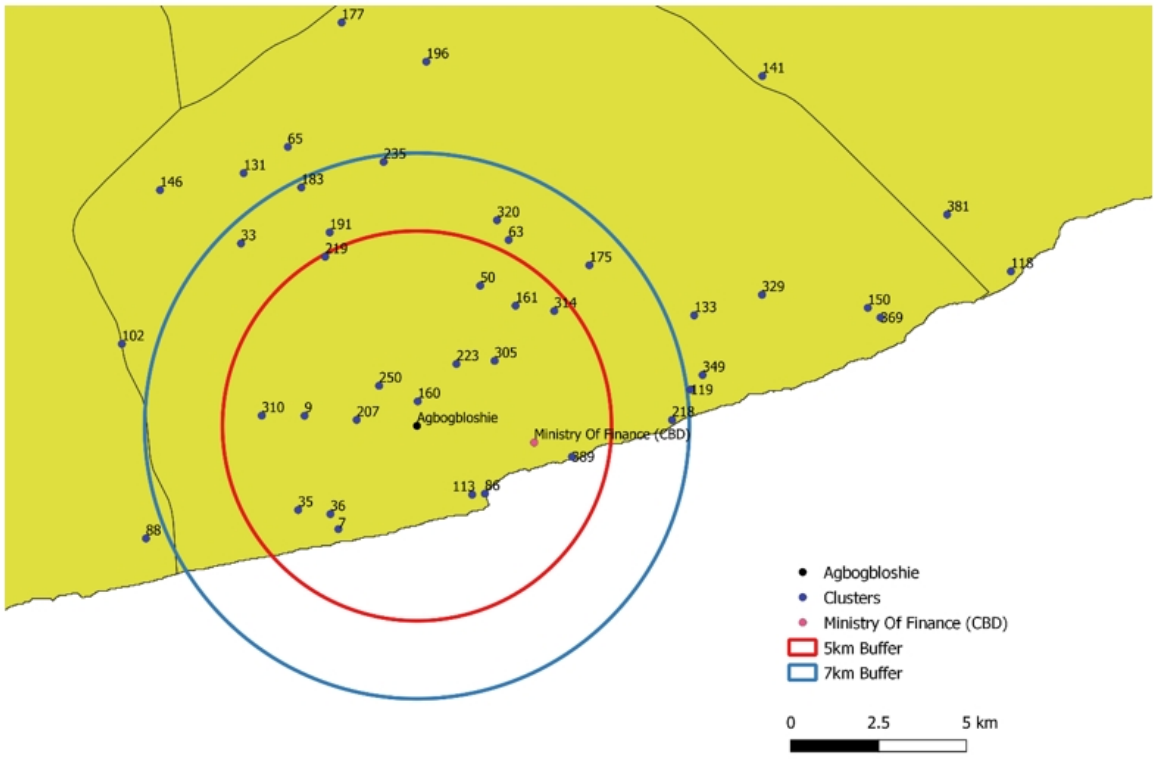
implements a 5 km noise to the geocode from the rural clusters. Using survey clusters exclusively in the urban area helps reduce the magnitude of noise in the measurement.

Second, I need to determine the distance from the e-waste dumpsite, within which the residents are considered to live in the exposure area. To do this, I have extensively reviewed the literature in the fields of public health and epidemiology. Some researchers, for example Dolk et al. (1998), used buffer zones of 3 km and 4 km. Some epidemiologists considered a distance of 2–4 km between the exposure areas and hazardous-waste landfill sites to be informative (Jarup et al., 2002). Based on a waste management report by the World Health Organization (2007), similar choices have also been made by other research papers. They adopted 2–3 km circles, and they do not use a 1 km circle because of large data fluctuations (World Health Organization, 2007). The same report by the World Health Organization also points out that a distance of 5 km between the exposure areas and landfill site was considered by some studies in order to acquire a more "powerful" study because it allows for a larger potentially exposed population (World Health Organization, 2007). There is sufficient evidence suggesting that epidemiologists and public health experts have a consensus on the circles of exposure of being 2-5 km from the center of the e-waste dumpsite (World Health Organization, 2007).

Based on the consensus in the rich volume of research in public health and epidemiology, I choose to use a radius of exposure of 5 km from the center of the Agbogbloshie as the primary definition of the treated area. For robustness check purposes, I also show the results when using 4 km, 3 km, and 2 km to be the buffer zones for the treated group.[3]

[3]Butts (2021) suggests that to identify unbiased average treatment effect on the treated using DID method with geocoded microdata requires two assumptions: (1) parallel trends; and (2) knowing how far treatment effects extend. To relax the second assumption, Butts (2021) proposes a nonparametric estimation that uses infinitely partitioned distance intervals. Based on the TWFE event study result, my DID approach satisfies the parallel trends assumption. Given the prior research on e-waste, I arguably have knowledge of the exposure distance from an e-waste dumpsite. For further robustness checks, I also utilize and report the result of continuous treatment with $1/distance$ and $1/distance^2$ to the dumping site, which is consistent with the nonparametric approach that Butts (2021) proposes.

Figure 3.1: GIS Layers of the Geography of Accra, Ghana and DHS Clusters



Note. This graph generated using QGIS consists of layers of DHS Clusters in 2003 (blue dots), Agbogbloshie (the black dot), Buffer Zones from The Center of Agbogbloshie, the Ministry of Finance representing the location of the CBD, and the Shapefile of the Greater Accra Region of Ghana.

To show an example of how treatment and control clusters are selected, the graph presented in Figure 3.1, generated by the GIS, illustrates the DHS clusters in 2003, the e-waste dumpsite Agbogbloshie and the buffer zones from the e-waste dumpsite layered on the map of the Greater Accra region of Ghana.

Figure 3.1, which presents two buffer zones and the survey clusters in 2003, serves as an example of how the treated and control clusters are defined. The survey clusters that fall within the 5 km buffer zone are assigned to the treated group. These clusters are all in the urban area because Agbogbloshie is located fairly close to the city center of Accra, Ghana. I use the Ministry of Finance of Ghana to represent an approximate location of the central business district of Accra. In contrast, the clusters that lie outside of the 7 km circle, while remaining in the urban area of the Greater Accra Region of Ghana, are assigned to the control group. To account for the 2 km noise in the cluster-level GPS information

implemented by the DHS, the clusters that fall between the 5 km and the 7 km buffer zones are excluded from the analysis. I exclude the clusters in between because they can either fall into the 5 km buffer zone or fall outside of the 7 km buffer zone due to the noise in the geocode.

The clusters that are on the periphery of the 5 km buffer zone, in the most extreme case, may be 7 km away from the center of Agbogbloshie. However, it is very unlikely to happen because first, the noise of a cluster's geocode must extend exactly in the direction away from the dumping site; second, the noise has to be exactly 2 km from a continuous interval from 0 to 2 km. Statistically, it is extremely unlikely for both of the conditions to be satisfied at the same time. The same reason also holds for selecting the control groups. I repeat the same selection process for all the other waves other than 2003.

To test the balance of key covariates that may impact health outcomes after infants are born, I generate a balance table (Table 3.1) to compare the key covariates between the treatment and the control groups. All the clusters are in urban areas. Wealth quantile, reported by the DHS and MICS, is a categorical variable ranging from 1 to 5. The wealth quantile equal to 1 indicates that the household belongs to the lowest wealth quantile, and 5 indicates that the household belongs to the highest wealth quantile. Mother's years of education is defined as the total number of years of education that she had completed in the year of the child's birth. Mother's weight refers to her body weight, measured in kilograms, in the year of the child's birth. Mother's age at birth measures a mother's age in the year of the child's birth. The variable "Female Child" captures the share of female children within the treatment or control group. The variable "Twins" indicates the proportion of twin births among all births in the treatment or control group. As shown in Table 3.1, there are no significant differences in key covariates between the treated and the control group.

Table 3.1: Descriptive Statistics and Balance Table

	Treatment	Control	Difference
	(1)	(2)	(3)
Urban	1	1	0
			(0)
Wealth Quantile	4.25	4.30	-0.05
			(0.12)
Mother's Years of Education	7.78	8.06	-0.28
			(0.18)
Mother's weight (kg)	72.24	71.72	0.52
			(0.46)
Mother's Age at Birth	26.52	26.30	0.22
			(0.23)
Female Child	0.51	0.48	0.03
			(0.02)
Twins	0.05	0.04	0.01
			(0.01)
Observations	1,636	2,055	

Note: This is the covariate balance table that compares key characteristics between the treated and the control groups from 1997 to 2012. In parentheses, they present cluster-adjusted standard errors. Significance: * 0.10; ** 0.05; *** 0.01.

3.3 Empirical Framework

In order to assess the impact on infant health of being born within the circle of exposure of the e-waste dumpsite Agbogbloshie, I estimate the following two-way fixed effect difference-in-differences (DID) model.

$$Y_{ijt} = \beta_0 + \alpha_j + \lambda_t + \delta^{DD} * D_{jt} + \gamma' * X_{it} + \epsilon_{ijt} \tag{3.1}$$

where Y_{it} is the outcome of interest for individual i in cluster j at birth year t. The outcomes of interest are diarrhea, respiratory illness, infant mortality, birth weight, weight-for-age z-score, and height-for-age z-score. The estimate of interest δ^{DD} is the coefficient of the

DID estimate, α_j is the cluster fixed effect which controls for all time-invariant cluster characteristics, λ_t is the year fixed effect which controls for all time-varying factors common across clusters, the vector X_{it} represents the set of household level control variables, and ϵ_{ijt} is the error term.

The vector of control variables includes the mother's education, mother's weight, sex of the child, whether the child is born as one of the twins, and the mother's age at birth.[4] The analysis of the DID model spans from 5 years before the year of treatment (2002) to 10 years since the year of treatment (including the year 2002). Note that the mother fixed effect is collinear with the binary treatment status because the treatment status depends on the geographic information of the household, but it does not collinear with the interaction term of the treatment and child's year of birth, which is the estimate of interest D_{jt}. The key is that knowing the year of interview does not necessarily tell a child's year of birth. As I discuss previously in the Data section, despite being surveyed only once, each wave of the survey contains information about a mother's entire birth history.

I also employ a second identification strategy which utilizes the distance from the clusters to the dumping site and implements continuous doses of treatment. The treatment is $1/distance^2$ based on the inverse-square law in physics. For robustness, I also show the result of using $1/distance$ to the dumping site as the treatment. The continuous dose of treatment becomes larger if a cluster is closer to the dumping site, and it becomes smaller as a cluster locates further from the dumping site. No cluster in my sample has zero distance to the dumping site. The identification strategy using a continuous dose of treatment is as follows.

$$Y_{ijt} = \beta_0 + \alpha_j + \lambda_t + \delta^{DD} * After_t * \frac{1}{Distance_j^2} + \gamma^{'} * X_{it} + \epsilon_{ijt} \tag{3.2}$$

where the estimate of interest δ^{DD} is the DID estimate, $After_t$ is 1 if the year is or after 2002 and vice versa.

To test the pre-trends and to gain more insights into the dynamics of how living in the vicinity of the dumpsite affects health outcomes over time, I extend the DID framework given in Equation (3.1) and perform a two-way fixed effect event study analysis, estimating

[4]Note that household wealth quantile is not included as a covariate because it is a bad control that can also be a part of the outcome from working in or living in the vicinity of the dumping site, which is close to the city center of the Accra.

the following specification:

$$Y_{ijt} = \beta_0 + \alpha_j + \lambda_t + \sum_{\substack{t=-5 \\ t\neq 0}}^{9} \delta_t * D_{jt} + \gamma^{'} * X_{it} + \epsilon_{ijt} \tag{3.3}$$

where the series of δ_t are the estimates of interest, D_{jt} is 1 if the child was born in a treated cluster in the corresponding t, and it is 0 if otherwise. In this event study specification, children's health outcomes for 5 years before 2002 and 10 years since 2002 are examined, given that 2002 is the reference year.

Lastly, I employ a difference-in-difference-in-differences (DDD) estimator to investigate the heterogeneous effects based on family characteristics and to examine potential mechanisms that cause the adverse effects of e-waste dumping on children's health outcomes as follows.

$$Y_{ijt} = \beta_0 + \alpha_j + \lambda_t + \delta^{DD} * D_{jt} + \beta_1 * M_{ijt} + \delta^{DDD} * D_{jt} * M_{ijt} + \gamma^{'} * X_{it} + \epsilon_{ijt} \tag{3.4}$$

where δ^{DD} is the DID estimate, M_{ijt} stands for the family characteristic or the mechanism that links the e-waste pollution to a child's health outcomes, and δ^{DDD} is the DDD estimate. I explore the following family characteristics for heterogeneous effects: family's wealth level, mother's education, and mother's occupation. For mechanisms, I explore the sources of drinking water and whether a mother received at least one dose of tetanus vaccine before or during pregnancy. These two potential mechanisms have been reported by public health experts and epidemiologists for their mechanisms of addressing the potentially detrimental effects of e-waste pollution (Ihedioha et al., 2017; Tooher et al., 2005).

A family's wealth level is a binary variable measuring whether a household is above the threshold of average family wealth; it is 1 if a household's wealth level is above average and vice versa. The wealth index, calculated by the DHS, is a composite measure of a household's ownership of selected assets, which is an effective measure to identify the household wealth level in a third-world country. Mother's education is a binary variable coded as 1 if a mother's highest completed degree is secondary education or above; it is entered as 0 if a mother's education is below secondary. A mother's occupation is also a binary variable measuring whether a mother has a job that is less likely to be directly exposed to e-waste pollution. Occupations such as professionals, clerical workers, and managers are considered less likely to be directly exposed to the pollution, while jobs such as

manual workers, retailers, and agricultural workers are considered more likely to be directly exposed to the pollution. The job binary variable is coded as 1 if a mother is employed by a position that is less likely to be directly exposed to the pollution and vice versa.

The source of drinking water is a binary variable that equals 1 if a household regularly consumes safe alternatives of drinking water; it is 0 if a household regularly consumes unsafe sources of drinking. The second mechanism to explore is whether a mother received at least one dose of tetanus vaccine before or during pregnancy. It is coded as 1 if a mother received a least one tetanus vaccine and it is 0 otherwise.

3.4 Main Results

I study the impact of living in the exposure area of an e-waste dumping site on the following health outcomes: infant mortality, birth weight (kilogram), whether a child had diarrhea in the past two weeks, whether a child had a respiratory illness in the past two weeks, a child's weight-for-age (age in months) z-score, and height-for-age (age in months) z-score. Tables 3.2 to 3.7 present the average impact on the health outcomes for those who live in the vicinity of the dumping site, showing the results using different buffer zones (5 km, 4 km, 3 km, and 2 km) and two continuous treatment estimates ($1/distance$ and $1/distance^2$ to the dumping site), respectively. Tables 3.8 to 3.13 present the heterogeneous effects of a household's wealth level, mother's education, mother's occupation, and a related potential mechanism on the health outcomes, respectively. Section 3.4.1 presents the main results. Section 3.4.2 discusses the channel that links sources of drinking water to a child's health outcomes. Section 3.4.3 discusses the mechanism that connects women's tetanus vaccine take-up and infant's health outcomes

3.4.1 The Main Results

As shown in Table 3.2, living in the exposure area has a statistically significant (at the 1% level) impact on the probability of an infant (0-12 months old) suffering from diarrhea across all specifications (changing the buffer zones and using the continuous treatment strategy). A child who lives in the 5 km buffer zone is 52 percentage points more likely to have diarrhea in the past two weeks than a child living in the control area.[5] The adverse effect increases as households live closer to the dumping site. A child who resides in the 2

[5]The average probability of children having diarrhea in the control group is about 27%.

km buffer zone becomes 61 percentage points more likely to suffer from diarrhea than a child in the control area. The DID estimates are all significant at the 5% or 1% level and the percentage points are all larger than 40, revealing a concerningly higher risk of diarrhea living in the exposure area. The continuous treatment specification also demonstrates a worsening effect as the household lives at a closer distance to the dumping site.

I find varying mitigating effects of having a higher household's wealth level, higher mother's education, more decent mother's occupation, and safer sources of drinking water on infants' diarrhea (Table 3.8).[6] The difference-in-difference-in-differences (DDD) estimates on the mother's occupation and the water source are significantly negative, meaning that when a mother is hired by a job that is less likely to be directly exposed to the pollution, her infant is 14 percentage points less likely to have diarrhea compared to an infant whose mother, holding all else constant, has an occupation that is more likely to be exposed to the e-waste. The magnitude of the mitigating effect of consuming safer alternatives of drinking water is the largest among all heterogeneous effects and it is statistically significant. Children who do not have access to secure drinking water in the exposed area are, on average, 66 percentage points more likely (at the 1% significance level) to have diarrhea than children who do not have access to secure drinking water in the unexposed area. Having safe sources of drinking water significantly (at the 5% and 10% level) mitigates the risk of an infant suffering from diarrhea by 47 percentage points (about 71% of the adverse effect). The link that connects sources of drinking water to child's health outcomes will be further discussed in Section 3.4.2.

[6] A more decent mother's occupation indicates whether the mother's job is less likely to be directly exposed to e-waste pollution. Occupations such as professionals, clerical workers, and managers are considered less likely to be directly exposed to the pollution, while jobs such as agricultural workers, manual workers, and retailers are considered more likely to be directly exposed to the pollution.

Table 3.2: Average Effect on Diarrhea

	(1)	(2)	(3)	(4)	(5)	(6)
DID Estimate	0.41***	0.41***	0.42***	0.42***	0.72***	1.20***
	(0.06)	(0.06)	(0.08)	(0.08)	(0.10)	(0.16)
Cluster Fixed Effect	Yes	Yes	Yes	Yes	Yes	Yes
Controls	No	No	No	No	No	No
Sample Size	446	384	204	155	476	476
DID Estimate	0.52***	0.56***	0.62***	0.61***	0.91***	1.51***
	(0.18)	(0.19)	(0.21)	(0.22)	(0.31)	(0.48)
Cluster Fixed Effect	Yes	Yes	Yes	Yes	Yes	Yes
Controls	Yes	Yes	Yes	Yes	Yes	Yes
Sample Size	446	384	204	155	476	476

Note: Significance: * 0.10; ** 0.05; *** 0.01. The columns present the DID estimator using different buffer zones to identify the treated group. Specification (1) reports the results of using 5km buffer zone, (2) using 4km buffer zone, (3) using 3km buffer zone, (4) using 2km buffer zone, (5) using continuous treatment with 1/distance to the dumping site, and (6) using continuous treatment with 1/distance squared to the dumping site. Standard errors are cluster robust at the DHS cluster level. The vector of control variables includes the mother's education, mother's weight, sex of the child, whether the child is born as one of the twins, and the mother's age at birth.

Table 3.3: Average Effect on Respiratory Illness (Cough)

	(1)	(2)	(3)	(4)	(5)	(6)
DID Estimate	0.39*	0.40*	0.41	0.41	0.50	0.83
	(0.22)	(0.23)	(0.27)	(0.28)	(0.37)	(0.65)
Cluster Fixed Effect	Yes	Yes	Yes	Yes	Yes	Yes
Controls	No	No	No	No	No	No
Sample Size	446	384	204	155	476	476
DID Estimate	0.48*	0.50*	0.50*	0.58*	0.71	1.17
	(0.26)	(0.28)	(0.31)	(0.32)	(0.46)	(0.77)
Cluster Fixed Effect	Yes	Yes	Yes	Yes	Yes	Yes
Controls	Yes	Yes	Yes	Yes	Yes	Yes
Sample Size	446	384	204	155	476	476

Note: Significance: * 0.10; ** 0.05; *** 0.01. The columns present the DID estimator using different buffer zones to identify the treated group. Specification (1) reports the results of using 5km buffer zone, (2) using 4km buffer zone, (3) using 3km buffer zone, (4) using 2km buffer zone, (5) using continuous treatment with 1/distance to the dumping site, and (6) using continuous treatment with 1/distance squared to the dumping site. Standard errors are cluster robust at the DHS cluster level. The vector of control variables includes the mother's education, mother's weight, sex of the child, whether the child is born as one of the twins, and the mother's age at birth.

Table 3.4: Average Effect on Infant Mortality

	(1)	(2)	(3)	(4)	(5)	(6)
DID Estimate	0.01	0.02	0.03	-0.04	-0.01	-0.05
	(0.02)	(0.02)	(0.03)	(0.04)	(0.05)	(0.04)
Cluster Fixed Effect	Yes	Yes	Yes	Yes	Yes	Yes
Controls	No	No	No	No	No	No
Sample Size	3,341	3,018	2,728	2,439	3,862	3,862
DID Estimate	0.01	0.02	0.02	-0.04	-0.01	-0.06
	(0.02)	(0.02)	(0.03)	(0.04)	(0.05)	(0.04)
Cluster Fixed Effect	Yes	Yes	Yes	Yes	Yes	Yes
Controls	Yes	Yes	Yes	Yes	Yes	Yes
Sample Size	3,341	3,018	2,728	2,439	3,862	3,862

Note: Significance: * 0.10; ** 0.05; *** 0.01. The columns present the DID estimator using different buffer zones to identify the treated group. Specification (1) reports the results of using 5km buffer zone, (2) using 4km buffer zone, (3) using 3km buffer zone, (4) using 2km buffer zone, (5) using continuous treatment with 1/distance to the dumping site, and (6) using continuous treatment with 1/distance squared to the dumping site. Standard errors are cluster robust at the DHS cluster level. The vector of control variables includes the mother's education, mother's weight, sex of the child, whether the child is born as one of the twins, and the mother's age at birth.

Living in the vicinity of Agbogbloshie increases the likelihood that an infant has a respiratory illness (cough) by about 50 percentage points (Table 3.3). It is statistically significant at the 10% level in all four binary treatment specifications. As presented in Table 3.9, the harmful effect that creates a respiratory illness is significantly (at the 5% level) alleviated when a mother has received at least secondary education. The adverse effect is also significantly (at the 1% level) mitigated by 50 percentage points (about 68% of the adverse effect) when a mother has an occupation that is less likely to be exposed to the pollution.

Infant mortality, on average, is not statistically higher in households living in the vicinity of the dumping site than in households that are far away (Table 3.4). Despite statistically insignificant due to a smaller sample size and a larger standard error, it is notable that mothers who did not receive a tetanus vaccine faced a much larger risk of infant mortality (25 percentage points higher) than those who received at least one dose of tetanus vaccine before or during pregnancy (Table 3.10). I also find a significant (at the 5% level) heterogeneous effect on infant mortality based on family wealth (Table 3.10). However, other household characteristics do not appear to exhibit a significant effect. Thus, the estimate of the heterogeneous effect on infant mortality based on family wealth can be

a spurious significance.

Table 3.5: Average Effect on Birth Weight

	(1)	(2)	(3)	(4)	(5)	(6)
DID Estimate	-0.19	-0.22	-0.12	-0.02	-0.18	-0.12
	(0.21)	(0.23)	(0.34)	(0.37)	(0.24)	(0.16)
Cluster Fixed Effect	Yes	Yes	Yes	Yes	Yes	Yes
Controls	No	No	No	No	No	No
Sample Size	732	642	461	390	816	816
DID Estimate	-0.15	-0.17	-0.07	-0.01	-0.10	-0.07
	(0.22)	(0.24)	(0.33)	(0.35)	(0.25)	(0.17)
Cluster Fixed Effect	Yes	Yes	Yes	Yes	Yes	Yes
Controls	Yes	Yes	Yes	Yes	Yes	Yes
Sample Size	732	642	461	390	816	816

Note: Significance: ∗ 0.10; ∗∗ 0.05; ∗∗∗ 0.01. The columns present the DID estimator using different buffer zones to identify the treated group. Specification (1) reports the results of using 5km buffer zone, (2) using 4km buffer zone, (3) using 3km buffer zone, (4) using 2km buffer zone, (5) using continuous treatment with 1/distance to the dumping site, and (6) using continuous treatment with 1/distance squared to the dumping site. Standard errors are cluster robust at the DHS cluster level. The vector of control variables includes the mother's education, mother's weight, sex of the child, whether the child is born as one of the twins, and the mother's age at birth.

Table 3.6: Average Effect on Weight-for-age Z-score

	(1)	(2)	(3)	(4)	(5)	(6)
DID Estimate	-0.29	-1.31	-1.44	-0.22	-1.42	-1.80
	(1.76)	(2.02)	(2.64)	(2.30)	(1.83)	(1.52)
Cluster Fixed Effect	Yes	Yes	Yes	Yes	Yes	Yes
Controls	No	No	No	No	No	No
Sample Size	1,022	842	443	356	1,081	1,081
DID Estimate	-0.50	-1.76	-2.60	-0.87	-0.03	-0.66
	(1.47)	(1.75)	(2.63)	(1.17)	(0.72)	(0.67)
Cluster Fixed Effect	Yes	Yes	Yes	Yes	Yes	Yes
Controls	Yes	Yes	Yes	Yes	Yes	Yes
Sample Size	1,022	842	443	356	1,081	1,081

Note: Significance: ∗ 0.10; ∗∗ 0.05; ∗∗∗ 0.01. The columns present the DID estimator using different buffer zones to identify the treated group. Specification (1) reports the results of using 5km buffer zone, (2) using 4km buffer zone, (3) using 3km buffer zone, (4) using 2km buffer zone, (5) using continuous treatment with 1/distance to the dumping site, and (6) using continuous treatment with 1/distance squared to the dumping site. Standard errors are cluster robust at the DHS cluster level. The vector of control variables includes the mother's education, mother's weight, sex of the child, whether the child is born as one of the twins, and the mother's age at birth.

Table 3.7: Average Effect on Height-for-age Z-score

	(1)	(2)	(3)	(4)	(5)	(6)
DID Estimate	0. 35	-0.05	0.22	-0.03	-0.05	-0.24
	(0.50)	(0.52)	(0.61)	(0.66)	(0.56)	(0.22)
Cluster Fixed Effect	Yes	Yes	Yes	Yes	Yes	Yes
Controls	No	No	No	No	No	No
Sample Size	1,002	825	426	340	1,056	1,056
DID Estimate	0.45	0. 11	0.61	0.46	0.17	-0.08
	(0.41)	(0.42)	(0.41)	(0.38)	(0.51)	(0.20)
Cluster Fixed Effect	Yes	Yes	Yes	Yes	Yes	Yes
Controls	Yes	Yes	Yes	Yes	Yes	Yes
Sample Size	1,002	825	426	340	1,056	1,056

Note: Significance: ∗ 0.10; ∗∗ 0.05; ∗∗∗ 0.01. The columns present the DID estimator using different buffer zones to identify the treated group. Specification (1) reports the results of using 5km buffer zone, (2) using 4km buffer zone, (3) using 3km buffer zone, (4) using 2km buffer zone, (5) using continuous treatment with 1/distance to the dumping site, and (6) using continuous treatment with 1/distance squared to the dumping site. Standard errors are cluster robust at the DHS cluster level. The vector of control variables includes the mother's education, mother's weight, sex of the child, whether the child is born as one of the twins, and the mother's age at birth.

Table 3.8: Heterogeneous Effect of Household Characteristics on Diarrhea

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
DID Estimate	0.55***	0.69**	0.43***	0.54**	0.53***	0.64***	0.62***	0.66***
	(0.20)	(0.28)	(0.09)	(0.22)	(0.04)	(0.17)	(0.13)	(0.19)
DDD Estimate	-0.12	-0.11	-0.09	-0.11	-0.14**	-0.14*	-0.48**	-0.47*
	(-0.20)	(0.19)	(0.14)	(0.14)	(0.07)	(0.08)	(0.24)	(0.27)
Cluster Fixed Effect	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Controls	No	Yes	No	Yes	No	Yes	No	Yes
Sample Size	446	446	446	446	445	445	440	440

Note: Significance: ∗ 0.10; ∗∗ 0.05; ∗∗∗ 0.01. The table presents the heterogeneous effect of various household characteristics on the outcome. Column (1) and (2) report the heterogenous effect by whether the household's wealth level is above average or not. Column (3) and (4) present the heterogenous effect by whether the mother received the degree of secondary education or not. Column (5) and (6) show the heterogenous effect by whether the mother was employed by decent jobs or not. Column (7) and (8) present the heterogenous effect by whether families consumed safe sources of drinking water or not. Standard errors are cluster robust at the DHS cluster level. The vector of control variables includes the mother's education, mother's weight, sex of the child, whether the child is born as one of the twins, and the mother's age at birth.

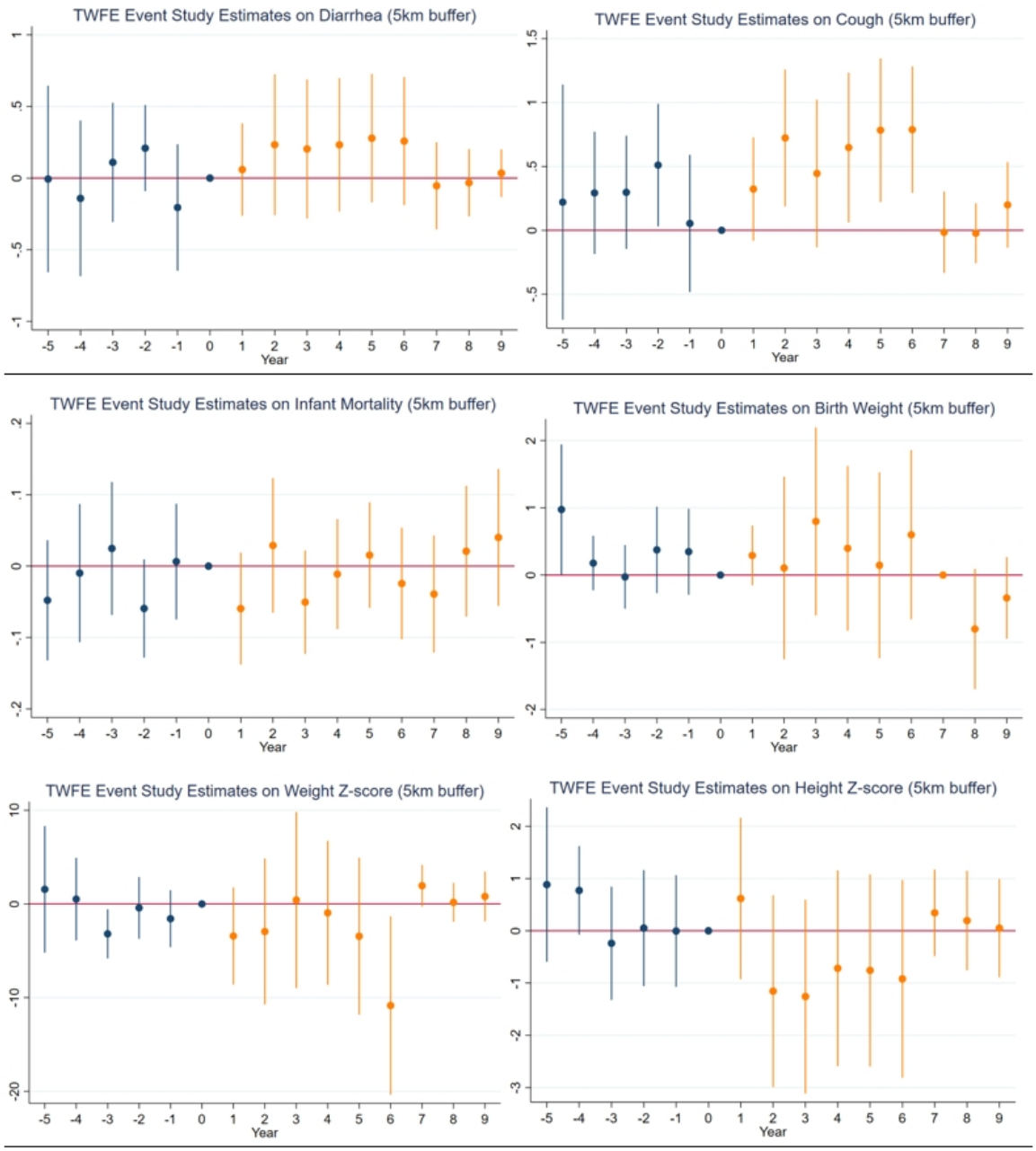
On average, I find a statistically insignificant but universally negative impact on an infant's birth weight of living in the exposure area (Table 3.5). However, when I examine heterogeneous effects on infants' birth weight by maternal tetanus vaccination uptake (Table 3.11), I discover that infants are, on average, 0.95 kilograms lighter (at the 10% level)

delivered by mothers who did not receive a tetanus vaccine in the exposed area compared to the birth weight delivered by mothers who did not receive a tetanus vaccine in the unexposed area. This adverse effect can be significantly (at the 1% level) mitigated by mothers taking the tetanus vaccine before or during pregnancy, which alleviates lower birth weight by 0.68 kilograms (about 72% of the adverse effect).

On average, there is a statistically insignificant and universally negative impact on a child's weight-for-age z-score and no effect on a child's height-for-age z-score of living in the exposure area (Table 3.6 and 3.7). The weight-for-age and height-for-age z-scores do not exhibit significant heterogeneous effects either (Table 3.12 and 3.13). Catch-up growth in low-birth-weight infants is a potential explanation for evidence that I found of lower birth weight in the exposed area but no statistically significant evidence of lower weight-for-age and height-for-age z-scores. Infants with low birth weight may experience rapid postnatal catch-up growth, especially in the first year of life, which can occur spontaneously or be promoted through interventions (Cooke et al., 2023).

I estimate the two-way fixed effect (TWFE) event study estimator for all six health outcomes and present the coefficient plots in Figure 3.2. As illustrated by the coefficient plots, the parallel trend assumption holds in the estimation. The likelihood of infants, whose households live in close vicinity of the e-waste dumping site, having a respiratory illness significantly rises between 2003 and 2008 and it returns to zero after 2009. Despite its weak statistical significance, I also find a consistent increase between 2003 and 2008 in the probability of infants suffering from diarrhea in the treated area. I do not observe significant changes in the coefficient plot of any other outcomes, except that there is a significant reduction in the weight-for-age z-score for those who were born in the year 2008.

Figure 3.2: Coefficient Plots for the TWFE Event Study Estimators



Note. These are the coefficient plots for the TWFE event study estimators for diarrhea, cough, infant mortality, birth weight, weight z-score, and height z-score from top left to down right, using 5 km as the buffer zone for treatment.

I acknowledge that some of the effects are very large, such as the effect on diarrhea, respiratory illness, and the heterogeneous effect of maternal tetanus vaccination on infant birth weight. One possible reason that the effects are large on these outcomes is that the

sample sizes are smaller due to missing values. A smaller sample size leads to less precise regression estimates, which makes the confidence intervals wider. Thus the actual impact is, with 95% probability, somewhere in the 95% confidence interval, which includes values that are much lower than these large point estimates.

3.4.2 The Sources of Drinking Water

In Accra, Ghana, residents consume two main types of water, pipe-borne water and sachet water (Monney et al., 2013). Pipe-borne water or tap water is regarded as unsafe and inappropriate to drink, and in contrast sachet water or bottled water have become the safer alternative source of drinking water (Monney et al., 2013; Kangmennaang et al., 2020). The pipe-borne water is contaminated mainly from two sources. First, the pipelines tend to be poorly maintained and broken at places. The broken pipelines are either unfixed or tied with plastic bags, and they run through sewage water and drains. For example, a picture is shown in Figure 3.3. The e-waste dumpsite is a point source of heavy metals, such as lead, arsenic, and chromium, that release into surrounding water and soil. Water in the leaking pipelines that run through sewage water exchanges substances with the contaminated open water systems (Monney et al., 2013). Lack of safe water storage is another source of contamination. Pipe-borne water is mostly stored in ground level open-air concrete tanks that local residents depend on, and some of the tanks are rarely cleaned (Monney et al., 2013). Ingesting heavy metals, such as cadmium and mercury, in higher amounts can lead to stomach irritation and result in vomiting and diarrhea (Jaishankar et al., 2014).

Figure 3.3: Concrete Water Tanks and Old pipelines in Accra, Ghana

a) b)

Note. Concrete water tanks and old pipelines that run through drains in Accra, Ghana (Monney et al. 2013).

A DHS survey question asks about the household's source of drinking water. It provides detailed options for piped water, tap water, bottled water, sachet water, and other water sources. I created a binary outcome of interest to determine whether a household consumes sources of safe water subject to the answers to the above question. The binary outcome is coded as 1 if the household's source of drinking water is from safe sources such as bottled water or sachet water; the variable is coded as 0 if the source of drinking water comes from contaminated sources such as piped water and tap water.

Table 3.8: Heterogeneous Effect of Household Characteristics on Diarrhea

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
DID Estimate	0.55***	0.69**	0.43***	0.54**	0.53***	0.64***	0.62***	0.66***
	(0.20)	(0.28)	(0.09)	(0.22)	(0.04)	(0.17)	(0.13)	(0.19)
DDD Estimate	-0.12	-0.11	-0.09	-0.11	-0.14**	-0.14*	-0.48**	-0.47*
	(-0.20)	(0.19)	(0.14)	(0.14)	(0.07)	(0.08)	(0.24)	(0.27)
Cluster Fixed Effect	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Controls	No	Yes	No	Yes	No	Yes	No	Yes
Sample Size	446	446	446	446	445	445	440	440

Note: Significance: * 0.10; ** 0.05; *** 0.01. The table presents the heterogeneous effect of various household characteristics on the outcome. Column (1) and (2) report the heterogenous effect by whether the household's wealth level is above average or not. Column (3) and (4) present the heterogenous effect by whether the mother received the degree of secondary education or not. Column (5) and (6) show the heterogenous effect by whether the mother was employed by decent jobs or not. Column (7) and (8) present the heterogenous effect by whether families consumed safe sources of drinking water or not. Standard errors are cluster robust at the DHS cluster level. The vector of control variables includes the mother's education, mother's weight, sex of the child, whether the child is born as one of the twins, and the mother's age at birth.

Table 3.9: Heterogeneous Effect of Household Characteristics on Cough

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
DID Estimate	0.17	0.15	0.53**	0.62**	0.68***	0.73***	0.29	0.36
	(0.30)	(0.32)	(0.23)	(0.27)	(0.21)	(0.27)	(0.24)	(0.28)
DDD Estimate	0.19	0.26	-0.31*	-0.33**	-0.49***	0.50***	0.06	0.21
	(0.22)	(0.22)	(0.16)	(0.16)	(0.18)	(0.19)	(0.21)	(0.08)
Cluster Fixed Effect	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Controls	No	Yes	No	Yes	No	Yes	No	Yes
Sample Size	446	446	446	446	445	445	440	440

Note: Significance: * 0.10; ** 0.05; *** 0.01. The table presents the heterogeneous effect of various household characteristics on the outcome. Column (1) and (2) report the heterogenous effect by whether the household's wealth level is above average or not. Column (3) and (4) present the heterogenous effect by whether the mother received the degree of secondary education or not. Column (5) and (6) show the heterogenous effect by whether the mother was employed by decent jobs or not. Column (7) and (8) present the heterogenous effect by whether the household consumed safe sources of drinking water or not. Standard errors are cluster robust at the DHS cluster level. The vector of control variables includes the mother's education, mother's weight, sex of the child, whether the child is born as one of the twins, and the mother's age at birth.

I discuss in Section 3.4.1 that having safe sources of drinking water significantly mitigates the risk of an infant suffering from diarrhea by 47 percentage points (about 71% of the adverse effect). To understand the dynamics of how differences in family characteristics lead to the take-up of safe or unsafe sources of water, I implement a test from Pei et al. (2019) by running a series of regressions with water take-up on the left-hand side of the regressions and regress it on a household characteristic variable, the DID treatment estimate, and an interaction term between the household characteristic variable and the DID treatment estimate. The results are reported in Table 3.14.

I find that, without the treatment in place, households having a higher wealth level, having mothers who have completed at least secondary education, and having mothers whose occupation is less likely to be exposed to the e-waste pollution have a significantly higher take-up of safe sources of water. Very interestingly, the gap in safe water take-up between the wealthier and the less wealthy households significantly closes by 15 percentage points in the treated group, but not fully mitigating the income effect (there is still a five-percentage-point difference in the take-up probability). This suggests that the less wealthy households are aware of the detrimental effects of drinking unsafe sources of water in the treated area because they would not spend money on the more expensive alternatives of water if they were in the untreated area. This leads to important policy implication that providing safe sources of drinking water at a more affordable rate to the less wealthy families is the key to close the gap in safe water consumption and to reduce child diarrhea in the exposed area.

3.4.3 The Risk of Tetanus Infection

Tetanus is caused by the bacterium *Clostridium tetani*, commonly found in soil, dust, and animal feces. Infection typically occurs when the bacteria enter the body through open wounds or punctures. Living in the vicinity of an e-waste dumping site, where poor sanitation and frequent injuries from sharp metal fragments are common, increases the risk of tetanus infections (Parvez et al., 2021). Maternal tetanus vaccination plays a crucial role in preventing tetanus infections in children, especially before the Tetanus, Diphtheria, and Pertussis (Tdap) vaccine is administered to children. When a mother is vaccinated, her immune system produces antibodies against the bacteria. These antibodies can cross the placenta and be transferred to the fetus, offering the child early protection until they are old enough to receive their own vaccinations (Campbell et al., 2018).

Having immunity against the tetanus bacteria is critical for mothers and children near an e-waste dumping site, as maternal and neonatal tetanus infections continue to be a major cause of neonatal mortality with high fatality rates among infants born to inadequately vaccinated mothers following unsanitary childbirth practices, particularly in low-income countries (Kanu, 2022). Maternal and neonatal tetanus infections can also lead to premature births, as infants who exhibit lower levels of anti-tetanus antibodies are more like to be born prematurely when there is a tetanus infection (Perin et al., 2012).

Therefore, tetanus vaccination and its potential mitigation effect is a key mechanism that I examine. I explore heterogeneous effects on infants' birth weight by maternal tetanus vaccination uptake and present the results in Table 3.11. I find that infants are, on average, 0.95 kilograms lighter when their mothers have not received at least one dose of tetanus vaccine in the exposed area compared to infants' birth weight of whose mothers have not received a tetanus vaccine in the unexposed area. This adverse effect can be significantly (at the 1% level) mitigated when mothers take at least one dose of tetanus vaccine before or during pregnancy. This mitigates infants' lower birth weight by 0.68 kilograms (about 72% of the adverse effect).

Table 3.10: Heterogeneous Effect of Household Characteristics on Infant Mortality

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
DID Estimate	0.01	0.02	0.03	0.03	-0.01	-0.01	0.24	0.24
	(0.02)	(0.02)	(0.04)	(0.03)	(0.03)	(0.03)	(0.26)	(0.26)
DDD Estimate	-0.06**	-0.03	-0.03	-0.02	0.05	0.05	-0.24	-0.25
	(0.03)	(0.02)	(0.03)	(0.03)	(0.06)	(0.06)	(0.26)	(0.26)
Cluster Fixed Effect	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Controls	No	Yes	No	Yes	No	Yes	No	Yes
Sample Size	3341	3341	3341	3341	1424	1424	545	545

Note: Significance: * 0.10; ** 0.05; *** 0.01. The table presents the heterogeneous effect of various household characteristics on the outcome. Column (1) and (2) report the heterogenous effect by whether the household's wealth level is above average or not. Column (3) and (4) present the heterogenous effect by whether the mother received the degree of secondary education or not. Column (5) and (6) show the heterogenous effect by whether the mother was employed by decent jobs or not. Column (7) and (8) present the heterogenous effect by whether the mother received at least one dose of tetanus vaccine before or during pregnancy or not. Standard errors are cluster robust at the DHS cluster level. The vector of control variables includes the mother's education, mother's weight, sex of the child, whether the child is born as one of the twins, and the mother's age at birth.

Table 3.11: Heterogeneous Effect of Household Characteristics on Birth Weight

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
DID Estimate	-0.23	-0.23	-0.25	-0.18	-0.15	-0.12	-0.89*	-0.95*
	(0.25)	(0.26)	(0.23)	(0.26)	(0.20)	(0.20)	(0.49)	(0.51)
DDD Estimate	0.06	0.11	0.07	0.03	-0.03	-0.32	0.60**	0.68***
	(0.22)	(0.21)	(0.15)	(0.17)	(0.27)	(0.27)	(0.24)	(0.22)
Cluster Fixed Effect	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Controls	No	Yes	No	Yes	No	Yes	No	Yes
Sample Size	732	732	732	732	657	657	611	611

Note: Significance: ∗ 0.10; ∗∗ 0.05; ∗∗∗ 0.01. The table presents the heterogeneous effect of various household characteristics on the outcome. Column (1) and (2) report the heterogenous effect by whether the household's wealth level is above average or not. Column (3) and (4) present the heterogenous effect by whether the mother received a degree of secondary education or not. Column (5) and (6) show the heterogenous effect by whether the mother was employed by decent jobs or not. Column (7) and (8) present the heterogenous effect by whether the mother received at least one dose of tetanus vaccine before or during pregnancy or not. Standard errors are cluster robust at the DHS cluster level. The vector of control variables includes the mother's education, mother's weight, sex of the child, whether the child is born as one of the twins, and the mother's age at birth.

Table 3.12: Heterogeneous Effect of Household Characteristics on Weight-for-age Z-score

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
DID Estimate	-4.60	-0.69	-1.18	2.44	-1.77	-0.64	3.00	6.13
	(3.69)	(0.26)	(2.55)	(2.71)	(2.24)	(-1.23)	(7.70)	(7.45)
DDD Estimate	3.31	1.48	1.01	-0.07	1.43	0.88	-3.73	-4.44
	(3.51)	(0.21)	(1.85)	(1.77)	(1.24)	(5.35)	(7.14)	(6.55)
Cluster Fixed Effect	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Controls	No	Yes	No	Yes	No	Yes	No	Yes
Sample Size	431	431	431	431	430	430	358	358

Note: Significance: ∗ 0.10; ∗∗ 0.05; ∗∗∗ 0.01. The table presents the heterogeneous effect of various household characteristics on the outcome. Column (1) and (2) report the heterogenous effect by whether the household's wealth level is above average or not. Column (3) and (4) present the heterogenous effect by whether the mother received a degree of secondary education or not. Column (5) and (6) show the heterogenous effect by whether the mother was employed by decent jobs or not. Column (7) and (8) present the heterogenous effect by whether the mother received at least one dose of tetanus vaccine before or during pregnancy or not. Standard errors are cluster robust at the DHS cluster level.

Table 3.13: Heterogeneous Effect of Household Characteristics on Height-for-age Z-score

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
DID Estimate	0.60	0. 52	0.20	0.34	0.73	0.74	-0.93	0.52
	(1.43)	(1.21)	(1.06)	(0.96)	(1.02)	(0.92)	(0.70)	(1.18)
DDD Estimate	-0.74	-0.46	-0.04	0.20	0.24	0.12	0.84	0.93
	(0.92)	(0.69)	(0. 56)	(0.45)	(1.15)	(0.69)	(0.93)	(0.70)
Cluster Fixed Effect	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Controls	No	Yes	No	Yes	No	Yes	No	Yes
Sample Size	415	415	415	415	414	414	343	343

Note: Significance: ∗ 0.10; ∗∗ 0.05; ∗∗∗ 0.01. The table presents the heterogeneous effect of various household characteristics on the outcome. Column (1) and (2) report the heterogenous effect by whether the household's wealth level is above average or not. Column (3) and (4) present the heterogenous effect by whether the mother received the degree of secondary education or not. Column (5) and (6) show the heterogenous effect by whether the mother was employed by decent jobs or not. Column (7) and (8) present the heterogenous effect by whether the mother received at least one dose of tetanus vaccine before or during pregnancy or not. Standard errors are cluster robust at the DHS cluster level.

To understand how tetanus vaccination is related to household characteristics, I implement a Pei et al. (2019) test by running a series of regressions with tetanus vaccine take-up on the left-hand side of the regressions and regress it on household characteristics variable, the DID treatment estimate and an interaction term between the household characteristics variable and the. DID treatment estimate. The results are presented in Table 3.15.

Table 3.14: Heterogeneous Effect of Household Characteristics on Safe Water Take-up

	(1)		(2)		(3)
Wealth	0.20***	Education	0.19***	Job	0.18**
	(0.04)		(0.07)		(0.09)
Treatment	0.03	Treatment	-0.08	Treatment	-0.06
	(0.07)		(0.08)		(0.06)
Wealth*Treatment	-0.15**	Education*Treatment	0.08	Job*Treatment	0.06
	(0.07)		(0.09)		(0.14)
Cluster Fixed Effect	Yes	Cluster Fixed Effect	Yes	Cluster Fixed Effect	Yes
Controls	No	Controls	No	Controls	No
Sample Size	2997	Sample Size	2997	Sample Size	1547

Note: Significance: ∗ 0.10; ∗∗ 0.05; ∗∗∗ 0.01. The table presents the results of testing whether the household's wealth level, whether the mother received a degree of secondary education or not, and whether the mother was employed by decent jobs or not affect the take-up of safe sources of drinking water. Standard errors are cluster robust at the DHS cluster level. The vector of control variables includes the mother's education, mother's weight, sex of the child, if the child is born as one of the twins, and the mother's age at birth.

Table 3.15: Heterogeneous Effect of Household Characteristics on Tetanus Vaccine Take-up

	(1)		(2)		(3)
Wealth	0.08**	Education	0.05	Job	0.07***
	(0.03)		(0.04)		(0.02)
Treatment	0.05	Treatment	0.05	Treatment	0.05**
	(0.04)		(0.04)		(0.02)
Wealth*Treatment	-0.03	Education*Treatment	-0.02	Job*Treatment	-0.25**
	(0.04)		(0.05)		(0.10)
Cluster Fixed Effect	Yes	Cluster Fixed Effect	Yes	Cluster Fixed Effect	Yes
Controls	No	Controls	No	Controls	No
Sample Size	670	Sample Size	670	Sample Size	1547

Note: Significance: * 0.10; ** 0.05; *** 0.01. The table presents the results of testing whether the household's wealth level, whether the mother received a degree of secondary education or not, and whether the mother was employed by decent jobs or not affect the take-up of safe sources of drinking water. Standard errors are cluster robust at the DHS cluster level. The vector of control variables includes the mother's education, mother's weight, sex of the child, if the child is born as one of the twins, and the mother's age at birth.

Household wealth level and mother's occupation both have significant positive effects on the vaccine take-up by 8 and 7 percentage points, respectively. More importantly, in the treated area, mothers whose jobs are riskier to tetanus infection are more likely to take the tetanus vaccine than mothers whose works are less risky to tetanus infection, even though occupations, such as clerical and managerial jobs, that are less directly exposed to e-waste and tetanus infection tend to pay at a higher rate. This provides evidence that, in the exposed area, there is not necessarily a gap in the tetanus vaccine take-up between the wealthier and less wealthy mothers. It means that mothers working in the e-waste recycling industry are aware of the potential risk of tetanus infection and they utilize the free vaccines offered by the Expanded Programme on Immunization (EPI) (Dow et al., 1999). Another explanation would be that certain government agency or other aid agency (local or international) made efforts to vaccinate women who work at e-waste sites. However, I have not found programs or agencies that provide additional aid targeting the e-waste workers in Ghana. With assistance from the EPI, immunization against the six childhood diseases (i.e., diphtheria, measles, pertussis, poliomyelitis, tetanus, and tuberculosis) has been instituted as part of Ghana's primary health care program available for the entire population (Nyarko et al., 2001). Therefore, the local government should keep providing free essential vaccines to the general public, especially to the households that have an e-waste worker and to the more disadvantaged households.

3.5 Robustness Checks

I perform multiple placebo tests and robustness checks to assess the validity of the specifications and the results. In Section 3.5.1, I run a placebo test using a placebo year of treatment (1995) holding every else the same in the specification. In Section 3.5.2, I perform a robustness check using a placebo outcome. In Section 3.5.3, I conduct a robustness check using only the DHS data (without the MICS data). In Section 3.5.4, I discuss the potential impact of missing values on the validity of the results. Section 3.5.5 discusses other potential threats to the validity of estimation.

3.5.1 Placebo Treatment Test

I implement placebo tests by running a series of regressions (Equation (3.1)) on a child's health outcomes but using a different year of treatment. The placebo test uses 1995 as the year when treatment started and uses the data from 1990 to 1995 as the pre-treatment period and 1995 to 1999 as the post-treatment period. Table B.1 reports the results of the placebo treatment test. The results show that none of the estimates of interest are significant across all specifications, and the effect sizes are also close to zero.

3.5.2 Placebo Outcome Test

I perform a robustness check using a placebo outcome - whether the child was born as one of the twins or not. To my knowledge, no research has shown that exposure to the pollutants released from e-waste causes women more likely to have babies in twins. The results are presented in Table B.2. I run a series of regressions to test this outcome, with everything else remaining the same. Across all specifications, the DID estimates are insignificant and close to zero.

3.5.3 Using Only the DHS Data

I conduct a robustness check for effects on diarrhea without including the MICS data. I chose to perform the test on diarrhea because this outcome exhibits the most statistically significant results in all specifications in Table 3.2. It is important to verify the robustness of this outcome. As shown in Table B.3, without including the MICS data, all of the estimates are still statistically significant at the 1% level, and the magnitude of the coefficients is very close to the coefficients in Table 3.2.

3.5.4 Potential Impact of Missing Values on Results

Admittedly, the DHS data are not short of missing values in many variables. The outcome variables that I have the most missing values are the infant's birth weight, diarrhea, the respiratory illness, weight-for-age z-score, and height-for-age z-score. For every outcome in the treatment and the control group, I construct a table (Table B.4) to show the percentage of households among the missing data in the corresponding outcome that are below the average household wealth, have mothers whose highest degree completed are lower than secondary education, and have mothers work in the industries that are more likely to be exposed to e-waste pollution.

For all household characteristics, the missing data in the treatment group has a lower percentage of households that are below the average level of wealth, but higher percentages of households that have mothers with lower education degrees and work in more concerning conditions. As shown in Table 3.8, household wealth level does not have a significant impact on diarrhea, while the mother's occupation has a significant mitigating effect on the adverse outcome. On the other hand, wealthier households are more likely to afford safe alternatives to drinking water, even though poorer households in the treated group are also more incentivized to spend money on safe sources of water and close the gap. Therefore, the impact of missing values on the estimates of diarrhea is ambiguous.

For the respiratory illness, household wealth level is not relevant, whereas the mother's education and occupation have a significant mitigating impact on reducing the likelihood of a respiratory illness (Table 3.9). Sources of water are showing no effect on the respiratory illness either. Therefore, the estimates of the respiratory illness are likely to be underestimated.

As presented in Table 3.10, household wealth level has a mitigating effect on infant mortality, and other variables do not appear to exhibit a significant effect. Thus, the estimates of infant mortality are more likely to be overestimated. Nonetheless, I do not make important policy implications based on the findings on infant mortality throughout the paper.

None of the three household characteristics are shown to have a significant impact on a child's birth weight (Table 3.11). Based on the heterogeneous effect of tetanus vaccines on birth weight presented in Table 3.11 and the effect of household characteristics on tetanus vaccine take-up in Table 3.15. It is evident that none of the household characteristics appear to have a substantial effect on a child's birth weight. Mothers in the treated area who work in e-waste-related jobs are strongly inclined to take the tetanus vaccine which is offered for free in Ghana thanks to the EPI program. Therefore, the missing values do not affect the

estimates of a child's birth weight.

I do not find any significant effect on a child's weight-for-age z-score and height-for-age z-score of living in the exposure area. Therefore, the potential impact of missing values on the weight-for-age z-score and height-for-age z-score is the least concerning. In conclusion, the missing values in the outcome variables do not affect the overall validity of the findings in this paper.

3.5.5 Discussion of Potential Threats to Estimation Validity

According to Callaway et al. (2024), for the results in a DID strategy with a continuous treatment to hold, two assumptions need to be satisfied. First, the parallel trends assumption must hold. Based on the coefficient plots generated from the TWFE event study and the results from various placebo tests, the parallel trends assumption for the setup of this paper holds. Second, the variation of treatment intensity can be partially endogenous if there is self-selection into different amounts of treatment intensity. Thus, a stronger assumption is required to rule out selection into different treatment intensities. This assumption also holds for this paper, because I verified the number of years each household had been living in the reported location, and I included only the households that had been living in the surveyed location as long as the woman was pregnant. No migration or selection into treatment intensity occurred after the women became pregnant. Therefore, the health outcomes of the child are not biased by self-selection.

The second potential threat is attenuation bias embedded in the geocode of the DHS data. For confidentiality, the DHS implements a minimum of 0 and a maximum of 2 kilometers of geographic noise for urban clusters in a random direction. Theoretically, a potential attenuation bias leads to underestimation of the coefficients due to addition measurement error (Wooldridge, 2016). The formula of attenuation bias is as follows,

$$\text{plim}\left(\hat{\beta}_1\right) = \beta_1 \left(\frac{\sigma^{2*}_{x_1}}{\sigma^{2*}_{x_1} + \sigma^2_{e_1}}\right), \tag{3.5}$$

where the estimated $\hat{\beta}$ has a smaller magnitude than the true β due to an additional variance from the attenuation bias in the denominator. Therefore, the attenuation bias resulting from the noise in geocode implemented by the DHS is unlikely to bias the estimate away from zero and lead to underestimation of the effects. Empirically, however, Michler et al. (2022) have found that spatial anonymization techniques, on average, have limited to no impact

on estimates of the relationship between weather and agricultural productivity. Therefore, for this paper, I argue that the spatial anonymization technique introduced by the DHS may lead to underestimation of the estimates or has no effect on the estimates.

3.6 Conclusion

This paper has assessed the effects of living in proximity to an e-waste dumpsite, Agbogbloshie, on infant health outcomes in Ghana. The estimates based on the difference-in-differences specifications provide evidence that living in the vicinity of Agbogbloshie significantly increases the probability of a child suffering from diarrhea by 52 percentage points. The adverse effect increases for households living closer to the dumping site. When households do not have access to safe sources of drinking water, children in the exposed area are 66 percentage points more likely to have diarrhea than children who do not have access to secure drinking water in the unexposed area. Having safe sources of drinking water significantly mitigates the risk of an infant having diarrhea by 47 percentage points (about 71% of the total adverse effect).

After studying the household characteristics and the consumption of safe alternatives of drinking water, I find evidence that the less wealthy households are aware of the deleterious effects of drinking unsafe sources of water in the treated area. Less wealthy families are less willing to spend money on the more expensive alternatives of water if they live in an unexposed area.

I find that living in the exposure area of the e-waste dumping site, Agbogbloshie, increases the likelihood that an infant has a respiratory illness by about 50 percentage points. The likelihood that an infant suffers from a respiratory illness is significantly alleviated by 50 percentage points (about 68% of the total adverse effect) when the mother has an occupation that is less likely to be exposed to the pollution.

I also find that infants are 0.95 kilograms lighter delivered by mothers who did not receive a tetanus vaccine in the exposed area compared to the birth weight delivered by mothers who did not receive a tetanus vaccine in the unexposed area. However, this adverse effect can be significantly mitigated by mothers taking the tetanus vaccine before or during pregnancy, which alleviates lower birth weight by 0.68 kilograms (about 72% of the total adverse effect).

A mother's occupation has a significantly positive effect on the vaccine take-up. In the treated area, mothers whose jobs are more susceptible to tetanus infection are more likely to

take the tetanus vaccine than mothers whose works are less susceptible to tetanus infection. Therefore, in the exposed area, there is not necessarily a gap in the tetanus vaccine take-up between mothers from the wealthier and the less wealthy households. It also shows evidence that people working in the e-waste recycling industry are aware of the potential risk of tetanus infection and they are willing to take vaccines to redress the corresponding risk.

The findings provide meaningful policy implications. In the long run, the international world should seek to ban the import and export of e-waste. The developed countries need to recycle their e-waste rather than dumping it into the developing countries at a trivial expense. The 1995 Basel Ban Amendment, a global waste dumping prohibition, has become an international law after being ratified by Croatia in 2019 as a result of the Basel Convention in 2011. We need to keep the effective execution of this international law.

In the short run, local governments should establish programs that assist e-waste collectors and handlers to find alternative employment. For those who cannot find alternative employment immediately, local governments are recommended to provide safe sources of drinking water at a more affordable rate to the less wealthy families, in order to close the gap in the consumption of safe drinking water and to reduce child diarrhea in the exposed area. Local government should also keep providing free essential vaccines to the general public, especially to the households that are more susceptible to the potential adverse effects of e-waste dumping sites.

Chapter 4

The Decision-Making of College Enrollment in an Increasingly Independent World

4.1 Introduction

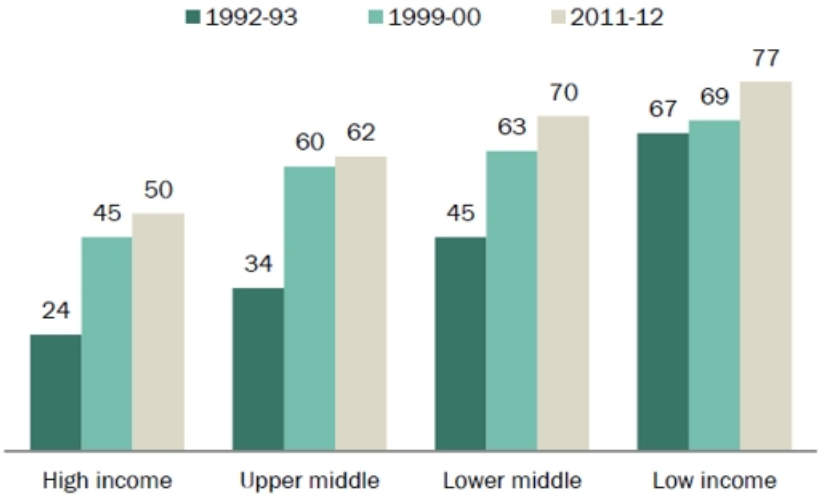
The decision-making process behind students' higher education choices has long been a central topic in education research. Understanding how students make these decisions, and why disparities in college attainment persist, is critical for identifying the forces that shape college access and for designing policies to reduce educational inequality. College education in the United States is among the most expensive in the world. According to the National Center for Education Statistics (2024), the average annual tuition and fees for full-time undergraduate students in 2020–21 were $9,400 at public four-year in-state institutions and $37,600 at private nonprofit four-year institutions, not including room, board, and other expenses. Without sufficient financial support, students from low-income households face significant financial barriers to accessing higher education, which contributes to persistent inequality in college attainment (National College Attainment Network, 2023).

Student loans are intended to reduce college enrollment disparities by providing financial support that enables students from disadvantaged backgrounds to pursue a college degree. In the United States, there has been a growing trend in student loan borrowing to finance higher education across all income levels (National Center for Education Statistics, 2015). As shown in Figure 4.1, the share of college graduates carrying student loan debt has

steadily increased since 1992 across every income group. A substantial and growing share of students now rely on loans to cover tuition costs, regardless of household income. Moreover, the distribution of student loan debt has become more balanced over time, with graduates from all socioeconomic backgrounds holding a more equal share of the overall debt burden (Fry, 2014).

This paper develops a two-period discrete choice model to examine how higher education decisions are shaped when student loans play an increasingly crucial role. The existing literature typically views parental investments in children's education as an instrument for transferring income between time periods, where the parent and child work as a single decision-making agent (Cai and Heathcote, 2022; Abbott et al., 2019; Keane and Wolpin, 2001). In our model, education is not perceived as a tool for intergenerational transfers but a decision that is independently made by the children to maximize their life-cycle utility, and children fully fund their higher education through borrowing. Our model is different from the setting in other papers, as we aim to evaluate disparities in higher education attainment under a constructed scenario in which all students, regardless of family income, rely on student loans. It is useful to model a scenario where all students rely completely on student loans, because the model highlights and helps us better understand whether disparities in higher education attainment are due to unequal access to financing or to deeper structural inequalities (for example, inequalities in K-12 preparation).

Figure 4.1: Percent of College Graduates with Student Loan Debt



Note: For the class of 2011-12, graduates in the lowest income quartile had parental incomes below $44,000 and those in the highest income quartile had parental incomes more than $125,700. Source: Pew Research Center (Fry, 2014).

The second contribution of this paper is that we model college quality as an endogenously selected component of students' higher education decisions. In reality, students do not simply choose whether to attend college, but also which college to attend, weighing considerations of quality, cost, and personal fit. The existing literature, for example Black and Smith (2004), documented substantial ability-based sorting across colleges of different qualities. However, much of these work treats college quality as an exogenous attribute rather than modeling the endogenous sorting process itself. These studies typically measure college quality using proxies such as the average SAT scores of entering classes, the high school GPA, the average salaries of faculty, or freshman retention rates (Black and Smith, 2004; Dale and Krueger, 2011; Hendricks et al., 2018). In contrast, our model incorporates college quality as an endogenous choice, affected by students' learning efficiency, expected labor market returns after graduation, the base price of college, and the interest rates associated with student loans. Endogenously modeling college quality allows for a more accurate assessment of how policies targeting these factors affect both students' college attendance decisions and the quality of the institutions they ultimately attend.

The main data used in this study come from the National Longitudinal Survey of Youth 1997 (NLSY97), a nationally representative survey of American men and women born between 1980 and 1984. We also incorporate data from the U.S. Department of Labor,

the Federal Reserve Economic Data (FRED), and the Education Data Initiative to obtain information on the federal minimum wage, the Federal Funds Effective Rate, and the average in-state tuition for community colleges, respectively. To estimate our model, we first construct the key variables using the available data, and then calibrate the model by matching it to the moments to estimate the model parameters.

Based on our model, we evaluate the effectiveness of various policy alternatives designed to promote educational attainment and close the gap in educational disparities across different income levels. We assess the impact of the following policy alternatives on college enrollment and the endogenously selected college quality: (1) investing in pre-college education to improve children's cognitive skills before college, (2) reducing college tuition costs, and (3) eliminating interest rates on student loans. Our findings indicate that the most effective strategy for reducing inequality in both college enrollment and the quality of college attended is to provide all children with more equitable access to high-quality education at an early age, which strengthens their cognitive abilities before college. Lowering the interest rate on student loans has only a limited impact on college attendance. Additionally, reducing baseline tuition costs disproportionately benefits students from higher-income families, offering relatively less support to those from lower-income backgrounds. The main driving force behind these findings is that the differences in predicted college enrollment rates and endogenously chosen college quality across different household income levels primarily come from variations in AFQT scores.

The structure of the rest of the paper is as follows. Section 4.2 demonstrates the structural model and explains the estimation strategy. Section 4.3 describes the data used for the analysis. Section 4.4 discusses variable measurement, presents model estimation results, and evaluates model performance using three validation approaches. Section 4.5 presents the policy simulations and implications derived from the model. Section 4.6 provides additional discussion, and Section 4.7 concludes.

4.2 Model

We construct a two-period discrete choice model. At the beginning of period 1, a high school graduate decides whether to attend college ($j = 1$) or enter the workforce ($j = 0$). Those who choose to attend college spend the entirety of period 1 in education. At the beginning of period 2, individuals who attended college in period 1 graduate from university. Regardless of their education decision at the beginning of period 1, all individuals work for

the entire duration of period 2. Each agent aims to maximize their lifetime expected utility U_{ij}, where i denotes a representative agent and j represents the choice made.

Before presenting the model, we first define the two periods more precisely. This paper uses data from the National Longitudinal Survey of Youth 1997 (NLSY97), which includes a nationally representative sample of 8,984 American men and women who were between 12 and 16 years old in 1996. Respondents were interviewed annually from 1997 to 2017. Typically, students graduate from a four-year college or university.[1] For respondents in the NLSY97, the average expected year of college graduation is 2008. In this two-period model, period 1 is defined as the first four years following high school graduation. For those pursuing higher education, this represents the time spent in college and the opportunity cost of not working during this period. Period 2 is defined as the ten years following the expected year of college graduation (approximately 2008–2017), during which individuals are assumed to be in the workforce.

Assuming a linear utility function, the lifetime expected utility of a two-period-lived individual is given by:

$$\mathbb{E}(U_{ij}) = V_{ij} = u(c_0) + \beta u(c_1) = C_{ij,t=1} + \delta C_{ij,t=2}, \tag{4.1}$$

where δ is the discount factor for future consumption, assumed to be constant across all individuals. Note that U_{ij} represents the true utility of a representative agent and consists of two components: the observable component V_{ij} and the unobserved error term ε_{ij}. Written as $U_{ij} = V_{ij} + \varepsilon_{ij}$, V_{ij} is the part of the true utility captured by the researcher, and ε_{ij} is the unobserved component, which is not explicitly included in this model. The nature of ε_{ij} will be further illustrated.

In period 1, all children receive a subsistence-level endowment y from their parents, regardless of whether they choose to attend college. It is assumed that no financial support from parents is expected in period 2.

If a child chooses to work in period 1, their consumption in period 1 is given by:

$$C_{i,j=0,t=1} = y_i + w_{i,j=0,t=1} = y_i + a_0 L_{i1}, \tag{4.2}$$

where L_{i1} is the expected total labor input in period 1, a_0 is the productivity of basic skills in the labor market, and $w_{i,j=0,t=1}$ is the total earnings of a child who does not attend

[1] In the NLSY97, among those who have a college degree, 72.1% students graduated from a four-year college instead of a two-year college between 1997 and 2017.

college in period 1. If a child chooses to work in period 1, their consumption in period 2 is:

$$C_{i,j=0,t=2} = w_{i,j=0,t=2} = a_0 L_{i2}, \tag{4.3}$$

where L_{i2} is the expected total labor input in period 2. Therefore, if a child chooses not to attend college, their expected lifetime utility is:

$$V_{i,j=0} = y_i + a_0 L_{i1} + \delta a_0 L_{i2}. \tag{4.4}$$

If a child chooses to attend college in period 1, their consumption in period 1 is:

$$C_{i,j=1,t=1} = y_i, \tag{4.5}$$

where the subsistence level remains unchanged regardless of the education choice. Students borrow loans from financial institutions to finance college expenses, so their period 2 consumption, if they choose to attend college in period 1, is given by:

$$C_{i,j=1,t=2} = w_{i,j=1,t=2} - (1 + r_i) p_i Q_i, \tag{4.6}$$

where $w_{i,j=1,t=2}$ is the wage of an individual after they complete college education. According to Glewwe (2002), higher college quality is associated with higher college costs. Following this, college expenses are expressed as $p_i Q_i$, where p_i is the base price of college in the student's state of residence and Q_i is an index of endogenous college quality. The prevailing interest rate at the time of decision-making is denoted by r_i. Students borrow loans from financial institutions in period 1 to cover tuition costs only and repay these loans in period 2. In this model, both the prevailing interest rate and the base price of college are assumed to be exogenous to each individual student.

The wage for a college graduate in period 2 is specified as:

$$w_{i,j=1,t=2} = a_{1i} L_{i2} = \alpha_i \pi f(Q_i) L_{i2}, \tag{4.7}$$

where a_{1i} is the productivity of a college graduate in the job market, α_i is the individual i's learning efficiency, π is the productivity of skills acquired in college (assumed to be constant and identical for all individuals), and L_{i2} is the expected labor input for person i in period 2. The decomposition of the productivity of a college graduate a_{1i} is based upon Glewwe (2002). We assume the functional form of $f(Q_i)$ to be $f(Q_i) = Q_i^{\beta}$ (Glewwe, 2002),

and impose $0 < \beta < 1$ to ensure concavity of $V_{i,j=1}$, as this guarantees g $f'(Q_i) > 0$ and $f''(Q_i) < 0$, thereby exhibiting a diminishing return of college quality.

The assumption is that the decision-maker is fully aware of their expected lifetime labor input L_{i1} and L_{i2} at the time of making the higher education decision, and they internalize this information in their decision-making process. Therefore, putting everything together, the expected lifetime utility of an individual who chooses to attend college can be written as:

$$V_{i,j=1} = y_i + \delta\alpha_i\pi Q_i^{\beta} L_{i2} - \delta(1 + r_i)p_i Q_i. \tag{4.8}$$

Based on the specified model, a child chooses to attend college if $\Delta V_i = V_{i,j=1} - V_{i,j=0} > 0$, or more specifically if:

$$\Delta V_i = \delta\alpha_i\pi Q_i^{\beta} L_{i2} - \delta(1 + r_i)p_i Q_i - a_0 L_{i1} - \delta a_0 L_{i2} > 0. \tag{4.9}$$

If a child decides to attend college, the optimal value of college quality Q_i can be obtained by the first-order condition that maximizes their expected lifetime utility from attending college:

$$Q_i^* = \left(\frac{\beta\alpha_i\pi L_{i2}}{(1 + r_i)p_i}\right)^{\frac{1}{1-\beta}}. \tag{4.10}$$

Given that $\beta \in (0, 1)$, the implications of Equation (4.10) are straightforward: students choose to attend a higher-quality college if they expect greater learning efficiency (larger values of α_i), higher expected labor input after college graduation, a lower base price of college, or a lower interest rate. Next, we replace college quality Q_i in Equation (4.9) with its optimal value, Q_i^*, which is determined by a set of student characteristics, the exogenous interest rate, and the exogenous base price of college in the student's state of residence.

We then use a logistic model to estimate how the endogenous and exogenous parameters affect individual decision-making, as formulated in Equation (4.9). To do so, recall that the true utility function for individual i is:

$$U_{ij} = V_{ij} + \varepsilon_{ij}, \tag{4.11}$$

where U_{ij} is nonstochastic and reflects the individual's true utility, ε_{ij} is stochastic and captures idiosyncratic variations in individual tastes for alternative j, and V_{ij} is the "repre-

sentative utility" observed by the researcher (Train, 2009).

The probability that individual i chooses to attend college ($j = 1$) is denoted as P_{i1}, where

$$\begin{aligned} P_{i1} &= P[V_{i1} + \varepsilon_{i1} > V_{i0} + \varepsilon_{i0}] \\ &= P[\varepsilon_{i0} - \varepsilon_{i1} < V_{i1} - V_{i0}] \\ &= F[V_{i1} - V_{i0}]. \end{aligned} \tag{4.12}$$

Assume that all ε_{ij} are independent and identically distributed (i.i.d.) draws from a type I Extreme Value Distribution, so that the difference between two randomly drawn error term $\varepsilon's$ $(\varepsilon_{i1} - \varepsilon_{i0})$ follows a logistic distribution (McFadden, 1973). Given that the difference between two random error terms $(\varepsilon_{i1} - \varepsilon_{i0})$ and the corresponding difference between the utility levels $(V_{i1} - V_{i0})$ both follow a logistic distribution with $mean = 0$ and $variance = 1$, the probability of attending college P_{i1} can be expressed as follows:

$$\begin{aligned} P_{i1} &= F[V_{i1} - V_{i0}] \\ &= \frac{1}{1 + exp\,(V_{i0} - V_{i1})}, \end{aligned} \tag{4.13}$$

which can be written in the following complete closed-form expression in full terms:

$$\frac{1}{1 + exp\left(a_0 L_{i1} + \delta a_0 L_{i2} - \delta \alpha_i \pi L_{i2} \left(\frac{\beta \alpha_i \pi L_{i2}}{(1+r_i)p_i}\right)^{\frac{\beta}{1-\beta}} + \delta(1+r_i)p_i \left(\frac{\beta \alpha_i \pi L_{i2}}{(1+r_i)p_i}\right)^{\frac{1}{1-\beta}}\right)}. \tag{4.14}$$

To summarize the model, it can be presented as the following system of equations:

$$\begin{cases} (U_{ij}) = V_{ij} = C_{ij,t=1} + \delta C_{ij,t=2}, \\ V_{i,j=0} = y_i + a_0 L_{i1} + \delta a_0 L_{i2}, \\ V_{i,j=1} = y_i + \delta a_{1i} L_{i2} - \delta(1+r_i)p_i Q_i, \\ a_{1i} = \alpha_i \pi Q_i^{\beta}, \\ Q_i^* = \left(\frac{\beta \alpha_i \pi L_{i2}}{(1+r_i)p_i}\right)^{\frac{1}{1-\beta}}, \\ P_{i1} = \frac{1}{1+exp(V_{i0}-V_{i1})}. \end{cases} \tag{4.15}$$

The interpretation of the role of the error terms in Equation (4.12) is as follows. Having $V_{i1} > V_{i0}$ does not necessarily mean that the individual chooses to attend college, as there are other unobserved factors that influence the decision. An individual would choose to

attend college if the unobserved factors associated with entering the workforce are not large enough to outweigh the observed advantage of attending college.

Equation (4.14) can be estimated using either Maximum Likelihood Estimation (MLE) or generalized Least Squares Estimation (LSE). These methods choose parameter estimates such that the predicted probability of attending college P_{i1} matches the observed market share for decision $j = 1$ in the data. In the estimation process, MLE maximizes the likelihood function, while LSE minimizes the sum of squared residuals, but these two methods are equivalent in the exponential family (Charnes et al., 1976).

4.3 Data

To answer the empirical and policy questions outlined in the introduction, we use data from the National Longitudinal Survey of Youth 1997 (NLSY97). The NLSY97 includes a nationally representative sample of 8,984 American men and women born between 1980 and 1984. Participants were between 12 and 16 years old as of December 31, 1996. Interviews were conducted annually from 1997 to 2011, and biennially thereafter. The survey provides extensive information on respondents' labor market and educational experiences, as well as their family and community backgrounds.

A respondent is considered to have a high school degree if they have either a high school diploma or an equivalent General Educational Development (GED) certificate. There are various types of postsecondary institutions, such as 2-year colleges that offer associate degrees, but this paper only focuses on 4-year college degrees. A respondent is considered to be enrolled in college only if they attend a 4-year college.

The binary choice of whether to attend college is coded as 1 in the year of college enrollment for individuals who eventually earn a four-year college degree. It is coded as 0 in the year of high school graduation for individuals who never enroll in college. For those who obtain a college degree, the decision is assumed to occur in the year they enroll in college. For those who never attend college, the decision is assumed to occur in the year of high school graduation. This distinction is necessary because the timing of high school graduation and college enrollment may differ, as some students take gap years.

Our sample includes 1,247 respondents who earned a four-year college degree and 1,417 respondents whose highest level of education is a high school diploma or equivalent. Since individuals typically graduate from college in four years, the average expected graduation year in our sample is 2008. In this two-period model, period 1 is defined as the first four

years following high school graduation. For those pursuing higher education, this represents the time spent in college and the opportunity cost of not working during this period. Period 2 is defined as the ten years following the expected year of college graduation (approximately 2008–2017), during which individuals are assumed to be in the workforce.

4.4 Estimation

To empirically estimate the discrete choice model specified by the system of equations (4.15), first we need to quantify the key variables and calibrate some parameters using the available data. The variables we need to quantify are: expected labor input in period 1 (L_{i1}), expected labor input in period 2 (L_{i2}), child's learning efficiency (α_i), interest rates (r_i), productivity of basic skills in the labor market (a_0), and the base price of college p_i. The parameters to be obtained through the model estimation are: the discount rate (δ), return to college quality (β), and the productivity of skills attained from higher education (π).

4.4.1 Quantification of Variables

In the 1997 NLSY data, children's learning efficiency, denoted as α_i, is measured by their percentile score on the Armed Forces Qualification Test (AFQT), which combines math and verbal components, taken when they were between 12 and 16 years old. The AFQT score is a better measure of learning efficiency than SAT or ACT scores because the latter are region-specific: students from coastal states are more likely to take the SAT, while those from central states more often take the ACT (The New York Times, 2013). Additionally, the proportion of missing values is significantly higher for the SAT (68%) and ACT (73%) compared to the AFQT (0%) in the sample. Another drawback of using SAT or ACT scores is that they are closely tied to college quality.

A limitation of using the AFQT as a proxy for learning efficiency is that it reflects more than just innate cognitive ability. As noted by Cameron and Heckman (1998), AFQT scores are influenced by a range of socioeconomic factors, such as school quality, parental support, and neighborhood conditions. Therefore, AFQT scores may partially reflect early-life advantages, leading to an overestimation of the impact of ability among higher-income students and an underestimation among their lower-income peers.

The NLS program grouped respondents into three-month age cohorts. The oldest cohort

consisted of individuals born from January to March of 1980, while the youngest cohort included those born from October to December of 1984. This resulted in a total of 20 cohorts, with an average of approximately 350 respondents per cohort. Within each cohort, the NLS computed a cohort-specific percentile score on the AFQT, yielding a final value between 0 and 99. The mean AFQT percentile score for the sample is 50.6%.

The expected work hours for a representative individual are typically 40 hours per week and 48 working weeks per year, which translates to 1,920 working hours in a typical year. However, individuals may differ in their work styles; for example, some may perceive themselves as more hardworking and, as a result, may have higher expected working hours than the average. Assume that decision-makers are aware of their own work style and incorporate this information into their educational decision-making. Accounting for these differences in work styles makes the model more realistic. The NLSY97 dataset includes a survey item that captures a respondent's tendency to work longer hours. The question is phrased as follows: "I do what is required, but rarely anything more." Respondents are asked to choose from a scale of 1 to 7, where "1" indicates "Disagree strongly" and "7" indicates "Agree strongly."

To make an arbitrarily scaled ordinal variable interpretable, we use a directly related outcome that has a well-defined cardinal scale to "anchor" the arbitrarily scaled variable, thereby making the scale interpretable (Cunha et al., 2010). Specifically, we use the respondent's reported annual hours of work to anchor their tendency to spend more time working. First, we calculate the average annual work hours for each group of respondents who report a given value on the work tendency question (e.g., "1"), and we repeat this process for each scale point. We then compute the percentage difference in average reported annual work hours between those who report the median scale value of "4" and those who report other values. This percentage difference serves as an index to quantitatively represent how much more or less hardworking a respondent is relative to the average. We then apply this index as a multiplier to the standard 40 working hours per week to compute an individual's expected work hours for (L_{i1}) and (L_{i2}). It is not a good idea to directly use the reported hours of work as a person's expected labor input, because actual work hours may be endogenously affected by education status and income or substitution effects.

The expected productivity of basic skills in the labor market (a_0) is measured by the federal minimum wage rate in 2009, which was \$7.25 per hour (U.S. Department of Labor, 2009). The prevailing interest rate is represented by the Federal Funds Effective Rate for the corresponding decision-making year, as reported by the FRED database. The average

value of the Federal Funds Rate across the years in the sample is 3%, which is close to the historical average of the real funds rate and the equilibrium interest rate commonly used in the literature (Williams, 2003; Abbott et al., 2019).

The "base" price of higher education is represented by the most affordable tuition rates, those of in-state community colleges. Data on the average in-state tuition for community colleges in each U.S. state are reported by the Education Data Initiative. At this stage, we do not have information on the respondents' states of residence, as this data is only accessible through a special application process. The publicly available data from the NLSY97 only indicate the region where each respondent resides — Northeast, Midwest, South, or West. Therefore, in the current version of this paper, the base price of college is measured using the 2020 average in-state tuition for community colleges in the respondent's designated region. Once state-level data become available, we will use the average in-state tuition for each respondent's state to provide a more precise estimate of the base tuition costs at the time of their decision-making.

4.4.2 Model Estimation and Validation

To empirically estimate the decision-making of school attainment, we match the model to the moments in the data, which is the probability of college enrollment in this case. We first employ a grid search method to pinpoint the best combination of initial parameter guesses that minimizes the sum of squared residuals (SSR). For each parameter, we propose several reasonable initial guesses. The program then calculates the SSR for every possible combination of these initial guesses and returns the combination that minimizes the SSR within 20 iterations. Table 4.1 presents the estimation results based on the preferred initial guess that minimizes the SSR.

The parameters to be estimated are: return to college quality β, discount factor δ, and return to skills acquired from education π. We compare the estimated parameters and the predicted enrollment rate to the values we observe in the data and the assumptions we make in Table 4.1. As shown by the Model (1) of Table 4.1, the predicted enrollment rate of 4-year college computed by the model using the estimated parameters is 48.3%, close to the actual percentage of 46.8% we observe in the NLSY97 data. The estimated value of β is consistent with the assumption that it is between 0 and 1. The estimation of δ also follows our understanding of the discount factor, a value between 0 and 1. The estimated value of π is consistent with the assumption that the return to skills needs to be a finite

positive number.

There are typically three approaches to assess the validity of the model according to Todd and Wolpin (2020). The first traditional way is to examine within sample fit. Using the estimated parameters, we simulate the choices and outcomes of individuals to compare them with the actual choices and outcomes observed in the data. This is the primary approach this paper implements. We match the model predicted outcome to the college enrollment rate in the data. A second method is to check model robustness under different specifications and assumptions. The third approach is to estimate the model on a subsample of the data, and then use the model and estimated parameters to predict the agent's behavior of the holdout sample. To assess the validity of our model, we implement all three approaches and present the estimates in Table 4.1 as follows.

Table 4.1: Estimation Results, Model Validity, and Robustness Checks

	Model (1)	Model (2)	Data	Model (1')	Data (1')
College enrollment rate	48.3%	46.3%	46.8%	46.3%	47.6%
			Assumption		Assumption
Model Parameters					
β	0.83	0.93	(0,1)	0.97	(0,1)
	(0.05)	(0.07)		(0.11)	
δ	0.69	0.68	(0,1)	0.68	(0,1)
	(0.04)	(0.04)		(0.06)	
π	3.07	2.28	(0, C)	3.53	(0, C)
	(0.30)	(0.25)		(0.46)	
Observations	2,518	2,664	2,664	1,314	1,279
SSR	628.00	663.29		326.92	

Note : The table presents the estimated moments and parameter from two different model specifications, Model (1) and (2), compared to the assumptions and the moments in the data. Model (1') shows the estimates using a subsample, compared to the assumptions and observations from the holdout sample. Standard errors are in parentheses except for assumed ranges of parameter values.

Model (1) in Table 4.1 presents the estimation of the primary model discussed earlier in this paper (Equation (4.15)). Model (2) estimates the model parameters and predicts the college enrollment rate under an alternative model specification. In Model (2), instead of using the anchor method to adjust expected labor input by worker type, we assume that all workers have the same expected labor input of 40 hours per week in both periods 1 and 2. We also assess the model's validity using a holdout sample, and the estimation

results are presented as Model (1'). Specifically, we estimate Model (1) using a randomly selected subsample comprising roughly half of the primary sample, then use the estimated parameters to predict the college enrollment rate and compare it to the holdout sample (presented as Data (1')).

The estimated parameters from Models (1), (2), and (1') are consistent with the model assumptions. The predicted moments from the three approaches are also similar to those observed in the data. Overall, the model produces consistent and accurate estimates across the three approaches, closely matching the moments observed in the data.

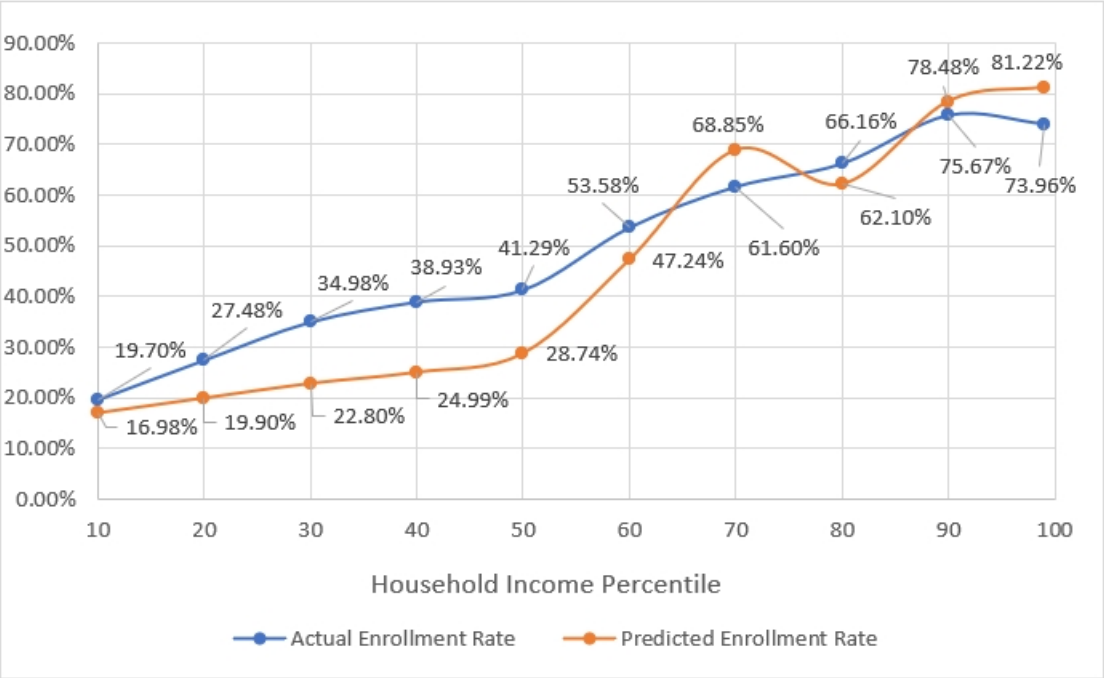
4.5 Policy Simulations and Implications

After estimating the model parameters, we use the model to simulate college enrollment rates among children from households at different income levels and compare these simulated rates to the actual enrollment rates by income. We then examine how variation in key variables affects the probability of college attendance and the endogenously chosen college quality. Specifically, we simulate the probability of college enrollment and the level of college quality by varying one key variable at a time while holding the others constant at their mean values. Lastly and most importantly, we simulate and evaluate the impacts of different policy alternatives aimed to increase college enrollment and improve the quality of colleges that students choose to attend. This analysis yields important policy implications regarding which strategies are most effective for increasing both college attendance and the quality of college selected by students.

4.5.1 College Enrollment

To study the disparity of education enrollment among households of different incomes, we investigate the 4-year university enrollment rate (among high school graduates) by the income percentile of their childhood households. In Figure 4.2, we also plot the model's predicted enrollment rate by income percentile to examine how much the model fits the data. The x-axis shows the income percentile at increments of 10%. The orange curve plots the enrollment rates predicted by the model at every 10th percentile, compared to the actual enrollment rates at every income percentile in blue.

Figure 4.2: College Enrollment Rate by Household Income

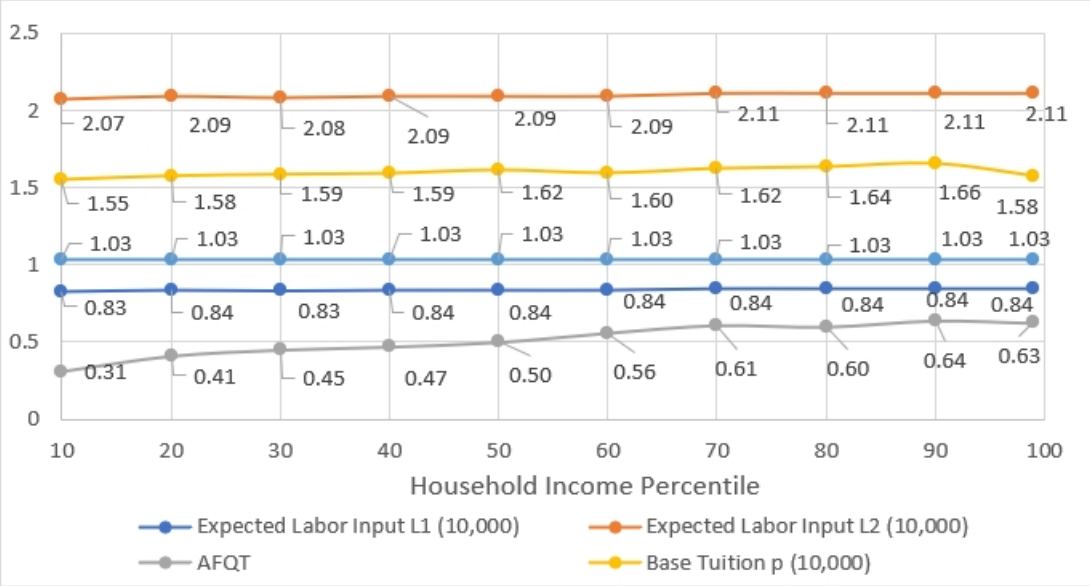


Note: The figure shows 4-year college enrollment rate (among high school graduates) by the income percentile of the households. The x-axis shows income percentile at an increment of 10%. The orange line shows the enrollment rates predicted by the model at every 10th percentile, compared to the true enrollment rates at every percentile in the data plotted in blue.

The predicted enrollment rates by family income match the overall pattern of the actual enrollment rates in the data. The model predicts a flatter increase in the enrollment rates at the lower-income family compared to the data and catches up at the higher-income households. The ups and downs in the predicted enrollment rate primarily come from variations in the AFQT score across income levels, as shown in Figure 4.3. Other key variables that influence college attendance do not vary as much as the AFQT score. The slight variation in base tuition is due to limitations in the geographical data, which may not accurately reflect the true patterns in the actual data. The predicted college attendance rate increases substantially between the 50th and the 60th income percentile, mostly due to a jump in the cohort-specific average AFQT score from 0.50 to 0.56. The unexpected drop in the enrollment rate at the 80th percentile is because the average AFQT score at this income percentile is lower than that in the previous income level. Admittedly, in addition to the observables, there are certainly other unobserved variables that affect the enrollment rate, such as family connections and discrimination, that are unobservable in the data and cause the differences between the predicted and the true enrollment rates. The model specification

is generalized in many aspects and does not perfectly reflect the real-world behavior, which is another reason that the predicted enrollment rates do not perfectly match the data. We further discuss the limitations of our model in Subsection 4.6.3.

Figure 4.3: Key Characteristics by Household Income Percentile



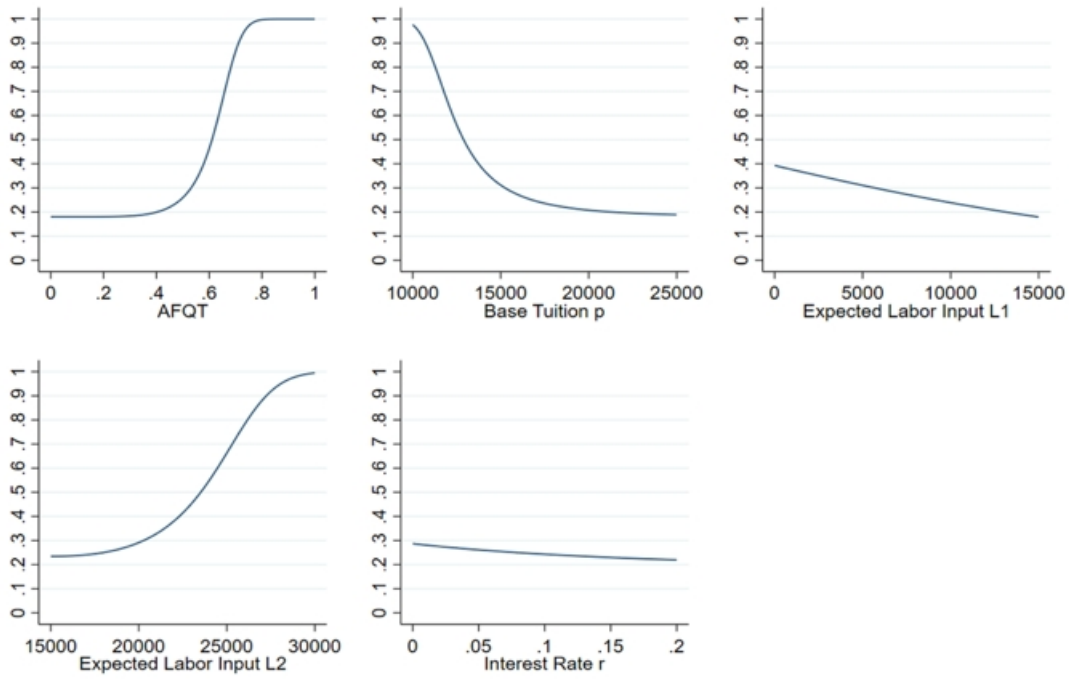
Note: The dots represent the mean value of a variable at a given household income percentile in the data.

According to the model, the key factors influencing college access are the children's learning efficiency (α_i) measured by cohort-specific AFQT percentile, the expected total labor input in period 2 (L_{i2}), the expected total labor input in period 1 (L_{i1}), the interest rate (r_i), and the base price of college (r_i)) measured by the average tuition of community colleges in the region of residence. By plotting these variables at every 10th percentile of family income, we examine whether disparities in college attendance can be explained by variations in these key characteristics.

As shown in Figure 4.3, the cohort-specific AFQT percentile increases as the family income rises. This is a very important contributor to a higher college enrollment rate. The "base" price of college, measured by average tuition cost of community college, (p_i) and expected labor input in period 2 (L_{i2}) rise slightly as the household income, but mostly flat. Interest rate (R_i), which is $(1 + r_i)$, and the expected total labor input in period 1 (L_{i1}) remain constant as the family becomes wealthier. This stability in interest rates is expected, as the timing of decision-making is exogenous to household income.

To illustrate how the predicted enrollment rate changes given the variation of every key variable, we simulate the probability of college enrollment based on the change of a key variable while holding the values of other key variables constant at their mean values. The variation of a variable ranges from the sample minimum to the sample maximum, except for the interest rate where we arbitrarily take 20% as the maximum. Figure 4.4 shows the simulation results. Consistent with the construction of the model, the enrollment rate monotonically increases given a higher value of AFQT percentile and a longer expected hours of work in period 2. The enrollment rate monotonically reduces as the expected labor input in period 1 (L_1), the base price of tuition (p), and as the interest rate (r) rises. Another important finding is that the AFQT, base price of tuition (p), and expected labor input (L_2) have a much larger impact on college attendance compared to that of L_1 and r.

Figure 4.4: Impact of Key Variables on College Enrollment Rate



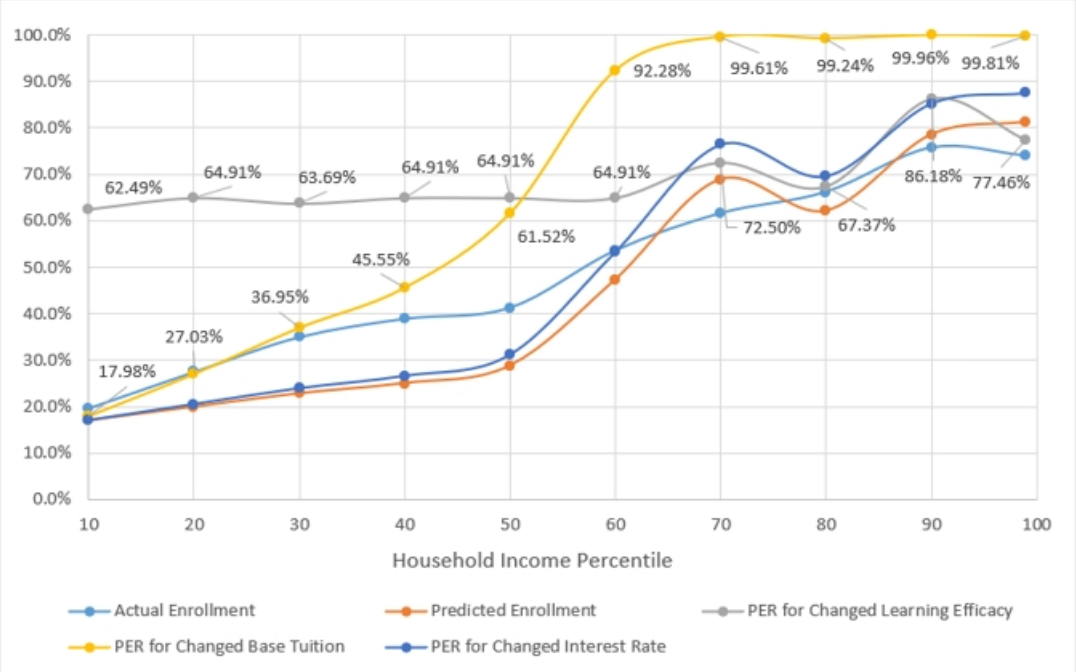
Note: The figures show the simulated probability of college enrollment rate based on the variation of a key variable while holding the values of other key variables constant at their mean values. The y-axis is the simulated probability of college enrollment rate from 0% to 100%.

Combining the findings from Figure 4.3 and 4.4, our model suggests that even in a setting where all students rely on loans to pay for tuition, substantial disparities in college enrollment persist across household income levels. Children from wealthier families are

more likely to attend college, largely because they have higher AFQT scores. These findings motivate an evaluation of three policy alternatives aimed at increasing college attendance. The first is to invest in primary and secondary education to improve students' pre-college cognitive skills. The second is to reduce the cost of college, making college education more accessible. The third is to lower the interest rate on student loans to ease the financial burden of borrowing.

We simulate college attendance rates under three policy scenarios, holding all other characteristics constant at the mean value for each income percentile. First, we simulate an increase in AFQT scores to the 75th percentile for students currently below that level, holding all else constant. Second, we lower the base price of college to the 25th percentile of the observed tuition distribution for all students. This models the effect of uniformly reducing tuition costs. Third, we set the student loan interest rate to zero for all individuals. Figure 4.5 presents the simulated enrollment rates under each policy, alongside the actual and baseline predicted enrollment rates.

Figure 4.5: Predicted Enrollment Rate (PER) for Changed Characteristics



Note: We simulate the college attendance rate, when increasing the learning efficiency to the upper quartile for those whose are lower than that, given all other characteristics fixed at the mean value of the income percentile. We repeat the same process when reducing the base price of schooling to the lower quartile all students and setting interest rate to zero for all students.

Increasing all students' learning efficiency to the current upper quartile significantly raises the probability of college attendance, especially among lower- and middle-income families, even when all other factors are held constant. In contrast, lowering the student loan interest rate has a very small effect on college enrollment. Comparing the predicted enrollment under a zero-interest scenario (shown in dark blue) with the original predicted enrollment (in orange), the impact appears slightly larger for higher-income students than for their lower-income peers.

Surprisingly, lowering the base tuition cost does not help students from the lower-income families as much as those from the more affluent families, holding all else constant. Comparing the original predicted enrollment in orange and the predicted enrollment in yellow for lower base tuition, the enrollment rate is promoted by a larger amount as the income level becomes higher. After lowering the base tuition, the predicted enrollment rate of students at the 60th percentile increases to over 90%, while children at the 30th percentile still have an enrollment rate lower than 40%. One possible reason is that students from wealthier families are more likely to respond to tuition cuts by enrolling in more or higher-quality programs, while lower-income students may still face academic, informational, or logistical hurdles that keep them from college enrollment even if tuition drops. However, students from more resourceful family backgrounds tend to receive better education during their childhood, and so a lower base tuition promotes their college access by a greater margin, because a higher learning ability makes them more capable of benefiting from the value of higher education, even if they also have to borrow student loans. Overall, our findings suggest that enhancing children's internal human capital and learning ability from an early age may be more effective in promoting college access than simply making college more affordable, especially given constraints on government resources.

It is important to note that the AFQT scores, which are used as a proxy for learning ability in our model, are correlated with a wide range of other factors such as school quality, parental support, and access to enrichment opportunities (Cameron and Heckman, 1998). Therefore, the significance of AFQT scores in our results may partly reflect the influence of these unobserved variables, rather than innate ability alone.

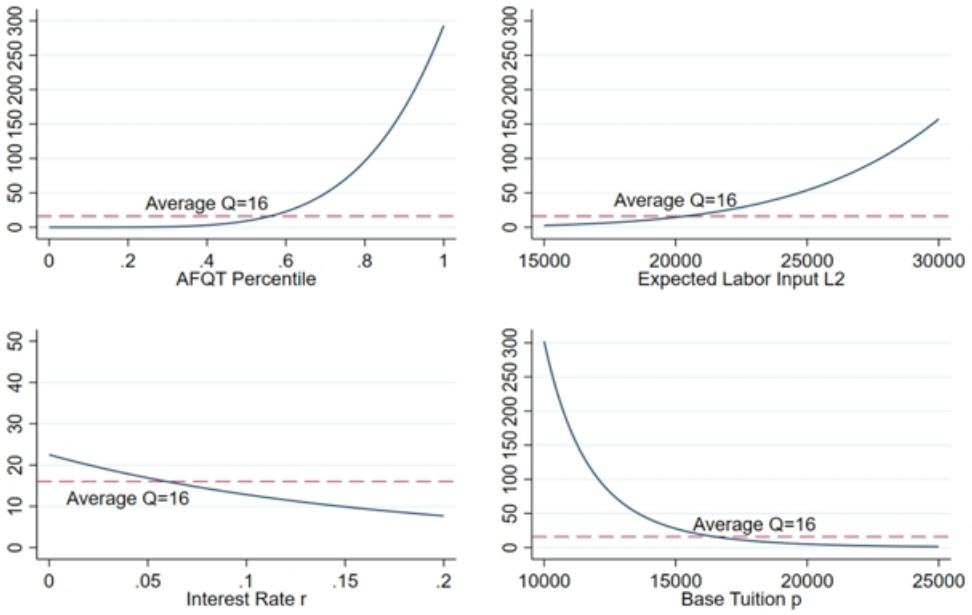
Two clarifications need to be made. First, the learning efficiency is measured by the cohort-specific AFQT percentile, and in practice, we cannot increase everyone's score percentile. However, we can think of the AFQT percentile as an index for knowledge or learning abilities. Instead of being a competition, every person can receive an "A" grade as long as they have "most of the answers correct".

Second, the implicit assumption that increasing students' human capital at the population level increases college attendance is that current universities are able to accept more incoming students, or new universities will be constructed to match the demand. As of 2017, 84.8% of the U.S. colleges admit at least 50% of the applicants, and 57.4% admit at least 70% of the applicants (DeSilver, 2019). From 2002 to 2017, more selective universities (above median) have been less able to keep pace with the soaring number of applicants, but the less selective ones (below median) have been sufficiently expanding to keep up with the climbing figure (DeSilver, 2019). The policy alternative we evaluated focused on investing in pre-college education for disadvantaged students, while holding the skill levels of more advantaged students constant. Therefore, it is reasonable to assume that the supply of higher education, particularly at less selective institutions, will adjust to meet the increased demand. Additionally, we acknowledge that providing high-quality pre-college education to all children and boosting their human capital to the 75th percentile of current levels is a long-term process.

4.5.2 Endogenous College Quality

In this section, we explore the impact of the key variables on the quality of the college that students choose and offer policy recommendations to effectively close the gap and reduce inequality in the selection of college quality. Assume that the agent knows their learning efficiency (α), return to college quality (β), return to skills acquired from higher education (π), their expected life-time hours of work after college graduation (L_2), interest rate of student loans, and the base tuition cost. Equation (4.10) suggests that, if students decide to attend college, there exists an optimal value of college quality (Q) that they will endogenously select based on their personal characteristics and the exogenous factors, including the interest rate and tuition cost.

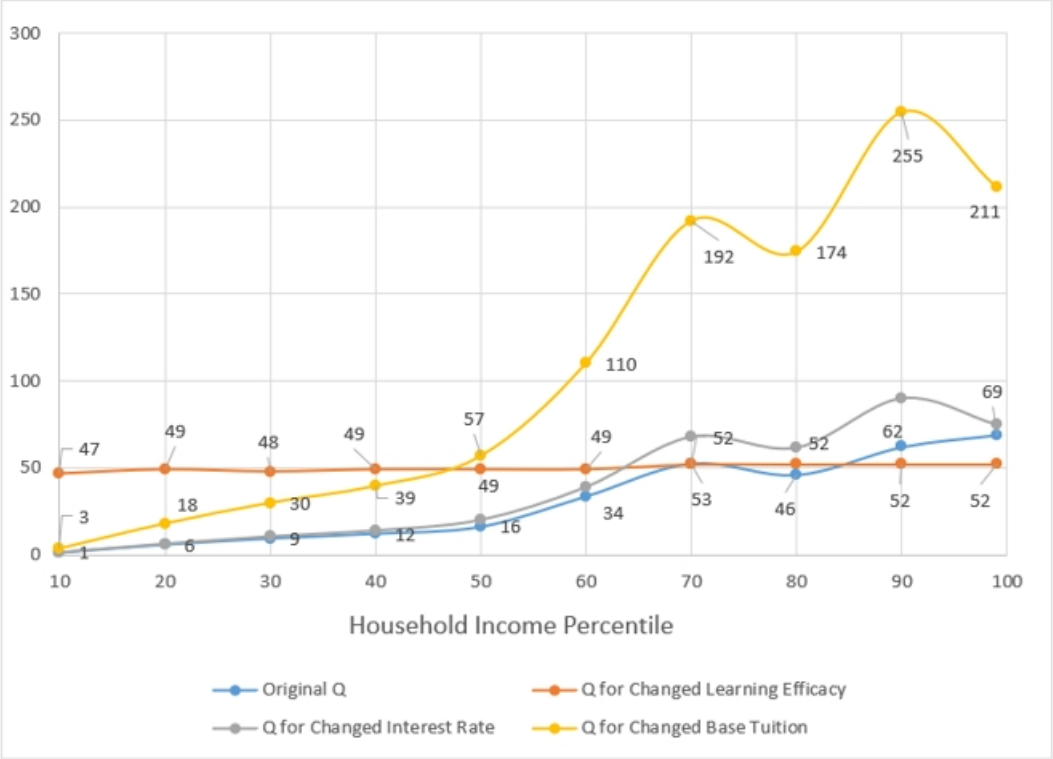
Figure 4.6: Impact of Key Variables on Endogenously Chosen School Quality



Note: These figures show the impact of key variables on the endogenously chosen school quality by students. Each figure is plotted against the changing of a key variable while holding the rest constant at the sample mean. The scale of the key variable ranges from the sample min to hypothetical max.

First, we simulate how changes in a key variable affect the chosen college quality, while holding the other variables constant at their sample mean. The simulation results are presented in Figure 4.6. Consistent with the implications of Equation (4.10), students endogenously choose higher-quality colleges when they have higher learning efficiency (α) and a greater expected labor input after graduation (L_2). Additionally, students opt for lower-quality colleges when the interest rate or base tuition cost is higher. In terms of the magnitude, learning efficiency (α), as measured by the AFQT, and the base price of tuition (p), measured by average community college tuition in the residential region, have a sizable impact on college quality. The expected hours of work in period 2 (L_2) also has a large effect, whereas college quality remains relatively unchanged as the interest rate on student loans increases from 0% to 20%. Next, we calculate and plot the mean endogenous college quality for each family income percentile, based on the mean observed characteristics, as shown in Figure 4.7.

Figure 4.7: School Quality (Q) for Changed Characteristics



Note: We simulate the school quality, when increasing the learning efficiency to the upper quartile for those whose are lower than that, given all other characteristics fixed at the mean value of the income percentile. We repeat the same process when reducing the base price of schooling to the lower quartile all students and setting interest rate to zero for all students.

The overall pattern, shown by the blue curve in Figure 4.7, indicates that the college quality students choose increases with family income. This rise is gradual in the lower income percentiles due to only modest gains in AFQT scores. A sharper increase in endogenous college quality begins around the 60th percentile, primarily because, as shown in Figure 4.3, the cohort-specific average AFQT score rises from 0.50 to 0.56. Interestingly, at the 80th percentile, the mean AFQT score is lower than that at the 70th percentile, resulting in a decrease in estimated college quality. Clearly, the AFQT score is the primary driver of disparities in endogenously selected college quality, as other key variables show little variation across income percentiles.

Similar to the policy simulation in Section 4.5.1, we first simulate an increase in AFQT scores to the 75th percentile for students currently below that level, while holding all other characteristics constant at the mean value for each income percentile. We repeat this process to calculate and plot the college quality at each income percentile under two additional

scenarios: reducing the base tuition cost to the lower quartile and setting the interest rate to zero for all individuals, again holding all else constant. As shown in Figure 4.7, improving adolescents' learning skills and knowledge levels proves to be a more effective strategy for narrowing the gap in endogenously chosen college quality, compared to lowering base tuition or interest rates. When the learning efficiency of disadvantaged students is raised to the upper quartile of the current distribution, the disparity in college quality between the 50th and 90th income percentiles is substantially reduced, closing a gap that previously reflected a fourfold difference between students from the wealthiest and middle-income families. Original values for college quality are 16 for the 50th income percentile and 69 for the 90th income percentile, as shown in the gray line. After disadvantaged students' AFQT scores is increased to the 75th percentile, the values for college quality are 49 for the 50th income percentile and 52 for the 90th income percentile, as shown in the orange line.

Lowering the base tuition to the current lower quartile substantially facilitates the overall endogenous college quality, but two issues come with it. First, the gap between higher- and lower-income percentiles widens compared to the original college quality curve. This mirrors the earlier finding that reducing tuition is less effective than improving learning efficiency. When students from lower-income families have lower learning abilities, cheaper tuition alone does little to improve their outcomes. In contrast, students from wealthier backgrounds, who generally receive better pre-college education, can better leverage reduced tuition, as their stronger learning skills allow them to realize greater returns from higher education, even when financed by loans. Second, although college quality rises notably for students in higher income brackets, a more practical concern is whether such high-quality institutions are available to accommodate this shift. Therefore, within the model's framework, enhancing human capital through increased investment in pre-college education is more effective than reducing tuition in narrowing disparities in endogenously chosen college quality.

It is important to note that the AFQT score, which we use as a proxy for learning ability in our model, is closely correlated with a variety of other factors, including school quality, parental involvement, and access to enrichment opportunities (Cameron and Heckman, 1998). As a result, the strong predictive power of AFQT scores in our findings may not solely capture innate learning ability, but also the cumulative effect of broader environmental and socioeconomic advantages.

4.5.3 Policy Implications

Suppose a local government has funding available to promote college attendance through several approaches: investing in pre-college education, lowering or subsidizing the interest rates on student loans, or providing funds to universities to reduce in-state tuition costs. The government must decide which approach is most cost-effective in increasing four-year college attendance. In this section, we focus on evaluating the effectiveness of these policy alternatives.

To compare and evaluate the effectiveness of various policy plans, we simulate the college attendance rate at each family income percentile by adjusting key variables. Specifically, we increase variables that facilitate college attendance to the upper quartile and decrease variables that hinder attendance to the lower quartile for every individual, while holding all other characteristics constant at the mean value for that income percentile.

The simulation results presented in Figure 4.5 show that improving the learning efficiency of all students to the upper quartile, while holding all other characteristics constant, significantly increases college attendance rates among children in the lower income percentiles. Enhancing students' learning ability or knowledge at an early age proves to be a more effective strategy for promoting college access than other policies evaluated, such as reducing base tuition costs or lowering interest rates. Contrary to common assumptions, lowering base tuition does not benefit students from lower-income families as much as it benefits those from more affluent backgrounds, when all other characteristics are held constant. This is primarily because a reduction in base tuition is marginally more beneficial to students who are more capable of taking full advantage of the value a university education offers, typically those with stronger learning abilities and better pre-college preparation.

The findings support a consistent core insight and policy implication for addressing the gap in endogenously determined college quality. Simulation results suggest that enhancing students' pre-college learning abilities is a more effective strategy than other interventions for reducing inequality in university quality. While reducing the base price of tuition significantly raises the overall quality of colleges selected, it also widens the gap between students from lower- and higher-income families. Moreover, despite a seemingly substantial increase in overall endogenous college quality, it remains difficult to determine which colleges genuinely meet that standard.

There are multiple approaches to reducing disparities in higher education, and ideally, one might wish to implement all policies that promote college access for disadvantaged families. Nonetheless, in reality, both society and governments face strict budget constraints,

and funding is a scarce resource. Therefore, it is essential to identify cost-effective strategies for expanding access to higher education and narrowing the gap in endogenously selected college quality. We acknowledge that our policy simulations only emphasize the effectiveness of policy alternatives, and we have not identified the policies that have the biggest effect per dollar spent. This is a limitation in the policy simulations and implications that we make in Section 4.5. Moreover, we do not account for market failures in the model, which presents an additional limitation for interpreting the results and drawing policy implications in this paper. As market failures are not considered in this paper, the results and policy implications may be the upper bound of the actual effect in reality.

4.6 Discussion

4.6.1 Robustness Check on Other Model Specification

One concern with our model is the assumption that students do not rely on parental transfers to pay for college tuition. However, the data show that some students do receive financial support from their parents. To address this, we conduct a robustness check in which we relax this assumption and allow a portion of students' tuition to be funded through parental transfers.

In the new model, the lifetime expected utility of an agent who chooses to attend college can be written as:

$$V_{i,j=1} = y_i + \delta\alpha_i\pi Q_i^{\beta} L_{i2} - \delta(1+r_i)p_i Q_i(1-E_i). \tag{4.16}$$

where E_i is the proportion of individual i's tuition paid by their family. The proportion of individual i's tuition paid by their family is an exogenous variable that we find in the NLSY97 data. It is worth noting that E_i can be endogenous to school quality and other student characteristics. We will consider working on this in the future.

Then, the optimal value of college quality, Q_i^*, derived from the first order condition that maximizes i's expected lifetime utility from attending college ($V_{i,j=1}$) is:

$$Q_i^* = \left(\frac{\beta\alpha_i\pi L_{i2}}{(1+r_i)p_i(1-E_i)}\right)^{\frac{1}{1-\beta}}. \tag{4.17}$$

Following what we do in Section 4.2, the probability of attending college, P_{i1}, can be

expressed as:

$$\frac{1}{1+exp\left(a_0L_{i1}+\delta a_0L_{i2}-\delta\alpha_i\pi L_{i2}\left(\frac{\beta\alpha_i\pi L_{i2}}{(1+r_i)p_i(1-E_i)}\right)^{\frac{\beta}{1-\beta}}+\delta(1+r_i)p_i(1-E_i)\left(\frac{\beta\alpha_i\pi L_{i2}}{(1+r_i)p_i(1-E_i)}\right)^{\frac{1}{1-\beta}}\right)} \tag{4.18}$$

Model (3) of Table 4.2 presents the estimated parameters of the new model, which allows for the possibility that some students have their tuition partially or fully covered by their families (Equation (4.18)). Model (4) of Table 4.2 also estimates the parameters of Equation (4.18). However, instead of adjusting for worker type, Model (4) uses a fixed measure of expected labor input, 40 hours per week, in periods 1 and 2. For comparison, Model (1) reports the estimates of the original model (Equation (4.14)) as discussed in Section 4.2. Model (2) also estimates the original model, but uses a fixed 40-hour workweek to measure expected labor input. The estimated parameters and predicted college enrollment rates remain consistent across all model specifications, suggesting that allowing a portion of tuition to be paid by families does not significantly alter the model's predictions.

Table 4.2: Estimation Results Considering Percent of Tuition Paid by Family

	Model (1)	Model (2)	Model (3)	Model (4)	Data
College enrollment rate	48.3%	46.3%	46.9%	47.2%	46.8%
					Assumption
Model Parameters					
β	0.83	0.93	0.95	0.96	(0,1)
	(0.05)	(0.07)	(0.07)	(0.09)	
δ	0.69	0.68	0.69	0.68	(0,1)
	(0.04)	(0.04)	(0.03)	(0.03)	
π	3.07	2.28	2.64	2.55	(0, C)
	(0.30)	(0.25)	(0.29)	(0.24)	
Observations	2,518	2,664	2,518	2,664	2,664
SSR	628.00	663.29	627.44	663.29	

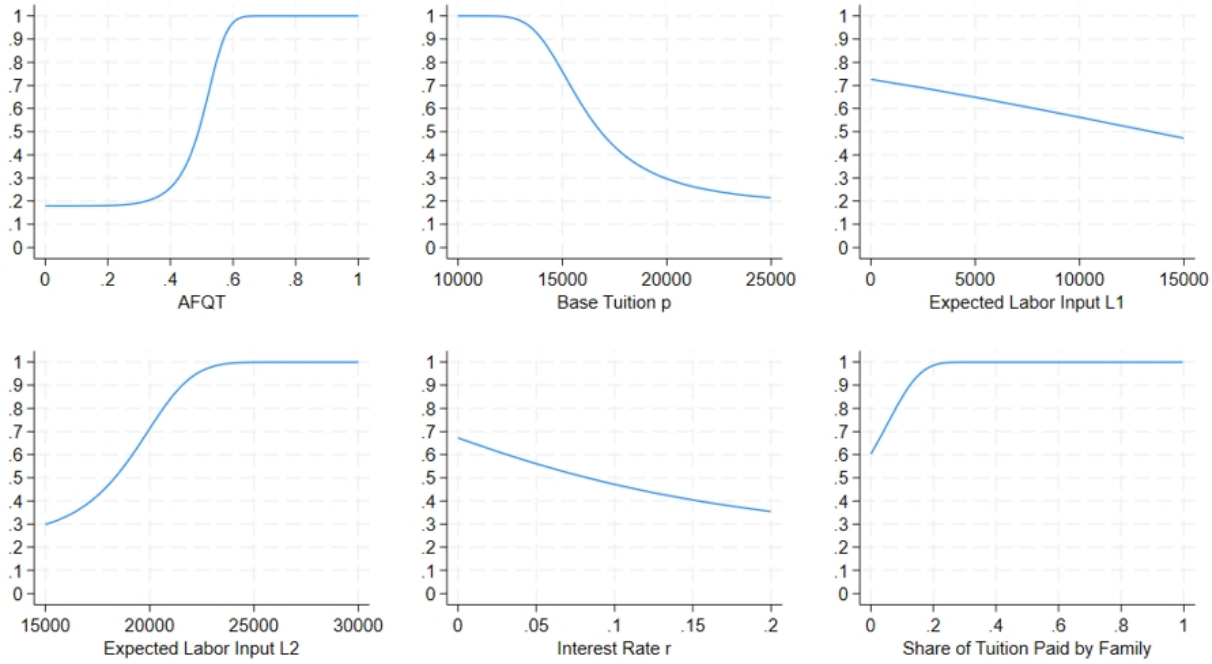
Note : The table presents and compares the estimated moments and parameter from the main specification and the specification that accounts for family tuition payment. Model (1) and (2) are the same as (1) and (2) in Table 1. They apply the main specification under different setting. Model (3) and (4) take the percentage of tuition paid by family into consideration. In Model (2) and (4), we use the typical 40 hours of work as the measure to calculate the expected labor input in period 1 and period 2 without adjusting it by the type of worker. Standard errors are in parentheses except for assumed ranges of parameter values.

Using the new model that incorporates parental transfers, we examine how variation in key factors influences the likelihood of college attendance. We then simulate college

enrollment rates for children from households across different income levels and compare these simulated rates to actual enrollment data by household income.

Figure 4.8 presents simulation results illustrating how the college enrollment rate responds to changes in a single key variable, while all other variables are held constant at their mean values. Each variable is varied from its sample minimum to maximum, except for the interest rate, which is capped at an arbitrarily chosen maximum of 20%. Consistent with the model setup, the enrollment rate increases as AFQT and expected labor input in period 2 (L_2) increases. Conversely, it decreases with higher expected labor input in period 1 (L_1), higher base tuition (p), and higher interest rate (r). Notably, AFQT scores, the base price of tuition, and expected labor in period 2 (L_2) exert a significantly stronger influence on college attendance decisions compared to L_1 and r. This pattern aligns with the simulation results from the original model (Figure 4.4).

Figure 4.8: Impact of Key Variables on Enrollment Rate – Considering Tuition Payment by Family



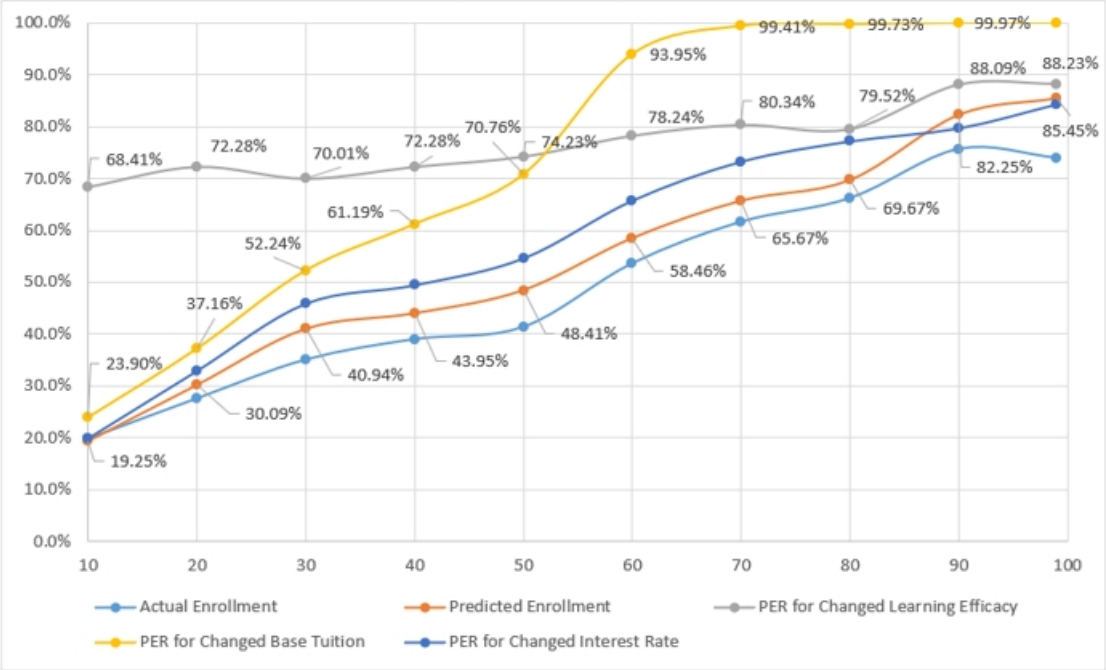
Note: The figures show the simulated probability of college enrollment rate based on the variation of a key variable while holding the values of other key variables constant at their mean values, including the percentage of tuition payment by children's family. The y-axis is the simulated probability of college enrollment rate from 0% to 100%.

We also simulate the effects of three policy alternatives aimed at increasing college attendance under the new model specification. The first policy is investing in primary or secondary education to improve children's cognitive skills before college. The second focuses on reducing the cost of college to encourage more students to take out loans and enroll.

The third explores the impact of lowering the interest rate on student loans. The simulation results are presented in Figure 4.9.

The orange line in Figure 4.9 represents the predicted college enrollment rate by income percentile, while the light blue line shows the actual enrollment rate. As shown in the figure, the predicted values follow a similar pattern to the true enrollment rates, with the only noticeable difference being that the predicted rates are slightly higher.

Figure 4.9: Predicted Enrollment Rate (PER) for Changed Characteristics – Considering Tuition Payment by Family



Note: Under the new model, we simulate the college attendance rate, when increasing the learning efficiency to the upper quartile for those whose are lower than that, given all other characteristics fixed at the mean value of the income percentile. We repeat the same process when reducing the base price of schooling to the lower quartile all students and setting interest rate to zero for all students. This simulation is different than that in Figure 4, as we consider the percentage of tuition payment by children's families at each family income percentile.

In Figure 4.9, the yellow line represents the predicted college attendance rate when the base tuition price for all students is uniformly reduced to the 25th percentile value observed in the data, while holding all other characteristics fixed at the mean income percentile. This simulates the potential impact on enrollment rates of making college tuition more affordable across the board. The grey line shows the simulated enrollment rate when students with AFQT scores below the 75th percentile are boosted to that level, with all other factors held

constant. Additionally, the dark blue line illustrates the effect of setting the interest rate to zero for all individuals, again holding all other variables constant.

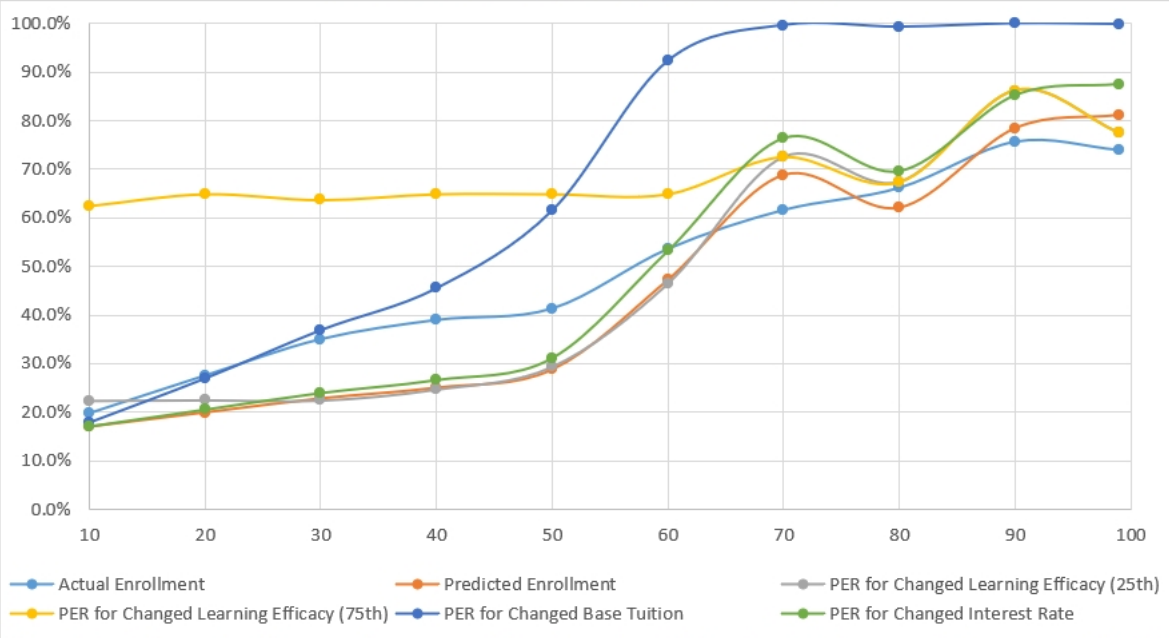
The simulation results presented in Figure 4.9 closely mirror those from the original model shown in Figure 4.5. Increasing all students' AFQT scores to the level of the current upper quartile, while holding other characteristics constant, significantly increases college attendance rates among lower- and middle-income families. In contrast, reducing the interest rate on student loans has only a minimal effect on college attendance. Similarly, lowering base tuition costs benefits students from higher-income families more than those from lower-income backgrounds, assuming all else remains equal.

In sum, our model remains robust across different specifications. Allowing a portion of students' tuition to be funded through parental transfers has minimal impact on the model's predictions and policy simulation results.

4.6.2 Robustness Check on Policy Experiment

Admittedly, it is unrealistic to raise every student's AFQT score to the 75th percentile for students below that level, as we perform various policy experiments in Figure 4.5. Therefore, in this subsection, we simulate college attendance rates under a policy scenario, where we are only able to raise every student's AFQT score to the 25th percentile for students currently below this level, holding all other characteristics constant at the mean value for each income percentile. We present the simulated college enrollment rate in Figure 4.10, alongside the three previous policy scenarios and the actual and baseline predicted enrollment rates.

Figure 4.10: Predicted Enrollment Rate (PER) for Changed Characteristics (Including Increasing Learning Efficicency to the Lower Quartile)



Note: We simulate the college attandance rate, when increasing the learning efficicency to the lower quartile (25th percentile) for those whose are lower than that, given all other characteristics fixed at the mean value of the income percentile. We also simulate the college attandance rate, when increasing the learning efficicency to the upper quartile for those whose are lower than that, given all other characteristics fixed at the mean value of the income percentile. We repeat the same process when reducing the base price of schooling to the lower quartile all students, and setting interest rate to zero for all students. This simulation is different than that in Figure 4, as we take into account of the percentage of tuition payment by children's family at each family income percentile.

Increasing all students' learning efficiency (AFQT score) to the 25th percentile raises the probability of college attendance for students from families in the 10th and 20th income percentile by a small magnitude, while the probability of college attendance for students from the 30th income percentile and above remains unchanged. This result is expected, as we only raise students' learning efficiency to the 25th percentile for those who are below this level, while leaving the learning efficiency of all other students unchanged. However, this simulation result provides us with a more realistic perspective of the policy experiment of raising students' learning efficiency. A more realistic case is that we are only able to raise the learning efficiency for students that are well-below the average, while raising the learning efficiency for all students to the previously above-average level is not as realistic. The result in this subsection suggests that the implications from the previous policy experiment reflect the upper bound of the actual effect. In reality, the actual effect of providing higher-quality early education more equally can be much smaller than the simulated results in Section 4.5.

4.6.3 Further Discussion on Limitations

The first limitation of the model is that it is a two-period framework in which the net present value of lifetime income is not specified as rigorously as in many other studies that employ more detailed multi-period models. Due to data limitations, specifically the structure of the NLSY97 dataset, we are only able to track approximately ten years of a respondent's career following college or university graduation. As a result, our model defines lifetime income as the earnings accumulated from the age of high school graduation through the first ten years following college or university graduation, applying a single discount factor to this entire period. In reality, students with an undergraduate degree stay in the workforce for more than ten years after graduation, allowing them to benefit from their college education for a longer period than model specified. The weight (ten years) that we place on the second period in our model is thus an underestimation of the reality. This can lead to an underestimated college enrollment rate, as the expected return to college education is likely higher in reality. Moreover, the inability to account for longer-term earnings may also cause us to underestimate the effectiveness of policies aimed at reducing the cost of college in our simulations. Because the model undervalues the long-term financial benefits of a college degree, it may not fully capture how strongly students would respond to cost-reducing policies in the real world, where actual lifetime returns are substantially higher.

The second limitation is that our model serves to describe an economy where higher education is not funded by parents and hence not used as a method of income transfer from period 1 to period 2. However, in the U.S., access to student loans, especially federal loans, depends heavily on parental financial resources, as reported through the FAFSA process. In practice, loan eligibility and amounts are negatively correlated with parental income and assets. Students from wealthier families generally qualify for smaller need-based loans or none at all. In reality, students from wealthier families may rely more on direct parental support rather than loans, while students from lower-income families may face stricter borrowing constraints despite being more loan-eligible, due to concerns about debt burden or limits on maximum loan amounts. Our model currently assumes that all students have equal access to borrowing, independent of parental background. This simplification may bias our results, particularly by underestimating the financial barriers faced by lower-income students and overestimating the financial flexibility of higher-income students. Given that access to student loans is limited in reality, our policy simulation may overestimate the effectiveness of policies aimed at increasing college enrollment by reducing tuition costs or lowering interest rates associated with student loans. In our simulations, these two policy

alternatives are not very effective. Thus, the "true" estimate is likely to be smaller.

The third limitation of our model is that we treat the productivity of high school graduates in the labor market (a_0) and the productivity of college graduates in the labor market (a_1) as constant across individuals. This assumption overlooks the reality that the return to education evolves with experience, and the rate of growth may differ between high school graduates and college graduates. Acemoglu and Autor (2011) show that individuals with a higher level of education tend to experience faster wage growth over time due to greater opportunities for career advancement, specialization, and skill accumulation. Our model does not capture the dynamic nature of wage progression, which could lead to inaccurate predictions regarding the expected return to education over time, potentially underestimating the benefits of obtaining a college degree. As a result, policy simulations based on this model may also underestimate the potential impact of interventions that reduce college costs, since the long-term financial advantages of higher education are not fully captured.

The fourth limitation of our model is that it considers only 4-year colleges and universities as the primary pathway for higher education, excluding options such as community colleges. In reality, community colleges are an important and more affordable alternative for many students, particularly those from lower-income families. By not modeling community college attendance, our analysis may overestimate the barriers to higher education and underestimate overall enrollment rates. This omission could also influence our policy simulations, potentially overestimating the impact of investment in pre-college education (measured by increased AFQT scores), because community colleges typically have lower entrance requirements, and students may be able to enroll without achieving the high AFQT scores necessary for four-year colleges.

Another limitation of our model is the use of AFQT scores as a measure of students' learning efficiency. While AFQT scores provide a convenient proxy for cognitive ability, they are also correlated with a wide range of other factors, such as school quality, parental support, neighborhood environment, and access to enrichment opportunities (Cameron and Heckman, 1998). As a result, the AFQT score may not isolate pure learning ability but instead reflect a broader set of socioeconomic advantages or disadvantages, which could affect our model's interpretation of the relationship between learning ability and educational outcomes. Specifically, the strong predictive power of AFQT scores in the model may partly capture the influence of these correlated factors, rather than innate learning efficiency alone. Since the AFQT score reflects not only students' learning efficiency but also

parental investments in early education, it tends to underestimate the learning efficiency of lower-income students relative to their higher-income counterparts. With this caveat in mind, policies that increase investment in pre-college education may be even more effective at improving learning outcomes and college attainment for lower-income students than our model's simulation results suggest.

Furthermore, our model has a key limitation in that it omits a number of factors that can significantly influence higher education decisions. While our findings suggest that the primary determinant of differences in college access and quality is the amount of human capital accumulated in earlier educational stages, it is important to recognize that this represents only one aspect of a complex decision-making process. For instance, Chetty et al. (2011) demonstrate that, even when controlling for test scores, students from lower-income families are still less likely to attend selective colleges. Additional factors, such as social networks, family background, access to information about college opportunities, and psychological barriers, may also play a crucial role in shaping educational outcomes, but they are not captured in our model. Therefore, our analysis may underestimate the broader socioeconomic challenges that affect students' higher education choices and could lead to an incomplete understanding of the drivers of college access and success.

4.6.4 The Return to Education

Another way to verify the usefulness of the model in this paper is to relate the estimated difference in life-time utility from Equation (4.9) to the return to college in the literature, as the estimated difference in life-time utility, ΔV_i, is expressed in dollar value. First, I calculate the premium of attending college in dollar value using Equation (4.9) for those who attend college, by entering the average values of the variables and the estimated parameters into the equation. The value of the total premium of attending college is \$123,477.46. I divide this value by 14 years (period 1 plus period 2) and then divide the result by 4 years of college to obtain the annual premium (in dollar value) of one additional year of college education, which yields \$2,204.95. In the sample, the average adult income per year is \$25,987.51.[2] This provides that the premium of one additional year of college education, using this model, is 8.5% (\$2,204.95/\$25,987.51).

The economics literature consistently has estimated that, on average, one additional year of college increases earnings by about 5% to 10% (Card, 1999; Heckman et al., 2006).

[2]I use the income data for persons who are 23–28 years old to compute the average annual adult income, following the notion of Cunha and Heckman (2008).

The fact that the estimated return to education using the model in this paper (8.5%) is consistent with the finding in the literature suggests that our model, to some extent, reflects the actual decision-making process of students' college attendance.

4.7 Conclusion

How to effectively increase college and university access, especially for low-income students, has been a long-standing and ongoing debate (Fack and Grenet, 2015; Page and Scott-Clayton, 2016). To contribute to this discussion, particularly at a time when children have become increasingly independent of parental financial assistance for college and university and when student loans play an increasingly central role in decision-making, we construct a two-period discrete choice model to investigate decisions regarding higher education acquisition. The model features children's independent decision-making aimed at optimizing their lifetime utility and endogenously determines college and university quality.

Using data from the NLSY97 survey, we estimate the model by matching it to moments in the data and run policy simulations based on different hypothesized policy plans. The values of key parameters are estimated using various approaches proposed by Todd and Wolpin (2020), and we compare these estimates to both the model assumptions and the moments observed in the data. The estimates are consistent with the assumptions made during model construction and closely align with the patterns observed in the data.

We find that even when all students lack parental financial assistance, significant disparities still exist in educational attainment and the endogenously chosen quality of colleges/universities between families with lower and higher incomes. The key factor driving differences in college access and quality is the amount of human capital accumulated from earlier educational stages. The base price of college tuition and the interest rate on student loans have negative but relatively small effects on college enrollment and the endogenous quality of colleges/universities.

After evaluating the effectiveness of multiple policy simulations, we found that fostering students' pre-college human capital is a more effective way to close the gap in college attendance and the endogenously selected quality of college/university between higher- and lower-income groups. Lowering the base price of college tuition or the interest rate does not help less wealthy students as much as it does wealthier students, which will widen the inequality in college attendance and the endogenous quality of college.

To conclude, even if all students were to borrow to fund their college or university

education, the differences in attendance and the endogenously selected quality of college or university are already deeply rooted in disparities in the quality of education they received during childhood. To close the gap in both access to college and the endogenously selected quality of college, the key approach is to provide primary and secondary education of higher quality more equally to all children.

Bibliography

Abbott, B., Gallipoli, G., Meghir, C., and Violante, G. L. (2019). Education policy and intergenerational transfers in equilibrium. *Journal of Political Economy*, 127(6):2569–2624.

Acemoglu, D. and Autor, D. (2011). Skills, tasks and technologies: Implications for employment and earnings. In *Handbook of Labor Economics*, volume 4, pages 1043–1171. Elsevier.

Agostinelli, F. and Wiswall, M. (2023). Estimating the technology of children's skill formation. *NBER Working Paper 22442*.

Amoyaw-Osei, Y., Agyekum, O. O., Pwamang, J. A., Mueller, E., Fasko, R., and Schluep, M. (2011). Ghana e-waste country assessment. *SBC e-waste Africa Project*, 66:111.

Arnett, J. J. (2023). *Emerging Adulthood: The Winding Road from the Late Teens through the Twenties*. Oxford University Press.

Attanasio, O., Meghir, C., and Nix, E. (2020). Human capital development and parental investment in India. *The Review of Economic Studies*, 87(6):2511–2541.

Aucejo, E. and James, J. (2021). The path to college education: The role of math and verbal skills. *Journal of Political Economy*, 129(10):2905–2946.

Bandowe, B. A. M., Bigalke, M., Boamah, L., Nyarko, E., Saalia, F. K., and Wilcke, W. (2014). Polycyclic aromatic compounds (PAHs and oxygenated PAHs) and trace metals in fish species from Ghana (West Africa): bioaccumulation and health risk assessment. *Environment International*, 65:135–146.

Black, D. A. and Smith, J. A. (2004). How robust is the evidence on the effects of college quality? Evidence from matching. *Journal of Econometrics*, 121(1-2):99–124.

Blair, C., Granger, D. A., Kivlighan, K. T., Mills-Koonce, R., Willoughby, M., Greenberg, M. T., Hibel, L. C., and Fortunato, C. K. (2008). Maternal and child contributions to cortisol response to emotional arousal in young children from low-income, rural communities. *Developmental Psychology*, 44(4):1095.

Boyatzis, R. E. (2008). Competencies in the 21st century. *Journal of Management Development*, 27(1):5–12.

Brunello, G. and Schlotter, M. (2011). Non-cognitive skills and personality traits: Labour market relevance and their development in education & training systems. *IZA Discussion Papers*.

Butts, K. (2021). Difference-in-differences with geocoded microdata. *arXiv preprint arXiv:2110.10192*.

Caetano, C. (2015). A test of exogeneity without instrumental variables in models with bunching. *Econometrica*, 83(4):1581–1600.

Caetano, C., Caetano, G., Nielsen, E., and Sanfelice, V. (2021). The effect of maternal labor supply on children: Evidence from bunching. *Working Paper*.

Cai, Z. and Heathcote, J. (2022). College tuition and income inequality. *American Economic Review*, 112(1):81–121.

Callaway, B., Goodman-Bacon, A., and Sant'Anna, P. H. (2024). Difference-in-differences with a continuous treatment. *NBER Working Paper No. w32117*.

Cameron, S. V. and Heckman, J. J. (1998). Life cycle schooling and dynamic selection bias: Models and evidence for five cohorts of American males. *Journal of Political Economy*, 106(2):262–333.

Campbell, H., Gupta, S., Dolan, G. P., Kapadia, S. J., Kumar Singh, A., Andrews, N., and Amirthalingam, G. (2018). Review of vaccination in pregnancy to prevent pertussis in early infancy. *Journal of Medical Microbiology*, 67(10):1426–1456.

Card, D. (1999). The causal effect of education on earnings. *Handbook of Labor Economics*, 3:1801–1863.

Carneiro, P., Crawford, C., and Goodman, A. (2007). The impact of early cognitive and non-cognitive skills on later outcomes. *CEE Discussion Papers*.

Carneiro, P. M. and Heckman, J. J. (2003). Human capital policy. *IZA Discussion Papers*.

Charnes, A., Frome, E. L., and Yu, P.-L. (1976). The equivalence of generalized least squares and maximum likelihood estimates in the exponential family. *Journal of the American Statistical Association*, 71(353):169–171.

Cherniss, C., Goleman, D., Emmerling, R., Cowan, K., and Adler, M. (1998). Bringing emotional intelligence to the workplace. *Consortium for Research on Emotional Intelligence in Organisations*.

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

Christensen, L. R., Jorgenson, D. W., and Lau, L. J. (1973). Transcendental logarithmic production frontiers. *The Review of Economics and Statistics*, 55(1):28–45.

Cohen, J. (2013). *Statistical Power Analysis for the Behavioral Sciences*. Routledge.

Cooke, R., Goulet, O., Huysentruyt, K., Joosten, K., Khadilkar, A. V., Mao, M., Meyer, R., Prentice, A. M., and Singhal, A. (2023). Catch-up growth in infants and young children with faltering growth: Expert opinion to guide general clinicians. *Journal of Pediatric Gastroenterology and Nutrition*, 77(1):7–15.

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

Cunha, F., Heckman, J. J., and Schennach, S. M. (2010). Estimating the technology of cognitive and noncognitive skill formation. *Econometrica*, 78(3):883–931.

Currie, J. and Almond, D. (2011). Human capital development before age five. In *Handbook of Labor Economics*, volume 4, pages 1315–1486. Elsevier.

Dale, S. and Krueger, A. B. (2011). Estimating the return to college selectivity over the career using administrative earnings data. Technical report, National Bureau of Economic Research.

Daum, K., Stoler, J., and Grant, R. J. (2017). Toward a more sustainable trajectory for e-waste policy: a review of a decade of e-waste research in Accra, Ghana. *International Journal of Environmental Research and Public Health*, 14(2):135.

De Fruyt, F. and Karevold, E. B. (2021). Personality in adolescence. *Handbook of Personality: Theory and Research*, pages 303–21.

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

DeSilver, D. (2019). A majority of US colleges admit most students who apply. *Pew Research Center*.

Diamond, A. (2013). Executive functions. *Annual Review of Psychology*, 64(1):135–168.

Ding, R., He, W., Wang, Q., and Qi, Z. (2022). Communicating emotional distress experienced by adolescents between adolescents and their mothers: Patterns and links with adolescents' emotional distress. *Journal of Affective Disorders*, 298:35–46.

Dolk, H., Vrijheid, M., Armstrong, B., Abramsky, L., Bianchi, F., Garne, E., Nelen, V., Robert, E., Scott, J. E., Stone, D., et al. (1998). Risk of congenital anomalies near hazardous-waste landfill sites in Europe: the EUROHAZCON study. *The Lancet*, 352(9126):423–427.

Dow, W. H., Philipson, T. J., and Sala-i Martin, X. (1999). Longevity complementarities under competing risks. *American Economic Review*, 89(5):1358–1371.

Fack, G. and Grenet, J. (2015). Improving college access and success for low-income students: Evidence from a large need-based grant program. *American Economic Journal: Applied Economics*, 7(2):1–34.

Flinn, C. J., Todd, P. E., and Zhang, W. (2018). Personality traits, intra-household allocation and the gender wage gap. *European Economic Review*, 109:191–220.

Florens, J.-P., Heckman, J. J., Meghir, C., and Vytlacil, E. (2008). Identification of treatment effects using control functions in models with continuous, endogenous treatment and heterogeneous effects. *Econometrica*, 76(5):1191–1206.

Fry, R. (2014). The changing profile of student borrowers: Biggest increase in borrowing has been among more affluent students. *Pew Research Center*.

Glewwe, P. (2002). Schools and skills in developing countries: education policies and socioeconomic outcomes. *Journal of Economic Literature*, 40(2):436–482.

Glewwe, P. and Todd, P. (2022). *Impact Evaluation in International Development: Theory, Methods, and Practice*. World Bank Publications.

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

Heckman, J. J. (1994). Is job training oversold? *Public Interest*, (115):91.

Heckman, J. J., Lochner, L. J., and Todd, P. E. (2006). Earnings functions, rates of return and treatment effects: The Mincer equation and beyond. *Handbook of the Economics of Education*, 1:307–458.

Hendricks, L., Koreshkova, T., and Leukhina, O. (2018). Sorting of students into colleges: Inefficiencies and policy implications. *Society for Economic Dynamics Working Paper*.

Hirshleifer, S., McKenzie, D., Almeida, R., and Ridao-Cano, C. (2016). The impact of vocational training for the unemployed: experimental evidence from Turkey. *The Economic Journal*, 126(597):2115–2146.

Hoff, A. (2002). The translog approximation of the constant elasticity of substitution production function with more than two input variables. WorkingPaper 14, Fødevareøkonomisk Institut.

Ihedioha, J., Ukoha, P., and Ekere, N. (2017). Ecological and human health risk assessment of heavy metal contamination in soil of a municipal solid waste dump in Uyo, Nigeria. *Environmental Geochemistry and Health*, 39:497–515.

Jaishankar, M., Tseten, T., Anbalagan, N., Mathew, B. B., and Beeregowda, K. N. (2014). Toxicity, mechanism and health effects of some heavy metals. *Interdisciplinary Toxicology*, 7(2):60.

Jarup, L., Briggs, D., De Hoogh, C., Morris, S., Hurt, C., Lewin, A., Maitland, I., Richardson, S., Wakefield, J., and Elliott, P. (2002). Cancer risks in populations living near landfill sites in Great Britain. *British Journal of Cancer*, 86(11):1732–1736.

Joshanloo, M. (2022). Neuroticism and openness moderate the relationship between negative affect and life satisfaction: a multi-level Bayesian analysis. *Applied Research in Quality of Life*, 17(6):3381–3391.

Kangmennaang, J., Bisung, E., and Elliott, S. J. (2020). 'We are drinking diseases': Perception of water insecurity and emotional distress in urban slums in Accra, Ghana. *International Journal of Environmental Research and Public Health*, 17(3):890.

Kanu, F. A. (2022). Progress Toward Achieving and Sustaining Maternal and Neonatal Tetanus Elimination — Worldwide, 2000–2020. *MMWR. Morbidity and Mortality Weekly Report*, 71.

Keane, M. P. and Wolpin, K. I. (2001). The effect of parental transfers and borrowing constraints on educational attainment. *International Economic Review*, 42(4):1051–1103.

Lindqvist, E. and Vestman, R. (2011). The labor market returns to cognitive and noncognitive ability: Evidence from the Swedish enlistment. *American Economic Journal: Applied Economics*, 3(1):101–128.

Losoncz, I. (2009). Personality traits in HILDA. *Australian Social Policy*, (8):169–198.

Lovo, S. and Rawlings, S. (2021). Garbage in, garbage out: the impact of e-waste dumping sites on early child health. *University of Reading Discussion Paper No. 2021-07*.

Lundborg, P., Plug, E., and Rasmussen, A. W. (2017). Can women have children and a career? IV evidence from IVF treatments. *American Economic Review*, 107(6):1611–1637.

Marek, S. and Dosenbach, N. U. (2018). The frontoparietal network: function, electrophysiology, and importance of individual precision mapping. *Dialogues in Clinical Neuroscience*, 20(2):133–140.

McFadden, D. (1973). Conditional Logit Analysis of Qualitative Choice Behavior. *Frontiers in Econometrics*.

Michler, J. D., Josephson, A., Kilic, T., and Murray, S. (2022). Privacy protection, measurement error, and the integration of remote sensing and socioeconomic survey data. *Journal of Development Economics*, 158:102927.

Monney, I., Buamah, R., Odai, S., Awuah, E., and Nyenje, P. (2013). Evaluating access to potable water and basic sanitation in Ghana's largest urban slum community: Old Fadama, Accra. *Journal of Environment and Earth Science*, 3(11):72–79.

National Center for Education Statistics (2015). The NCES fast facts tool provides quick answers to many education questions. *URL: https://nces.ed.gov/fastfacts/*.

National Center for Education Statistics (2024). Price of attending an undergraduate institution. *URL: https://nces.ed.gov/programs/coe/indicator/cua*.

National College Attainment Network (2023). The affordability of public colleges for Pell Grant recipients. *URL: https://www.ncan.org/page/pell*.

Nyarko, P., Pence, B. W., and Debpuur, C. (2001). Immunization status and child survival in rural Ghana. *Population Council, DOI: 10.31899/pgy6.1049*.

Onis, M. d., Onyango, A. W., Borghi, E., Siyam, A., Nishida, C., and Siekmann, J. (2007). Development of a who growth reference for school-aged children and adolescents. *Bulletin of the World health Organization*, 85(9):660–667.

Owusu Boadi, K. and Kuitunen, M. (2002). Urban waste pollution in the Korle Lagoon, Accra, Ghana. *Environmentalist*, 22:301–309.

Owusu-Sekyere, K., Batteiger, A., Afoblikame, R., Hafner, G., and Kranert, M. (2022). Assessing data in the informal e-waste sector: The Agbogbloshie Scrapyard. *Waste Management*, 139:158–167.

Page, L. C. and Scott-Clayton, J. (2016). Improving college access in the United States: Barriers and policy responses. *Economics of Education Review*, 51:4–22.

Parvez, S. M., Jahan, F., Brune, M.-N., Gorman, J. F., Rahman, M. J., Carpenter, D., Islam, Z., Rahman, M., Aich, N., Knibbs, L. D., et al. (2021). Health consequences of exposure to e-waste: an updated systematic review. *The Lancet Planetary Health*, 5(12):e905–e920.

Pei, Z., Pischke, J.-S., and Schwandt, H. (2019). Poorly measured confounders are more useful on the left than on the right. *Journal of Business & Economic Statistics*, 37(2):205–216.

Perin, M. C. A. A., Schlindwein, C. F., de Moraes-Pinto, M. I., Simão-Gurge, R. M., Mimica, A. F. d. M. A., Goulart, A. L., and dos Santos, A. M. N. (2012). Immune response to tetanus booster in infants aged 15 months born prematurely with very low birth weight. *Vaccine*, 30(46):6521–6526.

Rai, P. K., Lee, S. S., Zhang, M., Tsang, Y. F., and Kim, K.-H. (2019). Heavy metals in food crops: Health risks, fate, mechanisms, and management. *Environment International*, 125:365–385.

Rehman, K., Fatima, F., Waheed, I., and Akash, M. S. H. (2018). Prevalence of exposure of heavy metals and their impact on health consequences. *Journal of Cellular Biochemistry*, 119(1):157–184.

Sampson, K. (2015). How Ewaste Recycling is Creating a Lot of Jobs. *Hummingbird International*.

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

Schretlen, D. J., van der Hulst, E.-J., Pearlson, G. D., and Gordon, B. (2010). A neuropsychological study of personality: Trait openness in relation to intelligence, fluency, and executive functioning. *Journal of Clinical and Experimental Neuropsychology*, 32(10):1068–1073.

Summerfield, M., Garrard, B., Kamath, R., Macalalad, N., Nesa, M. K., Watson, N., Wilkins, R., and Wooden, M. (2023). HILDA user manual–release 22. *Melbourne Institute of Applied Economic and Social Research, The University of Melbourne*.

Thanomsangad, P., Tengjaroenkul, B., Sriuttha, M., and Neeratanaphan, L. (2020). Heavy metal accumulation in frogs surrounding an e-waste dump site and human health risk assessment. *Human and Ecological Risk Assessment: An International Journal*.

The New York Times (2013). Where the SAT and ACT Dominate. *URL: archive.nytimes.com/www.nytimes.com/interactive/2013/08/04/education/edlife/where-the-sat-and-act-dominate.html*.

Todd, P. E. and Wolpin, K. I. (2003). On the specification and estimation of the production function for cognitive achievement. *The Economic Journal*, 113(485):F3–F33.

Todd, P. E. and Wolpin, K. I. (2007). The production of cognitive achievement in children: Home, school, and racial test score gaps. *Journal of Human Capital*, 1(1):91–136.

Todd, P. E. and Wolpin, K. I. (2020). The best of both worlds: Combining RCTs with structural modeling. *Journal of Economic Literature*.

Todd, P. E. and Wolpin, K. I. (2023). The best of both worlds: combining randomized controlled trials with structural modeling. *Journal of Economic Literature*, 61(1):41–85.

Todd, P. E. and Zhang, W. (2020). A dynamic model of personality, schooling, and occupational choice. *Quantitative Economics*, 11(1):231–275.

Tooher, R., Griffin, T., Shute, E., and Maddern, G. (2005). Vaccinations for waste-handling workers. a review of the literature. *Waste Management & Research*, 23(1):79–86.

Train, K. E. (2009). *Discrete Choice Methods with Simulation*. Cambridge University Press.

UN Environment Programme (2019). UN report: Time to seize opportunity, tackle challenge of e-waste. *URL: https://www.unep.org/news-and-stories/press-release/un-report-time-seize-opportunity-tackle-challenge-e-waste*.

United Nations University (2020). Global E-Waste Surging: Up 21% in 5 Years. *URL: https://unu.edu/press-release/global-e-waste-surging-21-5-years*.

U.S. Department of Labor (2009). History of Federal Minimum Wage Rates Under the Fair Labor Standards Act, 1938 - 2009. *URL: https://www.dol.gov/agencies/whd/minimum-wage/history/chart*.

Van den Akker, A. L., Briley, D. A., Grotzinger, A. D., Tackett, J. L., Tucker-Drob, E. M., and Harden, K. P. (2021). Adolescent Big Five personality and pubertal development: Pubertal hormone concentrations and self-reported pubertal status. *Developmental Psychology*, 57(1):60.

Williams, J. (2003). The natural rate of interest. *FRBSF Economic Letter*, (oct31).

Wooldridge, J. M. (2010). *Econometric Analysis of Cross Section and Panel Data*. The MIT Press.

Wooldridge, J. M. (2016). *Introductory Econometrics: A Modern Approach 6rd ed.* Cengage Learning.

World Health Organization (2007). Population health and waste management: scientific data and policy options: report of a WHO workshop: Rome, Italy, 29–30 March 2007.

World Health Organization (2021). Soaring e-waste affects the health of millions of children, WHO warns. *URL: https://www.who.int/news/item/15-06-2021-soaring-e-waste-affects-the-health-of-millions-of-children-who-warns*.

Yoder, J. R., Grady, M., and Dillard, R. (2019). Maternal caregiving practices and child abuse experiences as developmental antecedents to insecure attachments: Differential pathways between adolescents who commit sexual and non-sexual crimes. *Sexual Abuse*, 31(7):837–861.

Zacharopoulos, G., Sella, F., and Cohen Kadosh, R. (2021). The impact of a lack of mathematical education on brain development and future attainment. *Proceedings of the National Academy of Sciences*, 118(24):e2013155118.

Appendix A

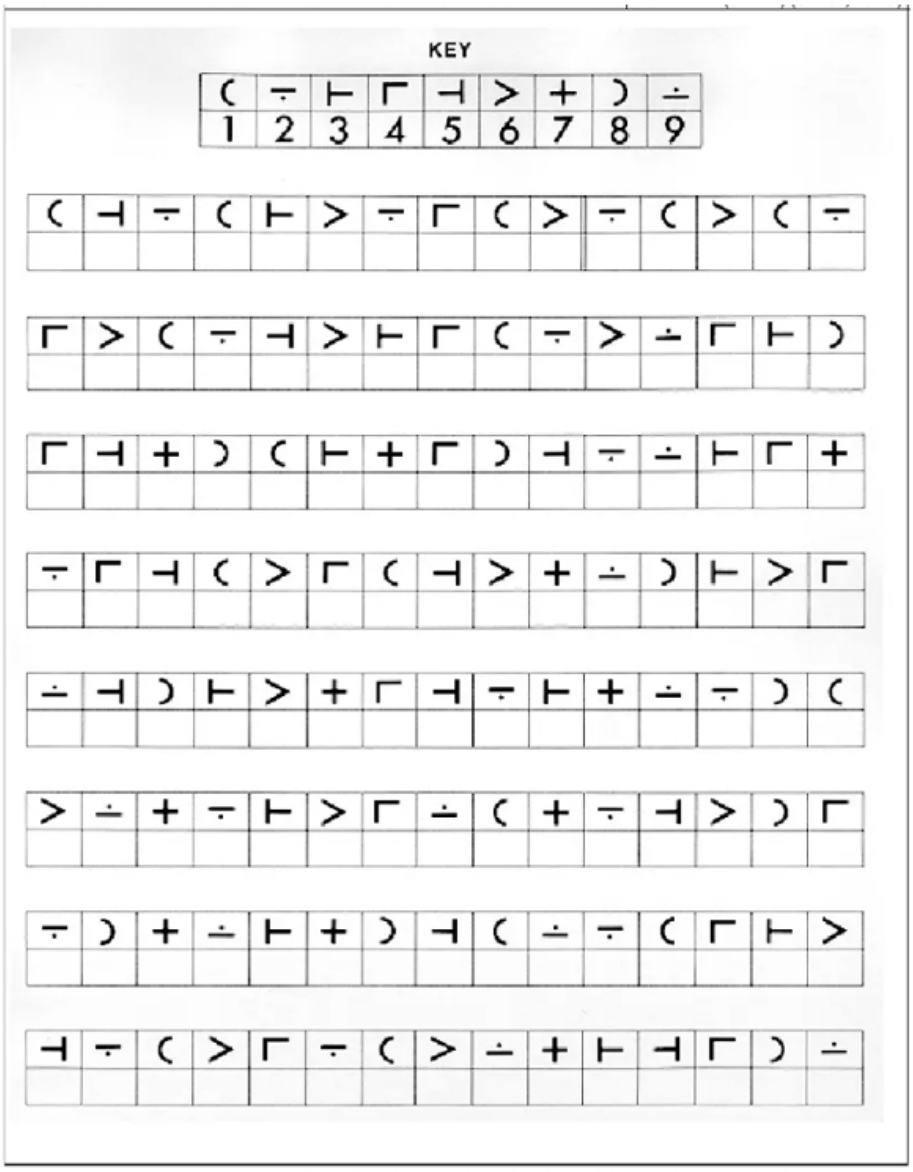
The Development of Personality Traits and Cognitive Skills in Adolescence: Evidence from a Skill Formation Model and a Control Function Approach

Figure A.1: Personality Traits Items in HILDA

	Personality trait items in HILDA and in the TDA-40					
	Emotional stability	–	**Extraversion**	**Conscientiousness**	**Agreeableness**	**Openness to experience**
TDA-40 items excluded from HILDA	Relaxed Unenvious		Bold Energetic	Organised Practical	Unsympathetic Rude	Uncreative Unintellectual
Items in HILDA from TDA-40	Envious Fretful Jealous Moody Temperamental Touchy		Bashful Extroverted Quiet Shy Talkative Withdrawn	Careless Disorganised Efficient Inefficient Sloppy Systematic	Cold Cooperative Harsh Kind Sympathetic Warm	Complex Creative Deep Imaginative Intellectual Philosophical
Items in HILDA from other sources	Calm		Enthusiastic Lively	Orderly Traditional	Selfish	

Note. Source of the table is from Losoncz (2009).

Figure A.2: Symbol Digit Modalities Test Example



Note. According to Summerfield et al. (2021), the HILDA adopts symbol digit modalities tests from Aaron Smith (2007), which an example is shown in this figure.

Table A.1: Predicting Self-rated Skills using Test Scores

	National Reading Test	Symbol-digit Modalities Score
Self-rated Reading Skill	0.437***	
	(0.014)	
Self-rated Math Skill		0.377***
		(0.016)

Note: Standard errors (clustered at household level) in parentheses * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. The table reports the coefficients of the OLS regression that predicts self-rated reading skill using the National Reading Test score and predicts self-rated math skill using the symbol-digit modalities score.

Appendix B

The Impacts of E-Waste Dumping on Infant Health Outcomes in Ghana

Table B.1: Placebo Test Using 1995 as the Year of Treatment (5km Buffer)

	(1) Infant Mortality	(2) Birth Weight	(3) Diarrhea	(4) Cough	(5) Weight-for-age Z-score	(6) Height-for-age Z-score
Placebo Treatment	0	0.04	-0.04	0.06	0.25	0.02
	(0.03)	(0.31)	(0.15)	(0.17)	(0.43)	(0.43)
Cluster Fixed Effect	Yes	Yes	Yes	Yes	Yes	Yes
Controls	No	No	No	No	No	No
Sample Size	2109	263	345	344	331	329
Placebo Treatment	0	0.11	-0.07	0.05	0.39	0.31
	(0.03)	(0.29)	(0.16)	(0.17)	(0.44)	(0.50)
Cluster Fixed Effect	Yes	Yes	Yes	Yes	Yes	Yes
Controls	Yes	Yes	Yes	Yes	Yes	Yes
Sample Size	2109	263	345	344	331	329

Note: Significance: * 0.10; ** 0.05; *** 0.01. The columns present the placebo DID using 1995 as the year of treatment with 5km buffer zone. Each column presents the coefficients of DID estimators for a single outcome. The table reports the results of all outcomes that have shown up in the paper. Standard errors are cluster robust at the DHS cluster level. The vector of control variables includes the mother's education, mother's weight, sex of the child, if the child is born as one of the twins, and the mother's age at birth.

Table B.2: Robustness Check Using a Placebo Outcome (Twins)

	(1)	(2)	(3)	(4)	(5)	(6)
DID Estimate	-0.04	0.00	0.01	0.03	0.04	0.08
	(0.03)	(0.02)	(0.03)	(0.05)	(0.06)	(0.06)
Cluster Fixed Effect	Yes	Yes	Yes	Yes	Yes	Yes
Controls	No	No	No	No	No	No
Sample Size	3,907	3,438	2,764	2,439	4,428	4,428
DID Estimate	-0.04	0.00	0.01	0.03	0.04	0.08
	(0.03)	(0.02)	(0.03)	(0.05)	(0.06)	(0.06)
Cluster Fixed Effect	Yes	Yes	Yes	Yes	Yes	Yes
Controls	Yes	Yes	Yes	Yes	Yes	Yes
Sample Size	3,907	3,438	2,764	2,439	4,428	4,428

Note: Significance: * 0.10; ** 0.05; *** 0.01. The columns present the DID estimator using different buffer zones to identify the treated group. Specification (1) reports the results of using 5km buffer zone, (2) using 4km buffer zone, (3) using 3km buffer zone, (4) using 2km buffer zone, (5) using continuous treatment with 1/distance to the dumping site, and (6) using continuous treatment with 1/distance squared to the dumping site. Standard errors are cluster robust at the DHS cluster level. The vector of control variables includes the mother's education, mother's weight, sex of the child, whether the child is born as one of the twins, and the mother's age at birth.

Table B.3: Robustness Check for Effects on Diarrhea (Exclude the MICS Data)

	(1)	(2)	(3)	(4)	(5)	(6)
DID Estimate	0.41***	0.41***	0.42***	0.42***	0.72***	1.20***
	(0.07)	(0.08)	(0.08)	(0.08)	(0.15)	(0.19)
Cluster Fixed Effect	Yes	Yes	Yes	Yes	Yes	Yes
Controls	No	No	No	No	No	No
Sample Size	222	204	186	155	252	252
DID Estimate	0.57***	0.57***	0.58***	0.61***	0.98***	1.56***
	(0.20)	(0.21)	(0.22)	(0.22)	(0.33)	(0.53)
Cluster Fixed Effect	Yes	Yes	Yes	Yes	Yes	Yes
Controls	Yes	Yes	Yes	Yes	Yes	Yes
Sample Size	222	204	186	155	252	252

Note: Significance: * 0.10; ** 0.05; *** 0.01. The columns present the DID estimator using different buffer zones to identify the treated group. Specification (1) reports the results of using 5km buffer zone, (2) using 4km buffer zone, (3) using 3km buffer zone, (4) using 2km buffer zone, (5) using continuous treatment with 1/distance to the dumping site, and (6) using continuous treatment with 1/distance squared to the dumping site. Standard errors are cluster robust at the DHS cluster level. The vector of control variables includes the mother's education, mother's weight, sex of the child, whether the child is born as one of the twins, and the mother's age at birth.

Table B.4: Percentage of Missing Data in the Treated and Control Groups

	Treatment	Control	Difference
	(1)	(2)	(3)
Birth Weight			
Family Wealth	29.85%	51.14%	-21.29%
Mother's Education Level	20.91%	10.25%	10.66%
Mother's Employment	94.91%	92.49%	2.42%
Diarrhea			
Family Wealth	40.69%	49.89%	-9.20%
Mother's Education Level	11.21%	8.39%	2.82%
Mother's Employment	94.30%	91.97%	2.33%
Cough			
Family Wealth	40.65%	49.89%	-9.24%
Mother's Education Level	11.43%	8.39%	3.04%
Mother's Employment	94.98%	91.97%	3.01%
Height Z-score			
Family Wealth	39.89%	49.16%	-9.27%
Mother's Education Level	12.80%	9.22%	3.58%
Mother's Employment	93.66%	91.71%	1.95%
Weight Z-score			
Family Wealth	39.83%	49.26%	-9.43%
Mother's Education Level	12.66%	8.98%	3.68%
Mother's Employment	93.56%	91.84%	1.72%

Note: For every outcome in the treatment and the control group, I show the percentage of households among the missing data in the corresponding outcome that are below the average of household wealth, have mothers whose highest degree completed are lower than secondary education, and have mothers work in the industries that are more likely to be exposed to e-waste pollution.

Appendix C

The Decision-Making of College Enrollment in an Increasingly Independent World

The transition from Equations (4.12) to (4.14) was made without a detailed proof that the difference $(\varepsilon_{i1} - \varepsilon_{i0})$ between two independently and identically Extreme Value distributed error terms follows a logistic distribution. In this section, we provide the proof for the transition from Equations (4.12) to (4.14), which relies heavily on McFadden (1973) and Train (2009).

First, the logit model is obtained by assuming that each ε_{ij} is independently and identically distributed according to an extreme value distribution, also known as the Gumbel and type I extreme value distribution. The subscript i denotes any individual i, and j denotes the discrete choice of college attendance. The probability density for each unobserved component of an individual's utility is:

$$f(\varepsilon_{ij}) = e^{-\varepsilon_{ij}} e^{-e^{-\varepsilon_{ij}}},$$

of which the cumulative distribution is

$$F(\varepsilon_{ij}) = e^{-e^{-\varepsilon_{ij}}}.$$

The probability that decision maker i will attend college ($j = 1$) is

$$\begin{aligned} P_{i1} &= P[V_{i1} + \varepsilon_{i1} > V_{ij} + \varepsilon_{ij}, \forall j \neq 1] \\ &= P[\varepsilon_{ij} < V_{i1} - V_{ij} + \varepsilon_{i1}, \forall j \neq 1]. \end{aligned}$$

This expression represents the cumulative distribution for each ε_{ij} evaluated at $V_{i1} - V_{ij} + \varepsilon_{i1}$. Therefore, $F(\varepsilon_{ij})$ can be expressed as:

$$F(\varepsilon_{ij}) = e^{-e^{-(V_{i1} - V_{ij} + \varepsilon_{i1})}}.$$

Since the ε's are independent, this cumulative distribution over all alternatives (where $j \neq 1$) is the product of the individual cumulative distributions:

$$P_{i1}|\varepsilon_{i1} = \prod_{j \neq 1} e^{-e^{-(V_{i1} - V_{ij} + \varepsilon_{i1})}}$$

Since ε_{i1} is not given, the choice probability is the integral of $P_{i1}|\varepsilon_{i1}$ over all values of ε_{i1} weighted by its density:

$$P_{i1} = \int_{\varepsilon_{i1}=-\infty}^{\infty} \left(\prod_{j \neq 1} e^{-e^{-(V_{i1} - V_{ij} + \varepsilon_{i1})}} \right) e^{-\varepsilon_{i1}} e^{-e^{-\varepsilon_{i1}}} \, d\varepsilon_{i1}$$

Then, by collecting terms in the exponent of e and noting that $V_{i1} - V_{i1} = 0$, we have

$$\begin{aligned} P_{i1} &= \int_{\varepsilon_{i1}=-\infty}^{\infty} \left(\prod_{j} e^{-e^{-(V_{i1} - V_{ij} + \varepsilon_{i1})}} \right) e^{-\varepsilon_{i1}} \, d\varepsilon_{i1} \\ &= \int_{\varepsilon_{i1}=-\infty}^{\infty} exp\left(-\sum_{j} e^{-(V_{i1} - V_{ij} + \varepsilon_{i1})} \right) e^{-\varepsilon_{i1}} \, d\varepsilon_{i1} \\ &= \int_{\varepsilon_{i1}=-\infty}^{\infty} exp\left(-e^{-\varepsilon_{i1}} \sum_{j} e^{-(V_{i1} - V_{ij})} \right) e^{-\varepsilon_{i1}} \, d\varepsilon_{i1}. \end{aligned}$$

Define $t = exp(-\varepsilon_{i1})$ such that $-exp(-\varepsilon_{i1})ds = dt$. Note that as ε_{i1} approaches infinity, t approaches zero, and as ε_{i1} approaches negative infinity, t approaches positive infinity. Using t as the new term, we have

$$
\begin{aligned}
P_{i1} &= \int_{\infty}^{0} exp\left(-t\sum_j e^{-(V_{i1}-V_{ij})}\right)(-dt) \\
&= \int_{0}^{\infty} exp\left(-t\sum_j e^{-(V_{i1}-V_{ij})}\right)dt \\
&= \left.\frac{exp\left(-t\sum_j e^{-(V_{i1}-V_{ij})}\right)}{-\sum_j e^{-(V_{i1}-V_{ij})}}\right|_0^{\infty} \\
&= \frac{1}{\sum_j e^{-(V_{i1}-V_{ij})}} \\
&= \frac{e^{V_{i1}}}{\sum_j e^{V_{ij}}}.
\end{aligned}
$$

We proved the general case where multiple alternatives exist beyond attending college, while the setup in this paper considers only two options: attending college or not attending, i.e., $j \in \{0, 1\}$. In our case, the probability that decision maker i choose to attend college $(j = 1)$ is

$$
\begin{aligned}
P_{i1} &= P[V_{i1} + \varepsilon_{i1} > V_{i0} + \varepsilon_{i0}] \\
&= P[\varepsilon_{i0} < V_{i1} - V_{i0} + \varepsilon_{i1}] \\
&= \frac{e^{V_{i1}}}{e^{V_{i1}} + e^{V_{i0}}} \\
&= \frac{e^{V_{i1}}}{1 + e^{(V_{i0}-V_{i1})}},
\end{aligned}
$$

which is equivalent to

$$
\frac{1}{1 + exp\left(a_0 L_{i1} + \delta a_0 L_{i2} - \delta\alpha_i\pi L_{i2}\left(\frac{\beta\alpha_i\pi L_{i2}}{(1+r_i)p_i}\right)^{\frac{\beta}{1-\beta}} + \delta(1+r_i)p_i\left(\frac{\beta\alpha_i\pi L_2}{(1+r_i)p_i}\right)^{\frac{1}{1-\beta}}\right)},
$$

as required.