Human Capital Production in Childhood

Essays on the Economics of Education

By

Maximiliaan Willem Pierre Thijssen

Thesis submitted in fulfilment of
the requirements for the degree of
PHILOSOPHIAE DOCTOR
(PhD)

Universitetet
i Stavanger

PhD Program in Social Sciences
Department of Economics and Finance
University of Stavanger Business School
2022

University of Stavanger
NO-4036 Stavanger
NORWAY
www.uis.no

ISBN: 978-82-8439-086-4
ISSN: 1890-1387
PhD: Thesis UiS No. 647

I dedicate this thesis to my parents, Annie and Wil.

Acknowledgements

The past four years have been an incredible journey. This journey has revealed many important lessons that have made me a more competent researcher and a better person. Make no mistake, I did not take this journey alone. If it wasn't for some very important people along the way, the thesis you are about to read would not be what it is today. Therefore, I'd like to acknowledge several people that have encouraged, inspired, and sustained me throughout the years and recognize that this thesis is as much their accomplishment as it is mine.

First and foremost, I'm deeply indebted to Mari Rege and Ingeborg Foldøy Solli. It is not easy to put into words how grateful I am for the experiences and opportunities both of you have provided. Without your continued support and enthusiasm for my work, this thesis would not be what it is today. Nor would I be the person I am today without your mentoring. Anyone interested in pursuing a Ph.D. could not wish for better supervisors. Mari, I have learned so many valuable lessons from you. You are extremely knowledgeable, and your ability to give critical yet constructive feedback is second to none. You have challenged my thinking on many occasions and made me see things from different perspectives. You have provided me with countless opportunities. While you are extremely busy, you always made the time if I needed your input or support. Ingeborg, you have always been available whenever I had questions or needed help with the data. The way you talk about child development and education is catching and has served as a great catalyst for my motivation and inspiration. You have shared your skills, knowledge, and expertise with me, for which I am very grateful. To both of you, I will be forever thankful. The lessons you both have taught me will be invaluable in all my future endeavors.

I would also like to take this opportunity to thank Eric Bettinger. While we have not had many opportunities to meet, I am thankful for the valuable insights and feedback you have given on some of the chapters in this thesis.

I would also like to extend my sincere thanks to Ingunn Størksen and Oddny Judith Solheim. Ingunn, I am thankful that you provided me with many insights on child development and the measurement of psychological constructs. I am grateful for our collaboration on the third chapter of this thesis. Oddny, I have been lucky to collaborate with you on the research project that makes up the fourth chapter of this thesis. I learned a lot from your insights and suggestions as an education scientist during the project, especially regarding the theory. The interdisciplinary perspectives both of you brought to my research have made this thesis richer than it otherwise would have been. It has been an absolute pleasure to learn from the insights you both have provided.

As part of this thesis, I have had the opportunity to work with data from the Agder Project and the Two Teachers Project. Dozens of people and hundreds of hours have been spent planning, organizing, and implementing these high-quality field experiments. Since these data serve as the foundation for two chapters in this thesis, I would like to give a special thanks to all who made these projects possible. In particular, I'd like to recognize the project leaders of the Agder Project, Mari Rege and Ingunn Størksen, and the project management and administration of the Two Teachers Project, Oddny Judith Solheim, Erin Margaret McTigue, Åse Kari Hansen Wagner, and Mari Rege. In addition, thanks to all the researchers, administrators, teachers, and children who made the project possible.

My thesis has benefited greatly from comments and suggestions along the way. I give special thanks to Erik Øiolf Sørensen, the discussant at my fifty percent seminar, and Torberg Falch at my ninety percent seminar. Both of you have carefully commented on my work, for which I am grateful. Also, thanks to the many participants at the Quantitative Forum. A special thanks to Knud Knudsen and Ulrich Dettweiler for their comments and suggestions. Also, thanks to the participants of the Ph.D. Brown Bag.

I would further like to acknowledge the support of the UiS Business School, our dean Ola Kvaløy, and Christian Jensen, head of the Department of Economics and Finance. Special thanks to Egil Clement Svela for helping with many non-research-related questions and Åse Lea for helping me with the many administrative procedures as well as organizing workshops, lunches, and presentations. I am also extremely grateful for my colleagues. I'd like to thank Ana Faria Lopes, Elin Svensen, Erika Christie Berle, Espen Sagen, Gorm Kipperberg, Helle Austvik Tholo, Heidi Victoria Skeiseid, Irene Kengen, Jason Deegan, Kwadwo Atta-Owusu, Nina Hjertvikrem, Silje Haus-Reve, Tom Brökel, and Venke Furre Haaland, for all the interesting conversations, coffee breaks, support, and for showing interest in my research throughout. You have made the past years a memorable experience, which I will never forget.

I would like to recognize Andreas Østbø Fidjeland, May Linn Auestad, and Simone Valerie Häckl-Schermer, in particular. Andreas, during the past years, my journey has been very enjoyable because you were around for interesting discussions, coffee breaks, and the occasional beer. We have been to courses together. Met up several times in San Francisco during our research stay in California, and we organized the annual Ph.D. workshop in Economics of Education together. I have learned a lot from you. May Linn, your presence in the office early in the morning brought me great joy. You were always willing to help me whenever I had questions or needed someone to bounce ideas off. I enjoyed our deep and engaging discussions early in the morning. Simone, you have given me a lot of moral support the past year and a half, for which I am very grateful. Your positive attitude has kept me going in times of struggle. I am thankful for all your constructive feedback, "quick" coffees, and having someone to share the good and bad.

I am also grateful for Xander Stegehuis. Xander, you always lent an ear when I needed to vent my frustrations. You have given me feedback on my writing, and you have helped me relativize the Ph.D. experience. I would also like to thank Roy Overbeek. Your music recommendations have helped me get through many hours of coding. Lastly, I would like to thank Carolin Straubel for offering a great perspective on work-life balance. I am grateful to have such friends.

Last but certainly not least, this thesis would not have been possible without the support of my parents, Annie and Wil Thijssen. I cannot begin to express how grateful I am for everything that the two of you have done for me. Thank you for always supporting me, irrespective of the choices I make. Thanks for being there when I need you. I also appreciate all your gentle reminders that there is more in life than work. You both have been an incredible inspiration and support throughout my journey. Thank you both for everything.

Maximiliaan Willem Pierre Thijssen
January, 2022

I received funding from Grant 270703 from the Norwegian Research Council. The Norwegian Research Council also funded my research stay at the University of California, Berkeley, application number: 291523.

Contents

Chapter 1

Introduction

By

Maximiliaan W. P. Thijssen

1. Overview

The skills demanded in the labor market have changed (see, *e.g.*, Acemoglu & Autor 2011).[1] Just as technological innovations have increased demand for workers capable of performing non-routine and complex tasks, demand for routine work has declined (Autor et al. 2003, Spitz-Oener 2006). Since lifelong learning can help individuals develop resilience and adapt to adverse shocks in changing labor markets, it is important to learn more about the best ways to support such learning.[2]

Skill formation starts in childhood, as does lifelong learning. By investing early, and with the support of teachers and caregivers, we can equip children with a strong foundation for further development (National Research Council Institute of Medicine 2000). Societies can invest in skill formation by improving the quality of children's environments. Indeed, the home environment, early childhood education and care (ECEC) centers, and schools can all affect children's skill formation (Almond & Currie 2011, Blau & Currie 2006, Cunha et al. 2006, Hanushek & Rivkin 2006, Heckman & Mosso 2014). This influential role of the environment has implications for policy because it suggests a role for investments that enhance its quality. Yet gathering evidence on environmental factors that affect skill formation is no simple task, and much remains to be learned about the relevant processes.

This thesis focuses on three key questions: First, what is the best way

[1]During the second half of the twentieth century, innovations such as the internet, home computers, and the World Wide Web were introduced in many Western economies. The twenty-first century has seen the emergence of a wealth of technological innovations, including artificial intelligence and autonomous factories.

[2]For a recent report on the importance of lifelong learning, see OECD (2021).

to gain a better understanding of how children acquire new skills? Econometric tools have been developed to evaluate the effectiveness of environmental factors that affect skill formation. However, these tools cannot be employed unless children are observed at equally spaced intervals. Second, which of the many possible skills should investments focus on? During early childhood, it is best to target skills that promote a child's further development, and in the context of the crucial transition from preschool to primary school, these could be skills that help prepare children for success in school. Third, once certain skills have been nurtured during early childhood, how can we provide effective education to sustain and build on this foundation? Although we know that teachers play a key role in education, much remains to be learned about what makes teachers effective. By addressing these questions, this thesis seeks to deepen our understanding of the processes driving skill formation in childhood, yielding new and better insights into public policy design. The individual chapters are described in more detail below.

Chapter 2 examines the literature on the econometrics of childhood skill formation. This literature comprises econometric tools for assessing the effectiveness of environmental factors and models influences on skill formation. In particular, it helps us understand what can be learned from a model of skill formation (identification), how we can best learn it (estimation), and how certain we can be about parameters we estimate (inference). Todd & Wolpin (2003), Cunha & Heckman (2008), Cunha et al. (2010), and Agostinelli & Wiswall (2016) develop tools to address several challenges. First, scales used to assess children may not have a cardinal interpretation. Second, given the difficulty of assessing children, any instrument is likely to be subject to measurement error. Third, unobserved inputs may correlate with observed inputs. Just as non-cardinality may result in biased interpretation, measurement error and the presence of unobserved inputs may cause biased estimation.

I evaluate the implications of a challenge that has received little attention in the past. When we estimate skill formation models, we typically assume that all children are observed in equally spaced intervals. The interval assumed by our model is therefore like the one observed in the data. In most longitudinal studies, however, observation intervals are not equally spaced (McKenzie 2001, Millimet & McDonough 2017). One might also argue that the substantive timing variable of child development is age, not the assessment wave. In this case, unequal spacing may occur even when survey waves are equidistant because (i) same-aged children may not be observed simultaneously, and (ii) simultaneously observed children may not be of the same age.

Most longitudinal studies target "same-aged" children based on birth year. The development of children born in the same year and assessed simultaneously can differ by as much as 12 months. If, in addition, assessments can occur at any time during the year, then the development of children born in the same year may differ by as much as 24 months. Such differences can have an impact on cognitive skill formation (Crawford et al. 2010). Of course, we would expect these developmental differences to become smaller as children grow older (Elder & Lubotsky 2009). Still, failing to account for unequally spaced intervals may lead to biased estimation, particularly for young children. Chapter 2 examines the implications of this estimation problem with the help of an often-used dataset for modeling child development.

Chapter 3 draws from the early childhood literature and seeks to determine which skills should be nurtured in early childhood. We know that children start school with different skill levels, and this insight has prompted an interest in early childhood education programs that boost "school readiness" (see, *e.g.*, Clements & Sarama 2011, Diamond & Lee 2011, Dillon et al. 2017, Rege et al. 2021). A challenge in designing these programs is deciding what skills to target, as not all skills are equally important for school success (Duncan et al. 2007, Lewit & Baker 1995).

An increasing number of studies have evaluated the effectiveness of early childhood education programs designed to reduce early skill disparities (*e.g.*, Attanasio et al. 2020, Conti et al. 2016, Heckman et al. 2013, Sylvia et al. 2020). These studies generally find that allocating resources to efforts to promote early skill formation can be an effective approach. However, depending on the type of skills targeted and the nature of the intervention, effects may not persist over time. For example, the effects of an investment that targets skills in early childhood that children will eventually develop irrespective of the investment may fade out. As a result, targeting such skills may not be optimal (Bailey et al. 2017).

Chapter 3 studies whether executive functions, defined as the cognitive control processes necessary for concentration and thinking (Diamond & Lee 2011), are skills that programs should focus on. As they involve fundamental skills, executive functions would appear to be a natural starting point (Diamond & Lee 2011, Howard-Jones et al. 2012). One could argue, however, that the structure provided in school enables children to develop the same level of executive functioning as they would have if they had been the beneficiaries of targeted investments in promoting such executive functions in preschool. But if executive functions are the basis for learning many other skills, children starting school with higher levels of executive functioning may be more efficient at learning other skills than their peers (*i.e.*, skill begets skill). Understanding the processes that drive skill formation in early childhood yields insights for public policy decisions about how educational resources should be used.

Chapter 4 looks at how skills nurtured in early childhood can be maintained through effective education, and examines the literature regarding the education production function. This literature focuses on school inputs that are effective in promoting children's development. While there are many kinds of school inputs (*e.g.*, class size, number of books, number of computers), Chapter 4 focuses specifically on teachers. Teacher effectiveness may vary widely, even in the same school (Aaronson et al. 2007, Araujo et al. 2016, Jackson 2018, Kraft 2019, Rivkin et al. 2005, Rockoff 2004). Moreover, effective teachers may have long-term impacts on children's education and labor market outcomes (Chetty et al. 2014*b*, Opper 2019). Lastly, teachers are the largest budgetary expense in most schools (Hanushek & Rivkin 2006).

While we can *identify* effective teachers through value-added estimates (Chetty et al. 2014*a*), we do not know how to *replicate* them (*cf*. Kane & Staiger 2012). The literature indicates that a teacher's readily observable characteristics, such as education, salary, or test scores, do not consistently predict children's academic achievement (Hanushek & Rivkin 2006). For this reason, researchers have started to focus on what goes on inside the classroom. The child development literature suggests that the quality of the child's relationship with the teacher and classmates, as perceived by the child, is particularly important (see, *e.g.*, Connell & Wellborn 1991, Hamre & Pianta 2001, Pianta 1997).

Still, the numerous studies that suggest that teacher relationship skills, as perceived by the child, are essential for learning may be biased by a child's (unobserved) preferences for a particular relationship. Recently, economists have started to refine our measures of investments in education to capture objective, detailed information about the quality of teacher-child interactions (Araujo et al. 2016, Kane et al. 2011). However, these classroom observations are costly and may fail to capture fundamental aspects of a child's perceptions that ultimately drive behavior (Connell & Wellborn 1991). Moreover, it is important to evaluate teachers and what goes on inside the classroom, with the help of a variety of assessments (Kane & Staiger 2012). Chapter 4 introduces and validates a new approach for measuring teachers' overall ability to form positive relationships in the classroom (based on the children's perspectives). This approach has implications for public policy, as such assessments can serve as tools for identifying teachers who need support, promoting development, conducting progress evaluations, and helping policymakers improve quality.

A key motivation for focusing on childhood skill formation is a desire to ensure equal opportunities for all children. Section 2 explains the concept of equality of opportunity and shows that skill disparities emerge early and persist over time. Since all three chapters of this dissertation conceptualize child development in keeping with the technology of skill formation, I introduce this technology and the related literature in Section 3. All chapters address a key challenge – namely, the fact that skills are inherently unobservable. Section 4 elaborates on how, despite that fact, we can learn about a child's skills. I begin by describing the general intuition underlying the measurement of unobserved variables. Since the measures used in each chapter have been validated in other studies, I briefly describe how such validation is typically achieved. Section 4 concludes with a description of a dedicated measurement model of the type used in each chapter. Section 5 summarizes each of the chapters.

2. Equality of Opportunity

Chapter 3 focuses on "school readiness," recognizing that children start school at different skill levels. These differences can likely be attributed to the early environments that have benefited children differently (Duncan

& Murnane 2011). Since these differences predict school success, not all children have the same opportunities to learn and benefit from school resources. Chapter 4 focuses on teacher quality. Teacher quality varies widely, and a child assigned to a more effective teacher has better learning opportunities. Whether because of a less favorable environment or a less effective teacher, some children are disadvantaged by circumstances beyond their control.

Helping children overcome these disadvantages is essential for creating equal opportunities (*i.e.*, leveling the playing field), a valued concept in Western society. When considering the concept of equality of opportunity, it is important to keep two things in mind (*cf.* Roemer 2000). First, while individuals eventually bear responsibility for their own achievement, it is generally agreed that this does not apply to children. Second, children should not be held responsible for circumstances beyond their control. In the case of skill formation, one such circumstance is the home environment as influenced by the parents' (in)actions; another is the classroom environment provided by the teacher to whom the child is assigned.

The educational disadvantages caused by these early environmental influences can be profound. Figure 1 plots average (standardized) scores in mathematics by family income quartile, using the data analyzed in the three chapters of this thesis. The plot in the top panel uses data from the United States (Chapter 2), and the plots at the bottom use data from Norway (Chapter 3 and 4). Figure 1 shows that differences across socioeconomic strata emerge early and persist over time. Such differences may arise because low-income families are less able to allocate resources to child development, causing their children to grow up in less enriched home environments.[3] It is interesting to note that while income inequality is much lower in Norway than in the United States, differences still emerge, though to a lesser extent.

[3]While it is true that children must exert effort, an environment can promote skill-building. For instance, children may perform at a more advanced level when the environment provides structure and support, enabling them to reach the next level of their capabilities (Vygotsky 1978), a process typically referred to as *scaffolding*. Conversely, risk factors such as abuse, homelessness, parental stress, or poverty can produce an unfavorable environment for skill formation (Masten & Coatsworth 1998). The responsible caregivers might simply be constrained (e.g., owing to income, information, stress) in their capacity to foster an environment conducive to child development.

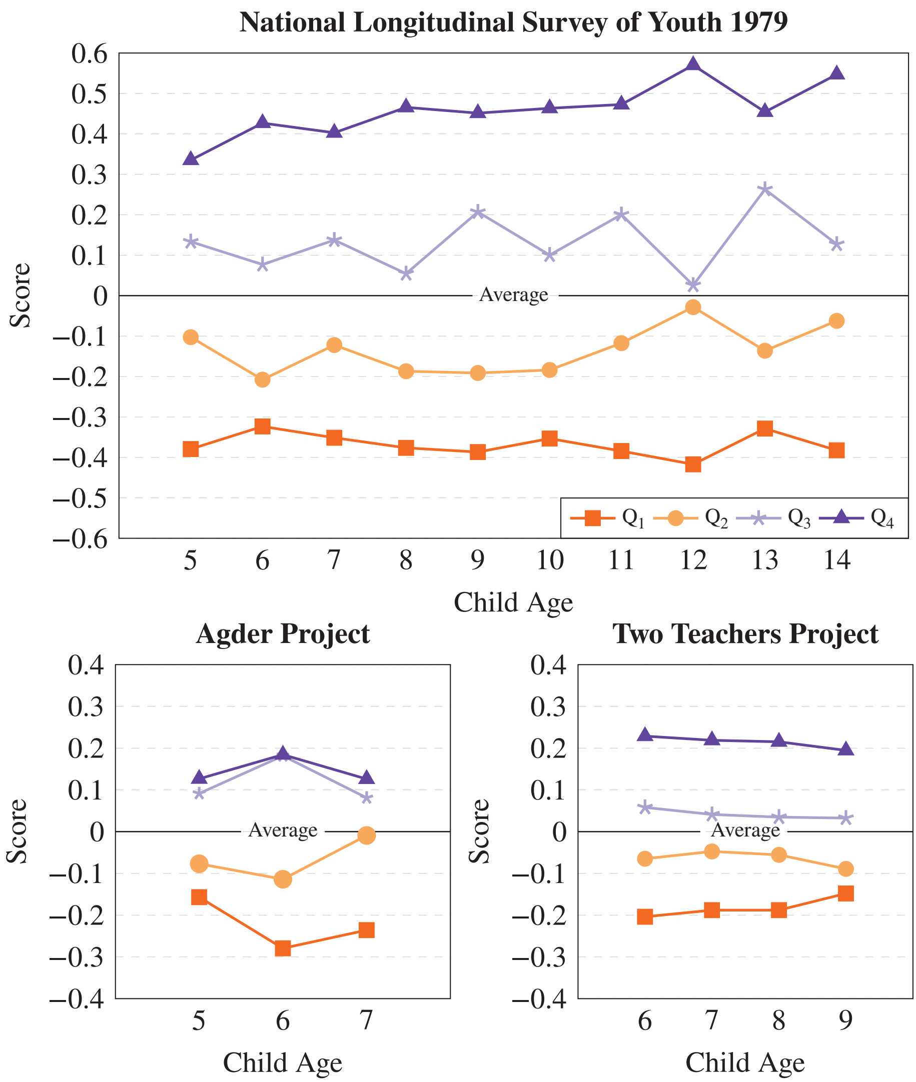

Figure 1. Average Score for Mathematics by Family Income Quartile

Notes. This figure reports average scores for mathematics by family income quartile. The mathematics test used in the National Longitudinal Survey of Youth 1979 is the Peabody Individual Achievement Test for Mathematics (Center for Human Resource Research 2009). The mathematics test used in the Agder Project is the Ani Banani Math Test (Størksen & Mosvold 2013). The mathematics test used in the Two Teachers Project is an arithmetic fact test (Klausen & Reikerås 2016). The test scores have been standardized to have a mean of zero and a standard deviation of one for each age.

Figure 1 is consistent with many other studies documenting gaps in cognitive skills, social-emotional skills, and health (*e.g.*, Adler & Ostrove 1999, Cunha et al. 2006, Duncan & Murnane 2011). These studies demonstrate that environments play a pivotal role in cultivating skill formation. The influential role of the environment in skill formation, combined with the goal of providing equal opportunities (Roemer 2000), has implications for public policy because it suggests a powerful role for investments that enrich these early environments, and hence a child's experiences at home (*e.g.*, Attanasio et al. 2020, Chetty et al. 2016, Sylvia et al. 2020), in ECEC facilities (*e.g.*, Campbell et al. 2014, Conti et al. 2016, Heckman et al. 2013, Rege et al. 2021) and during formal schooling (*e.g.*, Fredriksson et al. 2010, Iversen & Bonesrønning 2013, Schanzenbach 2006).

3. Conceptualizing Child Development

The literature on the economics of human development, which grew out of the early human capital literature (*e.g.*, Becker 1962, 1964), is relevant to the examination of childhood skill formation in the three chapters of this thesis. Human capital embodies the idea that individuals can develop resources in themselves and others through, for instance, schooling, parenting, on-the-job training, healthcare, or migration. Human capital can thus be improved through education, in the broadest sense of the word.[4] The term "human capital" is a hypernym, in that it comprises attributes such as cognitive skills (*e.g.*, phonetic coding, quantitative reasoning), social-emotional skills (*e.g.*, assertiveness, straightforwardness), and health capital (*e.g.*, physical and mental health).[5] These attributes may be synergistic. For instance, the ability to manage stress may positively affect mental health. Enhanced mental health may, in turn, boost mental energy, thereby stimulating cognitive growth. Taken together, these attributes represent the individual's total stock of skills.

[4]Note that childhood human capital enhances later learning efficiency. It differs from the concept of human capital first proposed in the early human capital literature (*e.g.*, Becker 1962), which involves enhancing workers' productivity.

[5]Social-emotional skills are often labeled "noncognitive skills," setting them apart from "cognitive skills." Noncognitive skills are a misnomer, however. Few human behaviors are devoid of cognition, as Borghans et al. (2008) rightly note. Attempts to justify this dichotomy lack a scientific basis (Howard-Jones et al. 2012).

I begin by introducing some notation and terminology. At age t (for $t = 0, 1, \ldots, T$), child i (for $i = 1, \ldots, N$) possesses a stock of skills, denoted by the vector $\mathbf{S}_{i,t}$.[6] Childhood lasts for T years. A period of adult working life may follow. As children mature, the stock of skills (and its dimensions) may change. Let H_t denote the dimensions of the skill stock at age t. The skill formation technology proposed in Cunha & Heckman (2007) formalizes how this skill stock evolves as a function of investments, denoted by vector $\mathbf{I}_{i,t}$, and the stock of skills of caregivers, denoted by vector $\mathbf{S}_{\mathrm{P},i,t}$,

$$\mathbf{S}_{i,t+1} = \mathbf{f}_k(\mathbf{S}_{i,t}, \mathbf{I}_{i,t}, \mathbf{S}_{\mathrm{P},i,t}), \tag{1}$$

where subscript k (for $k = 1, \ldots, K$) on $\mathbf{f}_k(\cdot)$, indexes a developmental stage with $K \leq T$.[7] Note that the function that maps the inputs ($\mathbf{S}_{i,t}$, $\mathbf{I}_{i,t}$, $\mathbf{S}_{\mathrm{P},i,t}$) onto $\mathbf{S}_{i,t+1}$ can be skill-specific.

The function in Equation (1) can take a variety of forms. For example, Cunha & Heckman (2008) consider a log-linear production function, Cunha et al. (2010) use a constant elasticity of substitution specification, and Agostinelli & Wiswall (2016) use a trans-log production specification. Each child i is assumed to be observed at time t and time $t+1$, thereby producing equally spaced intervals, to implement the estimation of these specifications. However, owing to the logistics involved in assessing many geographically dispersed children, intervals are unlikely to be equally spaced. Consequently, estimating empirical specifications of Equation (1) requires addressing unequally spaced intervals. In Chapter 2, I study the implications of not accounting for unequally spaced intervals.

We can conceive of investments and the caregivers' stock of skills as factors that influence the environment a child experiences. The investments in Equation (1) are relevant for policy, such as gathering evidence on which skills to target (Chapter 3) and on teacher effectiveness (Chapter 4). These investments may enrich the environmental influences children experience by improving their quality. The literature generally distinguishes between *structural* and *process* quality. Structural quality refers to regulated aspects of the environment (*e.g.*, the number of educational toys that are available).

[6] Other forms of human capital can include appearance, appropriate credentials, and reputation (see, *e.g.*, Becker & Tomes 1986, p. S6).

[7] Function, $\mathbf{f}_k(\cdot)$, is assumed to be twice continuously differentiable, increasing in its arguments, and concave in $\mathbf{I}_{i,t}$ (Cunha & Heckman 2007).

Process quality refers to the quality of influences a child experiences daily (*e.g.*, the quality of adult-child interactions and curricula).[8]

In Chapter 3, my coauthors and I consider such an investment in process quality. In particular, we study the Agder Project's structured curriculum, which consists of a variety of age-appropriate activities that target skill formation (Rege et al. 2021). In Chapter 4, we also consider an investment in process quality; the relationships the teacher creates with and among the students, as perceived by the students. In preschool and school, process quality seems to be more closely related to the development than structural equality (Blau & Currie 2006, Hamre & Pianta 2001, Hanushek & Rivkin 2006, Pianta 1997).

The technology of skill formation formalized in Equation (1) has two key features: (i) self-productivity and (ii) complementarity. The first, self-productivity, captures the idea that skills may reinforce themselves and cross-fertilize (or cross-produce) each other. Therefore, self-reinforcement involves skills that are alike, whereas cross-fertilization involves unalike skills. Consider a child's cognitive skills. Cognitive skills at age t may self-reinforce cognitive skills at age $t+1$ and cross-fertilize social-emotional skills at age $t+1$.

We can formally define self-productivity as a partial derivative,

$$\underbrace{\frac{\partial \mathbf{S}_{i,t+1}}{\partial \mathbf{S}_{i,t}} > 0.}_{\text{Self-productivity}} \tag{2}$$

The positive marginal effect defined in Equation (2) has three implications. First, if skills are self-reinforcing, then investments will not fully depreciate over a given period of time, all else being equal. Second, the stronger the self-reinforcement, the more stable the rank order of children from one age to the next, all else being equal. Lastly, if skills cross-fertilize, then investments in any particular skill will produce spillover effects on other skills, all else being equal.

[8] A measure of the quality of the home environment is the Home Observation for Measurement of the Environment. This captures the quantity and quality of stimulation and support at home (Bradley & Caldwell 1980). Empirical studies show that this measure explains a large share of the gaps between socioeconomic strata (Cunha et al. 2006).

The second feature is (direct) complementarity. Complementarity captures the idea that an investment in children with *higher* skill levels is more productive. The opposite of complementarity is substitutability. When an investment substitutes, it is said to "compensate" for *lower* skill levels. The following cross-partial derivatives can formally define these concepts,

$$\underbrace{\frac{\partial^2 \mathbf{S}_{i,t+1}}{\partial \mathbf{S}_{i,t} \partial \mathbf{I}'_{i,t}} > 0}_{\text{Complementarity}} \quad \text{and} \quad \underbrace{\frac{\partial^2 \mathbf{S}_{i,t+1}}{\partial \mathbf{S}_{i,t} \partial \mathbf{I}'_{i,t}} < 0.}_{\text{Substitutability}} \tag{3}$$

Intuitively, complementarity represents a relationship in which stocks of skills and investments enhance each other's qualities. By contrast, substitutability implies replaceability; one can replace (or compensate for) *lower* skill stocks with *higher* investment levels.

Self-productivity and complementarity (substitutability) jointly form dynamic complementarities (substitutions). If *current* investments are more effective at producing *future* stocks of skills because of investments made in the *past*, then dynamic complementarities exist. The opposite is true for dynamic substitutes. The following cross-partial derivative defines these concepts formally,

$$\underbrace{\frac{\partial^2 \mathbf{S}_{i,t+r+1}}{\partial \mathbf{I}_{i,t} \partial \mathbf{I}'_{i,t+r}} > 0}_{\substack{\text{Dynamic} \\ \text{Complementarity}}} \quad \text{and} \quad \underbrace{\frac{\partial^2 \mathbf{S}_{i,t+r+1}}{\partial \mathbf{I}_{i,t} \partial \mathbf{I}'_{i,t+r}} < 0,}_{\substack{\text{Dynamic} \\ \text{Substitutability}}} \quad r \geq 1. \tag{4}$$

Dynamic complementarity implies an equity-efficiency tradeoff for late as opposed to early investments (Cunha & Heckman 2007). By contrast, dynamic substitutability implies that one can compensate for low investments in the past by investing more in the present.

These features have important implications for understanding which skills to target, as explained in Chapter 3. Assume that some target skill (say executive functions) in preschool is cross-productive for other primary-school skills (say mathematical and language skills). Further assume that the higher levels of executive functions we observed at the start of primary school in Chapter 3 resulted from an investment in preschool (Rege et al.

2021). Investments made in primary school may complement executive functions in promoting other skills. If so, then a dynamic complementarity may exist because current investments (*i.e.*, those made in primary school) are becoming more effective at producing mathematical and language skills because of investments made in the past (*i.e.*, those made in preschool). Executive functions would then be the mechanism by which this dynamic complementarity is achieved. Evidence on such dynamic complementarities can inform public policy and the design of preschool education programs to ensure that all children are ready to learn at the time of school entry.

This section has briefly described the skill formation technology, self-productivity, and complementarity, representing the key theoretical constructs and relations of interest. See Cunha et al. (2006), Cunha & Heckman (2007, 2009), and Heckman & Mosso (2014) for additional details concerning the technology of skill formation. See also Bailey et al. (2020) for a description of how the concepts outlined in this section can explain the persistence and fade-out of (educational) interventions. In the next section, I note that the stock of skills is not necessarily (directly) observable, which is a challenge that comes up in each of the chapters. Investments, too, may not be directly be observable (*e.g.*, the quality of adult-child interactions).[9] In this introductory chapter, however, I focus only on the unobservability of skills.

4. Skills are Inherently Unobservable

The child's stock of skills and the model features described in Section 3 (*i.e.*, self-productivity and complementarity) represent the main theoretical constructs and relationships of interest. However, the child's skill stock is not necessarily (directly) observable. All we can hope to observe are

[9]Cunha & Heckman (2008) and Cunha et al. (2010) present an economic model that rationalizes measurement models for investments as derived demand equations. The parent's objective is to maximize lifetime utility, which depends on consumption and the child's adult stock of skills. In the first stage, parents decide how much to invest, based on their endowment and earnings as well as the child's stock of skills. In the second stage, they decide how much of the inputs to buy to achieve the desired level of investment. See Appendix 1 in Cunha & Heckman (2008) and Web Appendix A2 in Cunha et al. (2010) for more details.

manifestations, which we then assume to be consistent with a particular level of proficiency.

As a case in point, consider a child's motor skills (*i.e.*, the ability move body muscles to perform tasks). We can observe how a child walks, jumps, and runs, but not the motor skills themselves. As a result, problems arise when we attempt to test hypotheses concerning the theoretical constructs and relationships of interest, since empirics involve *observable* data. Rules of correspondence between the theoretical and empirical relate to the process of measurement (Torgerson 1958). Accordingly, I provide a brief discussion concerning measurement that is relevant to the economics of human development literature and the chapters that make up this thesis.

Section 4.1 presents a general model with a structure for conceptualizing the measurement models in Chapters 2 through 4. Section 4.2 briefly discusses how studies may produce evidence to support measurement strategies. The measures used in Chapters 2 through 4 have been validated in earlier work. Section 4.2 provides insight into how this validation is generally accomplished. In Section 4.3, I consider a dedicated measurement system of the type used in each chapter. Section 4.4 briefly discusses a common alternative to measurement modeling. For additional details on measurement modeling, see Almlund et al. (2011), Borghans et al. (2008), or Cunha et al. (2021), as well as the literature they cite.

4.1. A Model of Task Performance

Measuring skill formation in childhood is challenging. Researchers use three broad approaches when assessing child development. The first is to examine the interactions between a (trained) tester and the child, such as in the Ani Banani Math Test (Størksen & Mosvold 2013) used in Chapter 3. Second, observers (*e.g.*, caregivers, teachers) may report on skills or behaviors (*e.g.*, behavior problems index: Peterson & Zill 1986). Another case in point is Chapter 4, where we ask students to report on the relationship skills of the teacher. Third, children may self-report on their own skills or behaviors (*e.g.*, the reading self-concept used in Chapter 4: Chapman et al. 2000, Chapman & Tunmer 1995). All of these approaches are inevitably vulnerable to errors of measurement, given the difficulty of assessing (young) children.

In every case, the child performs some task (or action).[10] Performance on that task is then scored, and this allows us to draw conclusions about the child's level of proficiency. Let the variable $\mathrm{M}_{l,i,t}$ denote the *observed* performance on task l (for $l = 1, \ldots, L_t$) at age t, where L_t denotes the total number of tasks performed at age t. We assume that the observed outcome on task l is, at least in part, a manifestation of a child's stock of skills, which is *latent* (or unobserved). As $\mathrm{M}_{l,i,t}$ manifests the child's stock of skills, some refer to it as a *manifest variable*. Performance on task l may depend not only on the stock of skills, but also on other factors and measurement error, denoted by $\zeta_{l,i,t}$, which could be a vector.

We can define a model of task performance for task l performed at age t by child i as a function of the child's stock of skills and other factors (including measurement error),

$$\mathrm{M}_{l,i,t} = g_{l,t}(\mathrm{S}_{1,i,t}, \mathrm{S}_{2,i,t}, \ldots, \mathrm{S}_{H_t,i,t}, \zeta_{l,i,t}), \tag{5}$$

where H_t denotes the total number of skills at age t. One important implication of Equation (5), among others, is that the child's performance on task l is determined by a vector of skills, $\mathbf{S}_{i,t} = (\mathrm{S}_{1,i,t}, \mathrm{S}_{2,i,t}, \ldots, \mathrm{S}_{H_t,i,t})$, each of which can affect performance on task l differently at any given age t. Identifying a particular skill thus requires that performance on certain tasks can be *exclusively* attributed to that skill (*i.e.*, dedicated measures); otherwise, it is impossible to determine which skills produced the outcome on task l. The empirical studies discussed in Chapters 2 through 4 employ such validated measures. As all empirical studies in this thesis work with validated measures, I provide a brief discussion on the process of validating measurement strategies in the next section.

4.2. Validating Measurement Strategies

The inherent unobservability of skill stocks leads directly to the analysis of latent variables, factor models, and the econometrics of errors of

[10] We can interpret tasks more broadly, however. Tasks do not necessarily need to be designed by a researcher, as in the examples above. They might include achievement tests, survey questionnaires, enrollment in advanced mathematics courses, participation in extracurricular activities, and the like (see Almlund et al. 2011, Borghans et al. 2008, for more examples).

measurement (Bollen 1989, Skrondal & Rabe-Hesketh 2004, Wansbeek & Meijer 2000). Assuming Equation (5) is additively separable, we can propose the following multiple linear factor model for task l performed at age t by child i:

$$\mathrm{M}_{l,i,t} = \mu_{l,t} + \lambda_{1,l,t}\mathrm{S}_{1,i,t} + \lambda_{2,l,t}\mathrm{S}_{2,i,t} + \cdots + \lambda_{H_t,l,t}\mathrm{S}_{H_t,i,t} + \zeta_{l,i,t}, \qquad (6)$$

where the intercept, $\mu_{l,t}$, captures the mean performance on task l at age t. Since the child's stock of skills, $\mathbf{S}_{i,t} = (\mathrm{S}_{1,i,t}, \mathrm{S}_{2,i,t}, \ldots, \mathrm{S}_{H_t,i,t})$, is a H_t dimensional vector, the vector of coefficients, $\boldsymbol{\lambda}_{l,t} = (\lambda_{1,l,t}, \lambda_{2,l,t}, \ldots, \lambda_{H_t,l,t})$, can have as many as H_t nonzero factor loadings. These factor loadings measure the correlation between the performance on task l and any particular skill at age t. Skills that are more important for performance on task l have higher factor loadings. The last term, $\zeta_{l,i,t}$, captures other factors and errors of measurement (*i.e.*, "uniqueness"). Any correlation across tasks can arise because tasks depend on the same vector of skills, but to varying degrees.

Three pieces of evidence commonly support the validity of a set of tasks designed to measure a particular skill: *content-related validity*, *construct-related validity*, and *criterion-related validity* (VandenBos 2007). The most basic form of validity is content-related validity, a term that refers to the degree to which the content, as defined by experts, is well captured by the task (or action). Establishing construct- and criterion-related validity requires the use of factor analytic methods.

Construct-related validity has two components: *convergent* and *discriminant* validity. The former concerns whether a particular battery of tasks relates strongly as a group (*i.e.*, high intercorrelation), which is what one might expect if the battery of tasks is presumed to measure a particular skill. The factor loadings in Equation (6) capture the correlation between *unobserved* skills and *observed* performance on tasks. As convergent-related validity refers to the degree to which a battery of tasks correlates with a particular skill, the factor loadings should be high. In contrast, discriminant-related validity examines the degree to which tasks diverge from other tasks whose underlying skill is (conceptually) different. For example, consider the first skill in Equation (6), $\mathrm{S}_{1,i,t}$. Discriminant-related validity for $\mathrm{S}_{1,i,t}$ requires that the corresponding factor loading, $\lambda_{1,l,t}$, is the only nonzero component.

Criterion-related validity evaluates how well a battery of tasks that measure a skill correlates with established standards. An example of criterion-related validity is predictive validity: How well do tasks predict future (real-world) outcomes? Relative to content- and construct-related validity, predictive validity is more difficult to establish because of reverse causality and measurement error, as explained in Borghans et al. (2008) and Almlund et al. (2011).

Often, in empirical settings with multiple measures, we have some *a priori* knowledge of which skills affect which tasks. For example, the Ani Banani Math Test (Størksen & Mosvold 2013) used in Chapter 3 is designed to assess children's (early) mathematical skills, as substantiated in ten Braak & Størksen (2021). In such cases, we can assume a *dedicated system of measurements*, in which each task proxies only one skill. The point of departure in the next section is a dedicated measurement system used in all of the chapters in this thesis. In this section, I briefly discuss the main intuition underlying identification.

4.3. Dedicated Measurement System

I can represent a dedicated measurement system as a one-factor model for task l performed at age t by child i,

$$\mathrm{M}_{l,i,t} = \mu_{l,t} + \lambda_{l,t}\mathrm{S}_{i,t} + \zeta_{l,i,t}, \tag{7}$$

where $\mu_{l,t}$ denotes the intercept, $\lambda_{l,t}$ denotes the factor loading, $\mathrm{S}_{i,t}$ denotes the *unobserved* skill of interest, and $\zeta_{l,i,t}$ denotes the error term (or unique factor). Equation (7) could include a vector of observable covariates (see, *e.g.*, Skrondal & Rabe-Hesketh 2004, Williams 2020).

Three assumptions are typically made for the error term.[11] First, we assume that the error term is mean zero for all tasks and independent across children and over time. Second, in the case of every task the error term is independent of the unobserved skill (*i.e.*, classical measurement error). Lastly, the error term of task l at age t is independent of the error term of task l' at age t', where $l \neq l'$, conditional on the unobserved

[11]Factor analysis differs from principal component analysis because of this *a priori* structure of the error terms. Principal component analysis represents a singular value decomposition of an association matrix.

skill. Intuitively, we assume that the skill is causing the correlation across these (dedicated) tasks. Conditional on the skill, the error terms should thus be independent. As the right-hand-side variables in Equation (7) are unobserved, we require some normalizations to set a scale and location for the unobserved skill.

Location. One can normalize one of the intercepts (say the first) to zero, $\mu_{1,t} = 0 \; \forall \; t = 0, 1, \ldots, T$. The mean of the unobserved skill can then be identified, $\mathbb{E}(\mathrm{S}_{i,t}) = \mu_{\mathrm{S},t} \; \forall \; t = 0, 1, \ldots, T$, where $\mathbb{E}(\cdot)$ denotes the expectation operator. Alternatively, one can normalize the mean of the unobserved skill to zero, $\mathbb{E}(\mathrm{S}_{i,t}) = 0 \; \forall \; t = 0, 1, \ldots, T$, so that all intercepts are identified. Finally, one can normalize the sum of the intercepts to zero, $\mu_{1,t} + \mu_{2,t} + \cdots + \mu_{H_t,t} = 0$, so that all intercepts and the mean of the unobserved skill can be identified.

Scale. One can normalize one of the factor loadings (say the first) to one, $\lambda_{1,t} = 1 \; \forall \; t = 0, 1, \ldots, T$. The scale of the unobserved skill is now "anchored" in the first task. The unobserved skill variance is then identifiable, $\mathrm{Var}(\mathrm{S}_{i,t}) = \sigma_{\mathrm{S},t} \; \forall \; t = 0, 1, \ldots, T$, where $\mathrm{Var}(\cdot)$ denotes the variance operator. Alternatively, one can normalize the variance of the unobserved skill to one, $\mathrm{Var}(\mathrm{S}_{i,t}) = 1$, so that all factor loadings are identified. While these identification restrictions result in equivalent models, the former is preferred for "factorial invariance" (see, *e.g.*, Skrondal & Rabe-Hesketh 2004, and the literature they cite). Finally, one normalize the sum of the factor loadings to the total number of tasks, $\lambda_{1,t} + \lambda_{2,t} + \cdots + \lambda_{H_t,t} = L_t$, so that all factor loadings and the variance of the unobserved skill are identified.

Below, I consider a scenario in which we normalize one of the intercepts (say the first) to zero, $\mu_{1,t} = 0$, and one of the factor loadings (say also the first) to one, $\lambda_{1,t} = 1$. Under these normalizations and the assumptions made previously, one can identify the factor loadings and the distribution of factors. The identification requires a minimum of three tasks (*i.e.*, $L_t \geq 3$). If multiple periods are available, then a minimum of two tasks in each period is sufficient (see Cunha & Heckman 2008, Cunha et al. 2010). I consider the former case, in which we observe a minimum of three tasks. Under the imposed normalizations and assumptions, I can write the following covariances:

$$\mathrm{Cov}(\mathrm{M}_{1,i,t}, \mathrm{M}_{l,i,t}) = \lambda_{l,t}\mathrm{Var}(\mathrm{S}_{i,t}), \tag{8}$$

$$\mathrm{Cov}(\mathrm{M}_{l,i,t}, \mathrm{M}_{l',i,t}) = \lambda_{l,t}\lambda_{l',t}\mathrm{Var}(\mathrm{S}_{i,t}), \tag{9}$$

for $l, l' = 2, 3$, with $l \neq l'$. By taking the ratio of Equation (9) to Equation (8), one can identify the factor loadings,

$$\frac{\mathrm{Cov}(\mathrm{M}_{l,i,t}, \mathrm{M}_{l',i,t})}{\mathrm{Cov}(\mathrm{M}_{1,i,t}, \mathrm{M}_{l,i,t})} = \frac{\lambda_{l,t}\lambda_{l',t}\mathrm{Var}(\mathrm{S}_{i,t})}{\lambda_{l,t}\mathrm{Var}(\mathrm{S}_{i,t})} = \lambda_{l',t}, \tag{10}$$

for $l, l' = 2, 3$, with $l \neq l'$. With the factor loadings identified, one can (nonparametrically) identify the distribution of the factors by applying Kotlarski's lemma (Kotlarski 1967); see, for example, Cunha & Heckman (2008) and Hansen et al. (2004). Once the model is identified, estimation follows standard procedures for factor models (see, *e.g.*, Wansbeek & Meijer 2000).[12]

These identification results apply when the tasks we observe have an *interval* scale. An interval scale is such that the differences between points on the scale are equal (Torgerson 1958). Tasks may not have such a cardinal scale, however. A common example of a noncardinal scale is the Likert scale, which is ordinal. An ordinal scale is such that the distance between points on the scale is not equal. See (Torgerson 1958, pp. 15–21) for more information concerning these and related scales. An ordinal scale tells us that a particular point dominates another point, but not by how much. As a result, observed differences may turn out to be trivial. What is more, if the scale is ordinal, identification requires additional assumptions and normalizations (Skrondal & Rabe-Hesketh 2004).

[12]One can estimate the system of equations jointly, using parametric maximum likelihood estimation, by assuming normality. Alternatively, one can use instrumental variable approaches such as two-stage least squares (see, *e.g.*, Madansky 1964). Note that we can write Equation (7) as $\mathrm{S}_{i,t} = \mathrm{Y}_{1,i,t} - \mu_{1,t} - \zeta_{1,i,t}$. Plugging this in the model for the second task $\mathrm{M}_{2,i,t} = (\mu_{2,t} + \lambda_{2,t}\lambda_{1,t}) + \lambda_{2,t}\mathrm{M}_{1,i,t} + (\zeta_{2,i,t} - \lambda_{2,t}\zeta_{1,i,t})$. Estimating this equation using least squares does not produce consistent estimators of $\lambda_{2,t}$ because $\mathrm{M}_{1,i,t}$ correlates with the error term. However, under the imposed assumptions, we can use $\mathrm{M}_{3,i,t}$ as instrument and estimate using two-stage least squares. Analogously, we can consistently estimate $\lambda_{3,t}$ using $\lambda_{2,t}$ as an instrument.

An alternative approach to handling the noncardinality of scales is to estimate relationships between the observed task performance and a cardinal anchor outcome, resulting in a plausible interval scale. Cunha & Heckman (2008) and Cunha et al. (2010) take such an approach. Cunha et al. (2021) provides further details on anchoring. While intuitive, such forward-linking approaches also raise questions (Jacob & Rothstein 2016).

4.4. An Alternative Approach to Measurement Modeling

An alternative to measurement modeling is to use sum scores or averages of, for example, (standardized) test scores as a proxy for skills. Such "*measurement by fiat*" (Torgerson 1958, p. 22) is not recommended, however. First, it cannot be theoretically motivated (Skrondal & Rabe-Hesketh 2004). Second, tasks differ in the degree to which they provide information concerning the skills they measure (Cunha & Heckman 2008, Cunha et al. 2010), but sum scores or averages use arbitrary weights. Third, any task is imperfect (Borghans et al. 2008). The average only accounts for measurement error through simple averaging. Fourth, we cannot partial out other observed influences. In conclusion, it is not desirable to reject measurement modeling.

5. Summary of Chapters

In the introduction, I explained why society might invest in resources to foster socially productive skills. This objective raises questions. First, how can we best learn about the effectiveness of these investments in childhood skill formation? The study in Chapter 2 relates to this question (summarized in Section 5.1). Second, which of the many possible skills should be nurtured? This question is addressed in the study in Chapter 3 (Section 5.2). Lastly, how can we sustain the skills fostered in early childhood by providing effective education? The study in Chapter 4 relates to this question (Section 5.3).

5.1. Cognitive Skill Production and Unequal Intervals

Chapter 2 investigates an econometric challenge faced by researchers seeking to evaluate the effectiveness of investments within a model of skill formation (Equation 1). Several studies, most notably Todd & Wolpin (2003), Cunha & Heckman (2008), Cunha et al. (2010), and Agostinelli & Wiswall (2016), highlight econometric challenges that researchers who want to estimate the technology of skill formation need to acknowledge (see Cunha et al. 2021, for a recent survey).

This chapter addresses the challenge of "unequally spaced intervals," which has received little attention in the literature on skill formation in childhood. A straightforward interpretation of parameters may not be possible when intervals are unequally spaced. Moreover, inputs from *unobserved* periods may correlate with inputs from *observed* periods. In short, failing to account for unequally spaced intervals can result in biased estimation.

The observation that unequally spaced intervals are problematic is not new (see, *e.g.*, Baltagi & Wu 1999, McKenzie 2001, Millimet & McDonough 2017, Rosner & Muñoz 1988, Sasaki & Xin 2017). However, the approach in this chapter differs in at least three ways from these studies. First, I consider an economic model of skill formation with unobserved (or latent) dependent and independent variables based on Section 3. Second, the earlier papers presume that unequally spaced intervals are the result of the survey design. In other words, we observe all children in the same unequally spaced intervals. Suppose that the substantive timing variable of child development is age, however. In that case, unequally spaced intervals may occur even when survey waves are equidistant, as children can vary in age in any given wave. Third, following Jones & Boadi-Boateng (1991) and Voelkle & Oud (2013), I apply insights from the continuous-time modeling literature.[13]

In particular, I implement the exact discrete model introduced by Albert R. Bergstrom (Bergstrom 1988). Discrete-time model parameters relate *exactly* to continuous-time model parameters through the exact discrete model for a particular observation interval. By allowing this interval to be unequally spaced, I can estimate the underlying continuous-time model parameters. Once I obtain those parameters, I can solve the continuous-

[13]The approach taken in Voelkle & Oud (2013) is particularly similar to the one taken in Chapter 2. Voelkle & Oud (2013) use Monte Carlo simulations to study the impact of unequal intervals that vary across individuals on oscillatory and non-oscillatory processes.

time model for the interval *assumed* by the (discrete-time) model, but taking into account the unequally spaced intervals *observed* in the data. By extension, we arrive at the typically estimated parameters if the observation interval is defined as equally spaced. It follows that we can compare estimates obtained under equally and unequally spaced intervals.

To investigate the impact of (child-specific) unequal intervals, I analyze a frequently used dataset for studying children's skill formation: the National Longitudinal Survey of Youth 1979. I compare parameter estimates obtained for equally and unequally spaced intervals. Estimates based on the former can differ greatly from those based on the latter. Furthermore, the level of precision is generally higher for the self-reinforcement of skill when I account for the unequally spaced intervals. In conclusion, the study described in Chapter 2 suggests that (child-specific) unequal intervals are another challenge that researchers should acknowledge when estimating (early) childhood skill formation (*cf.* Cunha et al. 2021).

5.2. Cross-Productivities of Executive Functions

Chapter 3 investigates whether executive functions developed in early childhood lead to improvements in mathematical skills and language skills in primary school. This chapter thus pertains directly to one of the key features of the technology of skill formation outlined in Section 3.

The home environment a child experiences during the early years can provide disparate opportunities. A child's development may lag when the home environment is not conducive to fostering skill formation. As a result, children enter school with different skill levels (Duncan & Murnane 2011). Insight into such skill disparities has prompted an interest in early childhood education programs that boost "school readiness" (Clements & Sarama 2011, Diamond & Lee 2011, Dillon et al. 2017, Rege et al. 2021). One challenge in designing these programs is deciding what skills to target, as not all skills may be equally beneficial for school success (*e.g.*, Duncan et al. 2007). Evidence on what skills to target might help to inform public policy on the design of preschool education programs, in an effort to ensure that every child is ready to learn at the start of formal schooling.

The conceptual model in Section 3 is useful for understanding the economic argument for focusing on "school readiness." Consider the case in which a particular investment enriches the preschool environment in a way that improves a certain skill. Assume that the targeted skill has a favorable impact on other skills in primary school. If investments made in primary school (*e.g.*, teacher quality, routines) complement the targeted skill in promoting these other skills, then dynamic complementarities may exist. *Current* investments (primary school) become more effective in producing *future* (other) skills because of investments made in the *past* (preschool). The mechanism that gives rise to this dynamic complementarity is higher levels of the targeted skill.

Several studies suggest targeting executive functions as a key skill because they are fundamental to learning (*e.g.*, Blair 2002, Diamond & Lee 2011, Howard-Jones et al. 2012). Building on this hypothesis, we ask the following question: Do children with higher levels of executive functioning at the start of primary school develop more advanced mathematical and language skills? We provide evidence on this question by combining high-quality experimental data from the Agder Project (Rege et al. 2021), the economic model of skill formation (see Section 3), and the econometric decomposition framework in Heckman et al. (2013).

First, we find that program-induced gains in children's executive functions lead to improvements in mathematical and language skills in primary school. We also find that the proportions of mathematics and language skills that are attributable to executive functions increase over time. It therefore appears that executive functions become increasingly important as children progress through school. Finally, we provide empirical evidence that highlights executive functions as a fundamental skill, one that policies aimed at improving school readiness may want to target.

5.3. Teacher Relationship Skills and Student Learning

Chapter 4 investigates whether understanding a teacher's capacity to form positive relationships with and among students improves our understanding of how effective educational environments can be replicated. We refer to this capacity as teacher relationship skills. While there are many kinds of school inputs (*e.g.*, class size, number of books), we focus specifically on the teacher. Teacher effectiveness may vary widely, even in the same school (Aaronson et al. 2007, Araujo et al. 2016, Jackson 2018, Rivkin et al. 2005, Rockoff 2004). Moreover, effective teachers may have long-term impacts on students' education and labor market outcomes

(Chetty et al. 2014*b*, Opper 2019). Lastly, teachers are the largest budgetary expense in most schools (Hanushek & Rivkin 2006). Despite this extensive evidence on the value added by teachers, we need more research to gain a better understanding of why some teachers are more effective than others at promoting student learning. We are able to *identify* effective teachers, but we cannot *replicate* them.

The child development literature suggests that the teacher-child relationship and the relationships among classmates correlate with social, emotional, and academic development (Hamre & Pianta 2001, Parker & Asher 1987). Forming positive and avoiding negative relationships with and among students is ultimately the teacher's responsibility. These positive relationships may create an environment in which children feel competent, independent, and akin to others, which may increase their motivation to learn (Connell & Wellborn 1991). By contrast, negative interactions (*e.g.*, yelling) may result in emotional distress, leading to distractions and behavioral challenges (Parker & Asher 1987, Pianta 1997).

Studies in psychology and education science have examined teacher's relationship skills as perceived by students. The parameter estimates reported in these studies may be biased because of students' (unobserved) preferences for a particular type of relationship. Recently, economists have started to utilize classroom observations to measure teacher practices in a way that is less affected by such idiosyncrasies (Araujo et al. 2016, Kane et al. 2011).

However, these classroom observations are costly and may fail to capture the perceptions that ultimately drive student behavior (Connell & Wellborn 1991). Naturally, self-reports are not without limitations; they rely on the student's honesty, and students may also lack the capacity for introspection or may not fully understand the questions. Accordingly, both approaches have limitations. Nevertheless, both measurements are valuable, as it is important to rely on a variety of assessments when evaluating teachers (Kane & Staiger 2012).

In this context, Chapter 4 introduces and validates a new approach to measuring teachers' overall capacity to form positive relationships with and among students. To measure teacher relationship skills, we asked the students a broad set of questions that capture several dimensions of the teachers' ability to form such positive relationships. The questions we use have previously been validated (at the student level) in the psychology and education literature. We use a leave-out-mean specification to account for the bias that arises from students' preferences for a particular type of relationship.

We find that teacher relationship skills are highly stable over time. A degree of stability is what one would expect of teacher quality, as Chetty et al. (2014*a*) point out, and this is something typically assumed in policies addressing the issue of teacher quality (Goldhaber & Hansen 2010). Second, there is not only substantial variation in teacher quality, as measured by learning outcomes conditional on past achievement, but also in teacher relationship skills. Substantial variation is consistent with the value-added literature (Hanushek & Rivkin 2010). Finally, assuming (as-good-as) random class assignment of students, we show that teacher relationship skills affect various academic and social-emotional skills. These findings may inform policy discussions concerning the skills schools should focus on when hiring and evaluating teachers.

References

Aaronson, D., Barrow, L. & Sander, W. (2007), 'Teachers and student achievement in the Chicago public high schools', *Journal of Labor Economics* **25**(1), 95–135.

Acemoglu, D. & Autor, D. (2011), Skills, tasks and technologies: Implications for employment and earnings, *in* D. Card & O. Ashenfelter, eds, 'Handbook of Labor Economics', Vol. 4, Chapter 12, pp. 1043–1171.

Adler, N. E. & Ostrove, J. M. (1999), Socioeconomic status and health: What we know and what we don't, *in* 'Annals of the New York Academy of Sciences', Vol. 896, New York Academy of Sciences, pp. 3–15.

Agostinelli, F. & Wiswall, M. (2016), Estimating the technology of children's skill formation, Working Paper 22442, National Bureau of Economic Research, Cambridge, Massachusetts.

Almlund, M., Duckworth, A. L., Heckman, J. & Kautz, T. (2011), Personality psychology and economics, *in* 'Handbook of the Economics of Education', Vol. 4, Elsevier B.V., Chapter 1, pp. 1–181.

Almond, D. & Currie, J. (2011), Human capital development before age five, *in* D. Card & O. Ashenfelter, eds, 'Handbook of Labor Economics', Vol. 4, Chapter 15, pp. 1315–1486.

Araujo, M. C., Carneiro, P., Cruz-Aguayo, Y. & Schady, N. (2016), 'Teacher quality and learning outcomes in kindergarten', *Quarterly Journal of Economics* **131**(3), 1415–1453.

Attanasio, O., Cattan, S., Fitzsimons, E., Meghir, C. & Rubio-Codina, M. (2020), 'Estimating the production function for human capital: Results from a randomized controlled trial in Colombia', *American Economic Review* **110**(1), 48–85.

Autor, D. H., Levy, F. & Murnane, R. J. (2003), 'The skill content of recent technological change: An empirical exploration', *Quarterly Journal of Economics* **118**(4), 1279–1333.

Bailey, D., Duncan, G. J., Odgers, C. L. & Yu, W. (2017), 'Persistence and fadeout in the impacts of child and adolescent interventions', *Journal of Research on Educational Effectiveness* **10**(1), 7–39.

Bailey, D. H., Duncan, G. J., Cunha, F., Foorman, B. R. & Yeager, D. S. (2020), 'Persistence and fade-out of educational-intervention effects: Mechanisms and potential solutions', *Psychological Science in the Public Interest* **21**(2), 55–97.

Baltagi, B. H. & Wu, P. X. (1999), 'Unequally spaced panel data regressions with AR(1) disturbances', *Econometric Theory* **15**(6), 814–823.

Becker, G. S. (1962), 'Investment in human capital: A theoretical analysis', *Journal of Political Economy* **70**(5), 9–49.

Becker, G. S. (1964), *Human Capital: A Theoretical and Empirical Analysis with Special Reference to Education*, 3 edn, The University of Chicago Press, Chicago, Illinois.

Becker, G. S. & Tomes, N. (1986), 'Human capital and the rise and fall of families', *Journal of Labor Economics* **4**(3 Pt. 2), 1–47.

Bergstrom, A. R. (1988), 'The history of continuous-time econometric models', *Econometric Theory* **4**(3), 365–383.

Blair, C. (2002), 'School readiness: Integrating cognition and emotion in a neurobiological conceptualization of children's functioning at school entry', *American Psychologist* **57**(2), 111–127.

Blau, D. & Currie, J. (2006), Pre-school, day care, and after-school care: Who's minding the kids?, *in* E. A. Hanushek & F. Welch, eds, 'Handbook of the Economics of Education', Vol. 2, Elsevier B.V., Amsterdam, Netherlands, Chapter 20, pp. 1163–1278.

Bollen, K. A. (1989), *Structural Equations with Latent Variables*, John Wiley & Sons, Inc., New York City, New York.

Borghans, L., Duckworth, A. L., Heckman, J. J. & ter Weel, B. (2008), 'The economics and psychology of personality traits', *Journal of Human Resources* **43**(4), 972–1059.

Bradley, R. H. & Caldwell, B. M. (1980), 'The relation of home environment, cognitive competence, and IQ among males and females', *Child Development* **51**(4), 1140–1148.

Campbell, F., Conti, G., Heckman, J. J., Moon, S. H., Pinto, R., Pungello, E. & Pan, Y. (2014), 'Early childhood investments substantially boost adult health', *Science* **343**(6178), 1478–1485.

Center for Human Resource Research (2009), *NLSY79 Child and Young Adult Data User's Guide*, Center for Human Resource Research, Ohio State University, Columbus, Ohio.
URL: *https://search.library.wisc.edu/catalog/9910094201202121*

Chapman, J. W. & Tunmer, W. E. (1995), 'Development of young children's reading self-concepts: An examination of emerging subcomponents and their relationship with reading achievement', *Journal of Educational Psychology* **87**(1), 154–167.

Chapman, J. W., Tunmer, W. E. & Prochnow, J. E. (2000), 'Early reading-related skills and performance, reading self-concept, and the development of academic self-concept: A longitudinal study', *Journal of Educational Psychology* **92**(4), 703–708.

Chetty, R., Friedman, J. N. & Rockoff, J. E. (2014*a*), 'Measuring the impacts of teachers I: Evaluating bias in teacher value-added estimates', *American Economic Review* **104**(9), 2593–2632.

Chetty, R., Friedman, J. N. & Rockoff, J. E. (2014*b*), 'Measuring the impacts of teachers II: Teacher value-added and student outcomes in adulthood', *American Economic Review* **104**(9), 2633–2679.

Chetty, R., Hendren, N. & Katz, L. F. (2016), 'The effects of exposure to better neighborhoods on children: New evidence from the moving to opportunity experiment', *American Economic Review* **106**(4), 855–902.

Clements, D. H. & Sarama, J. (2011), 'Early childhood mathematics intervention', *Science* **333**(6045), 968–970.

Connell, J. P. & Wellborn, J. G. (1991), *Competence, Autonomy, and Relatedness: A Motivational Analysis of Self-System Processes*, The

Minnesota Symposia on Child Psychology, Volume 23., Lawrence Erlbaum Associates, Inc, Hillsdale, New Jersey, pp. 43–77.

Conti, G., Heckman, J. J. & Pinto, R. (2016), 'The effects of two influential early childhood interventions on health and healthy behaviour', *Economic Journal* **126**(596), F28–F65.

Crawford, C., Dearden, L. & Meghir, C. (2010), When you are born matters: The impact of date of birth on educational outcomes in England, Working Paper 10,06, Institute for Fiscal Studies, London, England.

Cunha, F. & Heckman, J. (2007), 'The technology of skill formation', *American Economic Review* **97**(2), 31–47.

Cunha, F. & 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. (2009), 'The economics and psychology of inequality and human development', *Journal of the European Economic Association* **7**(2–3), 320–364.

Cunha, F., Heckman, J. J., Lochner, L. & Masterov, D. V. (2006), Interpreting the evidence on life cycle skill formation, *in* E. A. Hanushek & F. Welch, eds, 'Handbook of the Economics of Education', Vol. 1, Elsevier B.V., Amsterdam, Netherlands, Chapter 12, pp. 697–812.

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

Cunha, F., Nielsen, E. & Williams, B. (2021), 'The econometrics of early childhood human capital and investments', *Annual Review of Economics* **13**, 487–513.

Diamond, A. & Lee, K. (2011), 'Interventions shown to aid executive function development in children 4 to 12 years old', *Science* **333**(6045), 959–964.

Dillon, M. R., Kannan, H., Dean, J. T., Spelke, E. S. & Duflo, E. (2017), 'Cognitive science in the field: A preschool intervention durably enhances intuitive but not formal mathematics', *Science* **357**(6346), 47–55.

Duncan, G. J., Dowsett, C. J., Claessens, A., Magnuson, K., Huston, A. C., Klebanov, P., Pagani, L. S., Feinstein, L., Engel, M., Brooks-Gunn, J., Sexton, H., Duckworth, K. & Japel, C. (2007), 'School readiness and later achievement', *Developmental Psychology* **43**(6), 1428–1446.

Duncan, G. J. & Murnane, R. J. (2011), *Whither Opportunity? Rising Inequality, Schools, and Children's Life Chances*, Russell Sage Foundation, New York City, New York.

Elder, T. E. & Lubotsky, D. H. (2009), 'Kindergarten entrance age and children's achievement: Impacts of state policies, family background, peers', *Journal of Human Resources* **44**(3), 641–683.

Fredriksson, P., Öckert, B. & Oosterbeek, H. (2010), Inside the black box of class size: Mechanisms, behavioral responses, and social background, Working Paper 8019, Institute for the Study of Labor, Bonn, Germany.

Goldhaber, D. & Hansen, M. (2010), 'Using performance on the job to inform teacher tenure decisions', *American Economic Review* **100**(2), 250–255.

Hamre, B. K. & Pianta, R. C. (2001), 'Early teacher-child relationships and the trajectory of children's school outcomes through eighth grade', *Child Development* **72**(2), 625–638.

Hansen, K. T., Heckman, J. J. & Mullen, K. J. (2004), 'The effect of schooling and ability on achievement test scores', *Journal of Econometrics* **121**(1–2), 39–98.

Hanushek, E. A. & Rivkin, S. G. (2006), Teacher quality, *in* E. A. Hanushek & F. Welch, eds, 'Handbook of the Economics of Education', Vol. 2, Elsevier B.V., Amsterdam, Netherlands, Chapter 18, pp. 1051–1078.

Hanushek, E. A. & Rivkin, S. G. (2010), 'Generalizations about using value-added measures of teacher quality', *American Economic Review* **100**(2), 267–271.

Heckman, J. J. & Mosso, S. (2014), 'The economics of human development and social mobility', *Annual Review of Economics* **6**, 689–733.

Heckman, J., Pinto, R. & Savelyev, P. (2013), 'Understanding the mechanisms through which an influential early childhood program boosted adult outcomes', *American Economic Review* **103**(6), 2052–2086.

Howard-Jones, P. A., Washbrook, E. V. & Meadows, S. (2012), 'The timing of educational investment: A neuroscientific perspective', *Developmental Cognitive Neuroscience* **25**, S18–S29.

Iversen, J. M. V. & Bonesrønning, H. (2013), 'Disadvantaged students in the early grades: Will smaller classes help them?', *Education Economics* **21**(4), 305–324.

Jackson, C. K. (2018), 'What do test scores miss? the importance of teacher effects on non–test score outcomes', *Journal of Political Economy* **126**(5), 2072–2107.

Jacob, B. & Rothstein, J. (2016), 'The measurement of student ability in modern assessment systems', *Journal of Economic Perspectives* **30**(3), 85–108.

Jones, R. H. & Boadi-Boateng, F. (1991), 'Unequally spaced longitudinal data with AR(1) serial correlation', *Biometrics* **47**(1), 161–175.

Kane, T. J. & Staiger, D. O. (2012), Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains, Report.

Kane, T. J., Taylor, E. S., Tyler, J. H. & Wooten, A. L. (2011), 'Identifying effective classroom practices using student achievement data', *Journal of Human Resources* **46**(3), 587–613.

Klausen, T. & Reikerås, E. (2016), *Regnefaktaprøven [Assessment of Arithmetic Facts]*, Norwegian Reading Centre, Stavanger, Norway.

Kotlarski, I. (1967), 'On characterizing the gamma and the normal distribution', *Pacific Journal of Mathematics* **20**(1), 69–76.

Kraft, M. A. (2019), ‘Teacher effects on complex cognitive skills and social-emotional competencies’, *Journal of Human Resources* **54**(1), 1–36.

Lewit, E. M. & Baker, L. S. (1995), ‘School readiness’, *The Future of children / Center for the Future of Children, the David and Lucile Packard Foundation* **5**(2), 128–139.

Madansky, A. (1964), ‘Instrumental variables in factor analysis’, *Psychometrika* **29**(2), 105–113.

Masten, A. S. & Coatsworth, J. D. (1998), ‘The development of competence in favorable and unfavorable environments: Lessons from research on successful children’, *American Psychologist* **53**(2), 205–220.

McKenzie, D. J. (2001), ‘Estimation of AR(1) models with unequally spaced pseudo-panels’, *The Econometrics Journal* **4**(1), 89–108.

Millimet, D. L. & McDonough, I. K. (2017), ‘Dynamic panel data models with irregular spacing: With an application to early childhood development’, *Journal of Applied Econometrics* **32**(4), 725–743.

National Research Council Institute of Medicine (2000), *From neurons to neighborhoods: The science of early childhood development*, National Academy Press, Washington, D.C.

OECD (2021), *OECD Skills Outlook 2021*, OECD Publishing, Paris, France.

Opper, I. M. (2019), ‘Does helping John help Sue? evidence of spillovers in education’, *American Economic Review* **109**(3), 1080–1115.

Parker, J. G. & Asher, S. R. (1987), ‘Peer relations and later personal adjustment: Are low-accepted children at risk?’, *Psychological Bulletin* **102**(3), 357–389.

Peterson, J. L. & Zill, N. (1986), ‘Marital disruption, parent-child relationships, and behavior problems in children’, *Journal of Marriage and the Family* pp. 295–307.

Pianta, R. C. (1997), 'Adult-child relationship processes and early schooling', *Early Education and Development* **8**(1), 11–26.

Rege, M., Størksen, I., Solli, I. F., Kalil, A., McClelland, M. M., ten Braak, D., Lenes, R., Lunde, S., Breive, S., Carlsen, M., Erfjord, I. & Hundeland, P. S. (2021), 'The effects of a structured curriculum on preschool effectiveness: A field experiment', *Journal of Human Resources* .

Rivkin, S. G., Hanushek, E. A. & Kain, J. F. (2005), 'Teachers, schools, and academic achievement', *Econometrica* **73**(2), 417–458.

Rockoff, J. E. (2004), 'The impact of individual teachers on student achievement: Evidence from panel data', *American Economic Review* **94**(2), 247–252.

Roemer, J. E. (2000), *Equality of Opportunity*, Harvard University Press, Cambridge, Massachusetts.

Rosner, B. & Muňoz, A. (1988), 'Autoregressive modelling for the analysis of longitudinal data with unequally spaced examinations', *Statistics in Medicine* **7**(1–2), 59–71.

Sasaki, Y. & Xin, Y. (2017), 'Unequal spacing in dynamic panel data: Identification and estimation', *Journal of Econometrics* **196**(2), 320–330.

Schanzenbach, D. W. (2006), 'What have researchers learned from Project STAR', *Brookings Paper on Education Policy* (9), 205–228.

Skrondal, A. & Rabe-Hesketh, S. (2004), *Generalized Latent Variable Modeling: Multilevel, Longitudinal, and Structural Equation Models*, Interdisciplinary Statistics, 1 edn, Chapman & Hall/CRC, Boca Raton, Florida.

Spitz-Oener, A. (2006), 'Technical change, job tasks, and rising educational demands: Looking outside the wage structure', *Journal of Labor Economics* **24**(2), 235–270.

Størksen, I. & Mosvold, R. (2013), 'Assessing early math skills with tablet computers: Development of the ani banani math test (ABMT) for young children [Utdanning2020]'.

Sylvia, S., Warrinnier, N., Luo, R., Yue, A., Attanasio, O., Medina, A. & Rozelle, S. (2020), 'From quantity to quality: Delivering a home-based parenting intervention through China's family planning cadres', *Economic Journal* **131**(635), 1365–1400.

ten Braak, D. & Størksen, I. (2021), 'Psychometric properties of the Ani Banani Math Test', *European Journal of Developmental Psychology* **18**(4), 610–628.

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

Torgerson, W. S. (1958), *Theory and Methods of Scaling*, John Wiley & Sons, Inc., New York City, New York.

VandenBos, G. R. E. (2007), *APA Dictionary of Psychology*, APA Dictionary of Psychology, American Psychological Association, Washington, D.C.

Voelkle, M. C. & Oud, J. H. L. (2013), 'Continuous time modelling with individually varying time intervals for oscillating and non-oscillating processes', *British Journal of Mathematical and Statistical Psychology* **66**(1), 103–126.

Vygotsky, L. S. (1978), *Mind in Society: The Development of Higher Psychological Processes*, Harvard University Press, Cambridge, Massachusetts.

Wansbeek, T. & Meijer, E. (2000), *Measurement Error and Latent Variables in Econometrics*, Advanced Textbooks in Economics, Elsevier Science B.V., Amsterdam, Netherlands.

Williams, B. (2020), 'Identification of the linear factor model', *Econometric Reviews* **39**(1), 92–109.

Chapter 2

Cognitive Skill Production and Unequally Spaced Panel Data

By

Maximiliaan W. P. Thijssen

Data collection often causes observed intervals to be unequal and child-specific. I estimate cognitive skill formation and analyze the role of the family environment at different stages in the life cycle of children while accounting for these unequally spaced intervals. I find a meaningful impact of failing to account for unequally spaced intervals, particularly as it relates to the self-productivity of cognitive skills, the influence of the family environment, and total factor productivity. For example, I find that a one-standard-deviation increase in parental investment raises children's cognitive skills by 0.086 of a standard deviation. When I account for the unequally spaced intervals, I estimate an effect of 0.045 of a standard deviation. These findings have implications for child development research, as I highlight the importance of accounting for child-specific unequal intervals when modeling cognitive skill formation.

Keywords. Cognitive skills, family influences, factor models, continuous-time modeling, unequally spaced intervals.

JEL codes. C13, C33, C38, J13.

Author note. I received funding from Grant 270703 from the Norwegian Research Council. I am thankful for helpful comments from Torberg Falch, Mari Rege, Erik Ø. Sørensen, and participants in the Ph.D. Workshop in Economics of Education (May 2021, Lund University) and the 43rd Annual Meeting of the Norwegian Association for Economists (January 2021, University of Bergen) for helpful comments.

1. Introduction

Children's cognitive skill formation is a primary focus of many social sciences.[1] When we estimate models of cognitive skill formation, we typically assume that all children are observed simultaneously and in equally spaced intervals. The interval assumed by our model is therefore like the one observed in the data. However, in most longitudinal studies, observation intervals are not equally spaced, and children are not observed at the same point in time. While several studies show that failing to account for unequally spaced intervals can lead to biased estimation, the implications of this estimation problem for estimating cognitive skill formation have not yet been studied.[2] I therefore analyze an often-used longitudinal dataset for investigating child development to explore these implications.

I use data on cognitive achievement and parental investments for the children who were part of the National Longitudinal Survey of Youth 1979 (NLSY79) to estimate a linear specification of the technology of skill formation (Cunha & Heckman 2007). I compare estimates obtained from models under equal and unequally spaced intervals, and find that estimates based on the former can differ greatly from those obtained under the latter. For example, a one-standard-deviation increase in parental investment boosts children's cognitive skills by 0.086 of a standard deviation. When I account for unequal spacing, I estimate an effect of 0.045 of a standard deviation, a 62.6-percent difference. I also find that precision is increased. For example, the standard error of the self-productivity estimate of 0.891 is 0.019. When I account for unequal spacing, the estimates become 0.960 and 0.010, respectively, a 62.1-percent difference in precision. These findings have implications for the econometrics of childhood human capital production (*cf*. Cunha et al. 2021).

To account for unequally spaced intervals, I apply insights from the continuous-time modeling literature. In particular, I implement the exact discrete model introduced by Albert R. Bergstrom (Bergstrom 1988). Discrete-time model parameters relate exactly to continuous-time model parameters through the exact discrete model for a particular observation

[1] See Heckman & Mosso (2014), and the studies they cite, for evidence from diverse literatures on the importance of children's skill formation.

[2] See, for example, McKenzie (2001), Sasaki & Xin (2017), and Millimet & McDonough (2017) for such evidence on (pseudo) panel data.

interval. By allowing this observation interval to be unequally spaced, I can estimate the underlying continuous-time model parameters. Once those are obtained, I can solve the continuous-time model for the interval assumed by the (discrete-time) model, taking into account the unequally spaced intervals observed in the data. By extension, I obtain the typically estimated parameters if I define the observation interval as equally spaced. It follows that I can compare estimates obtained under equal and unequally spaced intervals.

Using this insight and an economic framework as in Cunha & Heckman (2008) and Cunha et al. (2010), I estimate a production function for cognitive skills. The literature generally conceives of cognitive skills as factors relating to IQ and performance on achievement tests (*e.g.*, Borghans et al. 2008, pp. 979–980). In my model, cognitive skills formed at one stage depend on skills developed earlier in the child's life cycle, investments made by parents in earlier stages, the mother's cognitive skills, and other unobserved influences. As an "approximation" of household behavior, I allow these parental investments to depend (endogenously) on past levels of investments, the child's cognitive skills developed at an earlier stage, the mother's cognitive skills, and family income. In addition, I accommodate more general forms of endogeneity by modeling child-specific (time-invariant) heterogeneity in the skill formation model and parental investment function.

Many studies document the importance of cognitive skills in producing economic and social success (*e.g.*, Cawley et al. 2001). Therefore, understanding the factors that affect the evolution of cognitive skills is important for promoting successful lives. Several econometric frameworks exist for analyzing the impact of family and government inputs in producing cognitive skills (see Todd & Wolpin 2003, for an overview of these frameworks). In contrast, I use an alternative framework based on the models in Cunha & Heckman (2008) and Cunha et al. (2010).

Cunha & Heckman (2008) have expanded on models of cognitive skill formation (*cf.* Todd & Wolpin 2003, 2007) by identifying and estimating a dynamic factor model. Such models use covariability between measurements to form interpretable aggregates optimally weighted by the information content of each measure. Cunha et al. (2010) have extended the identification results and estimation in Cunha & Heckman (2008) to a nonlinear setting. In particular, they consider a constant elasticity of substitution technology and model the formation of cognitive and noncognitive skills, quantify the influence of the mother's skills and family inputs, and report evidence on important substitution parameters. Cunha et al. (2010) re-normalize technology parameters in each period to achieve identification. Agostinelli & Wiswall (2016) contrast such normalizations with normalizations of measurement model parameters.[3] They conclude that the latter type of normalization is preferable. Related literature uses these identification results and embeds experimental designs in models of skill formation to interpret treatment effects (*e.g.*, Attanasio, Cattan, Fitzsimons, Meghir & Rubio-Codina 2020, Sylvia et al. 2020, Thijssen et al. 2022).

I contribute to our understanding of estimating skill formation models by demonstrating that failing to account for unequally spaced intervals can be problematic. In subsequent work, Thijssen et al. (2022) apply the exact discrete model to account for unequally spaced intervals in a decomposition framework (*cf.* Heckman et al. 2013). They investigate the mechanisms underlying an early childhood intervention in Norway (Rege et al. 2021). The extent of unequal spacing in this intervention was limited because all centers met in a local science museum, where the preschool children were assessed. Nonetheless, even with limited variation in spacing along the time dimension, Thijssen et al. (2022) find meaningful differences for estimated self-productivity parameters. Consistent findings on the importance of accounting for unequally spaced intervals have implications for the econometrics of early childhood skill formation (Cunha et al. 2021).

While I do not contrast or develop new estimators for unequally spaced panel data, my study is relevant to literature with that aim (*e.g.*, Baltagi & Wu 1999, Jones & Boadi-Boateng 1991, McKenzie 2001, Millimet & McDonough 2017, Rosner & Muñoz 1988, Sasaki & Xin 2017). My paper differs from these papers in at least three ways. First, I consider models with *latent* dependent and independent variables. Second, the preceding papers (generally) presume that unequally spaced intervals can be attributed to survey design. In other words, children are observed in the same unequally spaced intervals. Suppose that the substantive timing

[3] The motivation for these normalizations is measurement invariance (Meredith 1993). Agostinelli & Wiswall (2016), in particular, consider a form of *weak* measurement invariance (Meredith 1993, pp. 530–532).

variable of child development is age, however. In that case, unequally spaced intervals can arise even when survey waves are equidistant, because children can vary in age in any particular wave. Third, I apply insights from the continuous-time modeling literature.[4] Accordingly, this paper is close in spirit to Voelkle & Oud (2013) in how it addresses the estimation problem of (child-specific) unequal intervals. Voelkle & Oud (2013) use Monte Carlo simulations to study the impact of unequal intervals that vary across individuals on oscillatory and non-oscillatory processes. In contrast, I study the impact of unequally spaced intervals that vary across children because of age, following a seminal economic model of skill formation (Cunha & Heckman 2007).

2. The Estimation Problem of Unequally Spaced Intervals

I explain the estimation problem of unequal intervals using a linear specification of the technology of skill formation (Cunha & Heckman 2007).[5] As this section focuses on the estimation problem, I make two simplifying assumptions. First, I assume that parameters are identified. Second, I assume inputs are observed without error. I return to the identification and measurement error in Section 3.

For simplicity, Consider a researcher interested in estimating children's cognitive skill formation as a function of previously developed cognitive skills, parental investments, and a shock. Let $\mathrm{S}_{i,t}$ denote cognitive skills at time t ($t = 0, 1, \dots, T$) for child i ($i = 1, \dots, N$), let $\mathrm{I}_{i,t}$ denote parental investments, and let $\epsilon_{i,t}$ denote a random shock. I assume that this shock, $\epsilon_{i,t}$, is independent across children and time (for the same child). Furthermore, I assume that the shock is independent of past inputs (*i.e.*, $\mathrm{S}_{i,t-1}$ and $\mathrm{I}_{i,t-1}$). A linear technology can then be defined as,

$$\mathrm{S}_{i,t} = \alpha_1 \mathrm{S}_{i,t-1} + \alpha_2 \mathrm{I}_{i,t-1} + \epsilon_{i,t}, \tag{1}$$

where the parameter α_1 measures the relationship between cognitive skills

[4]Jones & Boadi-Boateng (1991) also apply insights from the continuous-time modeling literature.

[5]For additional details on the problem of unequally spaced intervals, see Baltagi & Wu (1999), Jones & Boadi-Boateng (1991), McKenzie (2001), Millimet & McDonough (2017), Rosner & Muñoz (1988), and Sasaki & Xin (2017).

formed at time t and cognitive skills formed at time $t-1$ and α_2 measures the relationship between cognitive skills at time t and parental investments at time $t-1$.

Applying repeated substitution to Equation (1) yields (see Appendix A.1 for details),

$$\mathrm{S}_{i,t} = \alpha_1^{\Delta t}\mathrm{S}_{i,t-\Delta t} + \sum_{r=0}^{\Delta t-1} \alpha_1^r \alpha_2 \mathrm{I}_{i,t-1-r} + \sum_{r=0}^{\Delta t-1} \alpha_1^r \epsilon_{i,t-r}, \tag{2}$$

where t and Δt represent, respectively, the *assumed* time points and intervals. Equation (2) presents the general relationship between skills at time t and inputs at time $t-\Delta t$. If we take the unit interval, $\Delta t = 1$, then Equation (2) simplifies to Equation (1). If we take a larger interval, say $\Delta t = 2$, then we observe that the child's cognitive skill at time t depends (i) differently on past levels of cognitive skill and (ii) on all past inputs (*i.e.*, $\mathrm{I}_{i,t-1}$ and $\mathrm{I}_{i,t-2}$).

Next, let a ($a = 0, 1, \ldots, A$) index *observed* time points. These *observed* time points may (or may not) correspond to the *assumed* time points. In terms of *observed* inputs and *unobserved* inputs, Equation (2) can, after some rearranging, be written as follows (see Appendix A.1 for details),

$$\mathrm{S}_{i,a} = \underbrace{\alpha_1 \mathrm{S}_{i,a-1} + \alpha_2 \mathrm{I}_{i,a-1}}_{\text{Inputs from Observed Periods}} + \tilde{\epsilon}_{i,a}, \quad \text{with} \tag{3}$$

$$\tilde{\epsilon}_{i,a} \equiv \underbrace{(\alpha_1^{\Delta t_a} - \alpha_1)\mathrm{S}_{i,a-1} + \sum_{r=1}^{\Delta t_a-1} \alpha_1^r \alpha_2 \mathrm{I}_{i,t_a-r} + \sum_{r=0}^{\Delta t_a-1} \alpha_1^r \epsilon_{i,t_a-r}}_{\text{Inputs from Unobserved Periods}},$$

where t_a and $\Delta t_a \equiv t_a - t_{a-1}$ denote, respectively, the actual time points and intervals indicated by observed time point a.

Equation (3) shows that *unobserved* inputs (in $\tilde{\epsilon}_{i,a}$) depend on the *observed* interval and reduces to Equation (1) when (i) *observed* time points correspond to *assumed* time points and (ii) *observed* intervals are equally spaced (*i.e.*, $\Delta t_a = 1$ for all $a = 0, 1, \ldots, A$). Note that the summation operator in $\tilde{\epsilon}_{i,a}$ starts at $r = 1$, instead of $r = 0$, because we want to distinguish between inputs from *observed* and *unobserved* periods. The cognitive skills observed at time a are thus the sum of inputs previously observed (*i.e.*, $\mathrm{S}_{i,a-1}$ and $\mathrm{I}_{i,a-1}$) plus inputs from unobserved time points (in $\tilde{\epsilon}_{i,a}$) that may have occurred between a and $a-1$. Assume the relevant timing variable is age a (in years). If we observe children at ages five and seven ($\Delta t_a = 2$), then $\tilde{\epsilon}_{i,t}$ includes inputs from age six. If $\Delta t_a = 1$, then there are no unobserved inputs, and $\tilde{\epsilon}_{i,a}$ reduces to $\epsilon_{i,a}$.

I consider three scenarios to develop further intuition for Equation (3). I start with the "ideal" case, where we observe all children at the same age and in equally spaced intervals. Next, I consider a case where we observe all children at the same age but in unequally spaced intervals. Finally, I consider the NLSY79, the dataset I analyze in this article and where we do not observe children at the same age nor in equally spaced intervals.

Scenario 1. In the ideal case, a researcher may observe all children at ages $t_0 = 0$ and $t_1 = 1$, $t_2 = 2$, $t_3 = 3$, and $t_4 = 4$. I illustrate such data in Table 1 (Scenario 1). In this case, the observed time points correspond to assumed time points, and intervals are equally spaced. As a result, for each interval, Equation (3) simplifies to estimating Equation (1). For example, for the first interval (and similarly for the second, third, and fourth intervals), $\Delta t_1 = t_1 - t_0 = 1 - 0 = 1$, we have,[6]

$$\mathrm{S}_{i,1} = \alpha_1 \mathrm{S}_{i,0} + \alpha_2 \mathrm{I}_{i,0} + \epsilon_{i,1}. \tag{4}$$

Scenario 2. In the second case, the research still observes children of the same age, but the observed interval is not the interval assumed in the model (Table 1: Scenario 2). The suspension of data collection owing to a pandemic could be why researchers collected data later than intended. As a result, we may observe children at age $t_0 = 0$, $t_1 = 2$, $t_2 = 3$, and $t_3 = 4$.

[6]Scenario 1 appears ideal. Still, we require two additional assumptions to estimate Equation (4). First, we observe all children simultaneously. That is, the observed time points across children are typically coded as identical. Second, the observed time intervals are of substantive interest. While we generally have no reason to presume that a particular time interval (say 12 months) is more relevant than another (say 18 months), deciding on a particular interval affects parameter estimates and, by extension, their causal interpretation. Under these assumptions, we can estimate Equation (4) to obtain the parameters of interest as assumed by the model in Equation (1).

Consider the first observed interval, $\Delta t_1 = t_1 - t_0 = 2$. As the second time point does not correspond to the observed time point, Equation (3) reduces to estimating,

$$\mathrm{S}_{i,1} = \alpha_1 \mathrm{S}_{i,0} + \alpha_2 \mathrm{I}_{i,0} + [(\alpha_1^2 - \alpha_1)\mathrm{S}_{i,0} + \alpha_1 \alpha_2 \mathrm{I}_{i,1} + \epsilon_{i,2} + \alpha_1 \epsilon_{i,1}], \quad (5)$$

where the parameter that measures the relationship between cognitive skills formed at time t and cognitive skills formed at time $t-1$ (*i.e.*, α_1) has a different interpretation than the one assumed in Equation (1) because of the polynomial. Furthermore, the observed inputs (*i.e.*, $\mathrm{S}_{i,0}$ and $\mathrm{I}_{i,0}$) may

Table 1. Three Scenarios Concerning Unequally Spaced Panel Data

Scenario 1					Scenario 2					Scenario 3 (NLSY79)				
i	*a*	*t*	S	I	*i*	*a*	*t*	S	I	*i*	*a*	*t*	S	I
1	0	0	12	5	1	0	0	12	5	1	6.67	0	12	−0.51
1	1	1	15	6	1	–	1	–	6	1	8.83	1	12	−0.71
1	2	2	18	8	1	1	2	18	8	1	10.92	2	48	−0.59
1	3	3	20	9	1	2	3	20	9	1	12.83	3	18	−0.63
1	4	4	22	11	1	3	4	22	11	1	14.92	4	–	–
2	0	0	15	2	2	0	0	15	2	2	6.17	0	4	−0.68
2	1	1	16	6	2	–	1	–	6	2	8.17	1	2	−0.70
2	2	2	17	4	2	1	2	17	4	2	10.33	2	7	−0.63
2	3	3	22	3	2	2	3	22	3	2	12.33	3	–	−0.13
2	4	4	29	9	2	3	4	29	9	2	14.17	4	10	−1.16
3	0	0	18	5	3	0	0	18	5	3	5.00	0	6	−2.19
3	1	1	17	4	3	–	1	–	4	3	7.42	1	–	−0.85
3	2	2	30	3	3	1	2	30	3	3	9.00	2	31	−0.31
3	3	3	25	8	3	2	3	25	8	3	11.08	3	42	−1.21
3	4	4	26	6	3	3	4	26	6	3	13.08	4	–	–
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Notes. This table presents three scenarios concerning unequally spaced panel data. In Scenario 3, I use data from the NLSY79. I use the information on the child's age (in months) at the time of assessment, the child's score on the Peabody Individual Achievement Test: mathematics, and a (standardized) composite of the Home Observation Measurement of the Environment - Short Form.

correlate with the error term because of the inputs from missing periods.[7] An estimation problem thus arises when the observed interval is not the one assumed by the model.

Scenario 3. The NLSY79 Children and Young Adults is an often-used dataset for estimating models of cognitive (and noncognitive) skill formation (*e.g.*, Agostinelli & Wiswall 2016, Cunha & Heckman 2008, Cunha et al. 2010, Todd & Wolpin 2003, 2007). Table 1 (Scenario 3) depicts three children in the NLSY79, their age at the time of assessment (a), assessment wave (t), math score (S), and (standardized) investments (I). In Table 1 (Scenario 3), I followed Cunha et al. (2010) and Agostinelli & Wiswall (2016) in aggregating age as follows: age 5–6, age 7-8, age 9-10, age 11–12, and age 13–14.[8]

Assuming that the child's age is the relevant timing variable of child development, then the unequal intervals observed in Table 1 (Scenario 3) can arise for (a combination of) two reasons.[9] First, same-aged children are not assessed simultaneously. Second, simultaneously assessed children are not of the same age. For example, for child 1 ($i = 1$) in Table 1 (Scenario 3), the distance between the first and second observed time points is 2.16 years, whereas it is 2.42 years for child 3 ($i = 3$). For child 3, the distance between the first and second observed time points is 2.42 years, whereas the distance between the fourth and fifth observed time points is only 2.00 years. Scenario 3 is thus more complicated than Scenario 1 or Scenario 2 because the extent of unequally spaced intervals varies within and between

[7] In essence, we can think of the unequally spaced intervals as a missing data problem because the data are missing at time point $t = 1$. This missing data problem causes issues for conventional estimators (see, *e.g.*, Millimet & McDonough 2017).

[8] Parenthetically, this aggregation differs from Cunha & Heckman (2008), who aggregate age as follows: age 6-7, age 8-9, age 10-11, and age 12-13. The point made in this section does not depend on the aggregation decision.

[9] Most longitudinal studies target same-aged children based on birth year. In general, the development of children born in the same year and measured simultaneously can differ by as much as 12 months at the time of observation. Such differences can affect cognitive achievement (*e.g.*, Crawford et al. 2010). If, in addition, measurement can occur at any time during the year, then the development of children born in the same year can differ by as much as 24 months at the time of assessment. We would naturally expect these developmental differences to become smaller as children grow older. This expectation is consistent with the fade-out effect reported in Elder & Lubotsky (2009), who studied the effect of school starting age on test scores.

children. In terms of Equation (3), this implies that the *observed* interval is child-specific further complicating the estimation problem.

I propose to account for these (child-specific) unequally spaced intervals (*i.e.*, $\Delta t_{i,a} \equiv t_{i,a} - t_{i,a-1}$) using the exact discrete model. Suppose that the process by which children develop is continuous. We can then consider our observations as child-specific "snippets" of the continuous process. It follows that these snippets are not a problem but an opportunity, as variation provides information about the underlying continuous-time process (Voelkle & Oud 2013). Once I have estimated the continuous-time model, I can solve it for any substantive interval (*e.g.*, $\Delta t_{i,a} = 1$).[10]

In sum, there is a consensus that we should account for unequally spaced intervals (*e.g.*, Millimet & McDonough 2017, Voelkle & Oud 2013). We have yet to investigate the implications of this estimation problem for models of children's cognitive skill formation. I aim to account for unequally spaced intervals by incorporating child age information during each measurement occasion. I incorporate this information by applying the exact discrete model from the continuous-time modeling literature (Bergstrom 1988). I can examine the extent to which estimates are affected by contrasting estimates based on equal and unequally spaced intervals. The next section describes the model of skill formation I consider.

3. An Economic Model of Skill Formation

Section 3.1 describes a linear production function of cognitive skills. While a linear specification imposes perfect substitution, implying that skill remediation is always possible (Cunha et al. 2010), it is parsimonious. A linear specification is the simplest way to study the implications

[10]One may wonder whether the "vertical scales" discussed in Agostinelli & Wiswall (2016, pp. 14–16) can be helpful here because vertical scales combine *similar* domains at *different* educational levels into a common scale. Just as children at different educational levels vary in age, children can also be of different ages at the same educational level (*e.g.*, red-shirting). Consequently, it is not clear that vertical scales would be a solution. By extension, age variation at the measurement time could also have implications for the measurement invariance restrictions discussed in Agostinelli & Wiswall (2016). If child age varies at the time of measurement, we implicitly assume some degree of age invariance. This implication follows from the observation that we commonly assume that measurement properties are homogeneous across children.

of unequally spaced intervals for estimating models of cognitive skill formation.[11] Section 3.2 describes the measurement models. Section 3.3 discusses identification.

3.1. A Linear Production Function for Children's Cognitive Skills

In addition to the notation introduced above, let $\mathrm{S}_{\mathrm{P},i}$ denote the cognitive skills of child i's mother.[12] The investments and the mother's cognitive skills are characteristics of the childhood (home) environment. By enriching the child's experiences in the home, parents "invest" in the child's skill formation. The technology of skill formation (Cunha & Heckman 2007) uses these inputs to define human (capital) development formally,

$$\mathrm{S}_{i,t} = f_k(\mathrm{S}_{i,t-1}, \mathrm{I}_{i,t-1}, \mathrm{S}_{\mathrm{P},i}, \epsilon_{i,t}), \tag{6}$$

for child i ($i = 1, \dots, N$), time t ($t = 0, 1, \dots, T$), and developmental stage k ($k = 1, \dots, K$). Child development is thus defined as a dynamic and continuous interaction between a child's genetic makeup (*i.e.*, initial conditions) and experiences in the environment (National Research Council Institute of Medicine 2000).

It follows, then, that Equation (7) defines the process of child development when we assume a functional form that is linear-in-parameters,

$$\mathrm{S}_{i,t} = \alpha_0 + \alpha_1 \mathrm{S}_{i,t-1} + \alpha_2 \mathrm{I}_{i,t-1} + \alpha_3 \mathrm{S}_{\mathrm{P},i} + \epsilon_{i,t}, \tag{7}$$

for child i ($i = 1, \dots, N$) and time t ($t = 0, 1, \dots, T$).[13] In Equation (7), α_0 is an intercept (or total factor productivity), α_1 measures the relationship between cognitive skills formed at time t and cognitive skills formed at

[11]Linear-in-parameter specifications are common in the literature (*e.g.*, Agostinelli & Wiswall 2016, Attanasio, Cattan, Fitzsimons, Meghir & Rubio-Codina 2020, Cunha & Heckman 2008, Sylvia et al. 2020, Todd & Wolpin 2003, 2007). Furthermore, linear-in-parameters can be as flexible as the (nonlinear) constant elasticity of substitution technology considered in Cunha et al. (2010) and Attanasio, Meghir & Nix (2020) as it could be approximated using a linear translog production function.

[12]Mother's skills are measured only once in the NLSY79 dataset, and therefore $\mathrm{S}_{\mathrm{P},i}$ has no time subscript. The dataset provides no measure of fathers' cognitive skills.

[13]To avoid notational clutter, I kept the k ($k = 1, \dots, K$) subscript on the parameters (α_0, α_1, α_2, and α_3) implicit in Equation (7). These parameters may vary across the $K \leq T$ stages of development, however (*cf.* Cunha et al. 2010).

time $t - 1$ (*i.e.*, self-productivity of cognitive skills), α_2 measures the relationship between the mother's cognitive skills and the child's cognitive skills, and $\epsilon_{i,t}$ denotes a mean-zero error term.

I make two independence assumptions about the error term. First, the error term is independent across children and time (for the same child). Second, the inputs used to produce the child's cognitive skills (*i.e.*, $\mathrm{S}_{i,t-1}$, $\mathrm{I}_{i,t-1}$, and $\mathrm{S}_{\mathrm{P},i}$) are independent of the error term. Section 3.3.3 relaxes these assumptions and allows unobserved inputs to affect cognitive skill formation. For example, cognitive skills could depend on stable differences between children (*e.g.*, genetic material), which can cause the errors to be serially correlated.

3.2. A Linear System of Measurement Equations

The cognitive skills and parental investments in Equation (7) are not directly observable. One solution is to use (standardized) test scores to proxy cognitive skills and sum-scores of family inputs to proxy parental investments. In general, this "*measurement by fiat*" (Torgerson 1958, p. 22) is not recommended. First, we cannot theoretically motivate such sum scores as a measurement strategy (Skrondal & Rabe-Hesketh 2004, pp. 242–243). Second, there are many (standardized) tests and family inputs, each of which might be informative to very different degrees (Cunha & Heckman 2008, Cunha et al. 2010). Third, individual measures of cognitive skill and parental investments are imperfect (Borghans et al. 2008); they include variance related to the variable of interest as well as unrelated variance. We do not want to confound these different sources of variation. Fourth, we cannot partial out other observed influences with a sum-score approach. It is therefore desirable to specify a system of measurement equations.

This section discusses a measurement equation system that maps the unobserved inputs in Equation (7) onto observable measures. In what follows, I refer to the unobserved inputs as *factors* and the observable measures as *manifest variables*. The intuition is as follows: First, cognitive skills can manifest in observable measures such as achievement tests (Borghans et al. 2008). Second, parental investments can manifest in family inputs (*e.g.*, daily newspaper, museum visits) used to produce the desired investment level given endowments, earnings, and the child's skills (Cunha & Heckman 2008). While *common factors* manifest in related manifest variables, *unique factors* manifest uniquely in a particular manifest variable. In turn, unique factors could include *specific factors* – representing systematic variation – and *measurement errors*.

Suppose that the number of manifest variables related to the child's cognitive skill at time t is $L_{1,t}$, where l ($l = 1, \dots, L_{1,t}$) indexes a manifest variable. Let $\mathrm{M}_{1,l,i,t}$ denote the lth manifest variable for child i at time t. Furthermore, the number of manifest variables related to parental investment at time t is $L_{2,t}$ ($l = 1, \dots, L_{2,t}$), with $\mathrm{M}_{2,l,i,t}$ denoting the manifest variable. Lastly, the number of manifest variables related to the mother's cognitive skill is L_3 ($l = 1, \dots, L_3$), where $\mathrm{M}_{3,l,i}$ denotes the manifest variable. As is common in the literature (*e.g.*, Attanasio, Cattan, Fitzsimons, Meghir & Rubio-Codina 2020, Attanasio, Meghir & Nix 2020, Cunha & Heckman 2008), I assume that common factors manifest uniquely in one manifest variable in order to facilitate interpretation (*i.e.*, dedicated measures).[14]

As in Equation (7), I assume a functional form that is linear-in-parameters for the measurement equations,

$$\mathrm{M}_{1,l,i,t} = \mu_{1,l,t} + \lambda_{1,l,t}\mathrm{S}_{i,t} + \zeta_{1,l,i,t}, \quad l = 1, \dots, L_{1,t}, \tag{8}$$

$$\mathrm{M}_{2,l,i,t} = \mu_{2,l,t} + \lambda_{2,l,t}\mathrm{I}_{i,t} + \zeta_{2,l,i,t}, \quad l = 1, \dots, L_{2,t}, \tag{9}$$

$$\mathrm{M}_{3,l,i} = \mu_{3,l} + \lambda_{3,l}\mathrm{S}_{\mathrm{P},i} + \zeta_{3,l,i}, \quad l = 1, \dots, L_3, \tag{10}$$

for child i ($i = 1, \dots, N$) and time t ($t = 0, 1, \dots, T$). Equation (8), Equation (9), and Equation (10) map each manifest variable l to the corresponding common factor. Borghans et al. (2008) provide intuition for measurement Equation (8) and Equation (10). Appendix 1 and Appendix A2 in Cunha & Heckman (2008) and Cunha et al. (2010), respectively, provide further intuition for measurement Equation (9).[15]

[14] I do not require this assumption for identification. Suppose we assume that each of the common factors manifests uniquely in one of the manifest variables. We can then identify a factor model in which the common factors manifest jointly in the other manifest variables (*e.g.*, bi-factor models: Holzinger & Swineford 1937).

[15] Cunha & Heckman (2008) and Cunha et al. (2010) use derived demand equations to motivate the parental investment measurement equation, where the parent's objective

The parameters $\mu_{1,l,t}$, $\mu_{2,l,t}$, and $\mu_{3,l}$ are intercepts, $\lambda_{1,l,t}$, $\lambda_{2,l,t}$, and $\lambda_{3,l}$ are factor loadings, and the variables $\zeta_{1,l,i,t}$, $\zeta_{2,l,i,t}$, and $\zeta_{3,l,i}$ are unique factors. These factor loadings "weigh" the manifest variables based on their correlation with the common factor. The parameters and variables in the measurement equations are defined conditional on a vector of covariates, which I keep implicit. I assume that these covariates are independent of the common and unique factors. In addition, I assume that the unique factors are mean zero and independent of the common factor. Lastly, I assume that the unique factors are independent of one another conditional on the common factor. Unique factors can freely correlate over time, however. This serial correlation is important in the presence of specific factors.[16]

First, the assumption that unique factors are independent across time may not be tenable if a battery of tests measures the same specific factors. It could be important to allow unique factors to be serially correlated with such time dependencies.[17] Second, claims concerning "*substantial*" measurement error (Cunha & Heckman 2008, Cunha et al. 2010), based on the degree of noise (*i.e.*, $\text{Var}(\zeta_{1,l,i,t})/\text{Var}(\text{M}_{1,l,i,t})$, where $\text{Var}(\cdot)$ denotes the variance operator), are not necessarily appropriate since the numerator includes the specific factor variance. It is more appropriate to conceive of these noise computations as lower bounds on the "true" reliability (Skrondal & Rabe-Hesketh 2004, pp. 66–67). Third, age-invariance restrictions, as defined in Agostinelli & Wiswall (2016), may not be sufficient to ensure that the expected difference in two (consecutive) measurements is zero conditional on a constant level of skill.[18]

is to maximize lifetime utility, which depends on consumption and the child's skills in adulthood.

[16]This serial correlation is identifiable with a minimum of three measures (see, *e.g.*, Cunha et al. 2010). See also Appendix A.2.

[17]Identifying these specific factors requires longitudinal data or multitrait-multimethod designs (see, *e.g.*, Alwin 1989).

[18]Consider a situation in which measurements provide a basis for choosing children from an applicant pool to enroll in a comprehensive early childhood education program. If, for a constant level of skill, the conditional variance at a specific age is systematically larger (or smaller) than the conditional variance at an earlier age, then a comparison would be unfair, as the likelihood of being chosen would differ. Specific factors can cause such variation and may thus affect mean comparisons across time if one defines invariance based on the conditional expectation alone (Meredith 1993).

3.3. Identifying the Linear Production Function for Cognitive Skills

I first discuss the identification of the distribution of factors (Section 3.3.1), doing so only briefly because it follows standard identification arguments for factor models (*e.g.*, Anderson & Rubin 1956, Bollen 1989). For the sake of completeness, Appendix A.2 provides a detailed exposition of how we can identify the factor loadings and the distribution of factors. Second, Section 3.3.2 discusses accounting for the endogeneity of investments by specifying an investment function that approximates household behavior, similar to, among others, Agostinelli & Wiswall (2016), Attanasio, Cattan, Fitzsimons, Meghir & Rubio-Codina (2020), and Cunha et al. (2010). Lastly, I consider more general forms of (time-invariant) endogeneity in Section 3.3.3, similar to the approach taken in Balestra & Nerlove (1966).

3.3.1. Identifying the Distribution of Factors

Since none of the right-hand-side variables in Equation (8) through Equation (10) are observable, there is an inherent identification problem. In addition to the independence assumptions concerning the unique factors made in Section 3.2, identification requires some normalization to set a scale and location (Anderson & Rubin 1956). As a practical matter, I assume that I observe the same manifest variables in consecutive periods, as is the case in the NLSY79 (see Section 5).

For the child's cognitive skills (Equation 8), I normalize a factor loading on the same manifest variable (say the first) in each assessment wave to maintain a consistent interpretation of cognitive skill as children age. I set the location by normalizing the mean of the common factor in the initial period to zero and normalizing the intercept of the manifest variable used as an anchor to be invariant across time (Agostinelli & Wiswall 2016, Meredith 1993). I can then identify the intercept in Equation (7). Second, I set a scale for parental investment by normalizing the factor loading of the same manifest variable (say the first) in each period to one (Equation 9). To set a location, I normalize the mean of the parental investment factor to zero in each period. Lastly, for the mother's cognitive skill (Equation 10), I set the scale by normalizing a factor loading (say the first) to one, and I set the location by normalizing the mean of the common factor to zero.

Together, these normalizations (and the independence assumptions made in Section 3.2) are sufficient for identifying the distribution of factors. I can namely identify the factor loadings from the ratio of covariances.[19] With a minimum of three manifest variables for each common factor, it is possible to identify the serial correlation of the unique factors (Cunha et al. 2010). With the factor loadings identified, I can (nonparametrically) identify the distribution of factors by applying Kotlarski's lemma (see Lemma 1, Remark 4, and Remark 5 in Kotlarski 1967, pp. 70–73). I can thus identify the distribution of factors, up to a a change in sign, from the distribution of manifest variables, provided that the characteristic function does not vanish. Cunha et al. (2010, p. 893) show that (nonparametric) identification is also achievable under weaker independence assumptions concerning the unique factors (see also Ben-Moshe 2018).

3.3.2. Accounting for the Endogeneity of Parental Investments

As an "approximation" of household behavior, I allow parental investments to depend (endogenously) on past levels of investments, the child's cognitive skills developed at an earlier stage, the parent's cognitive skills, and family income. The related literature (*e.g.*, Agostinelli & Wiswall 2016, Attanasio, Cattan, Fitzsimons, Meghir & Rubio-Codina 2020, Cunha et al. 2010) generally uses a reduced-form approximation of household behavior instead of deriving the investment function from an explicit household model as in Del Boca et al. (2014). While such a reduced-form approximation yields no policy implications, it largely simplifies computation. Moreover, such a reduced-form approximation is sufficient for investigating the extent to which unequally spaced intervals affect the estimation of children's cognitive skill formation.

I consider the following linear-in-parameters investment function that specifies parental investment endogenously:

[19]For example, for the child's cognitive skills, we can write:

$$\frac{\mathrm{Cov}(\mathrm{M}_{1,l,i,t}, \mathrm{M}_{1,l',i,t})}{\mathrm{Cov}(\mathrm{M}_{1,1,i,t}, \mathrm{M}_{1,l',i,t})} = \frac{\lambda_{1,l,t}\lambda_{1,l',t}\mathrm{Var}(\mathrm{S}_{i,t})}{\lambda_{1,l',t}\mathrm{Var}(\mathrm{S}_{i,t})} = \lambda_{1,l,t},$$

for manifest variable $l, l' = 1, \dots, L_{1,t}$, $l \neq l'$, and time $t = 0, 1, \dots, T$. In Appendix A.2, I provide a detailed exposition of the identification.

$$\mathrm{I}_{i,t} = \beta_1 \mathrm{S}_{i,t-1} + \beta_2 \mathrm{I}_{i,t-1} + \beta_3 \mathrm{S}_{\mathrm{P},i} + \beta_4 \ln \mathrm{y}_{i,t-1} + \eta_{i,t}, \tag{11}$$

for child i ($i = 1, \dots, N$) and time t ($t = 0, 1, \dots, T$).[20] Variable $\mathrm{y}_{i,t-1}$ in Equation (11) is (log) family income, which I assume to follow a first-order auto-regressive process, and $\eta_{i,t}$ is a mean-zero error term.[21] I also assume that the error term is independent across children and time (for the same child). Furthermore, I assume that the inputs in the investment function (*i.e.*, $\mathrm{S}_{i,t-1}$, $\mathrm{I}_{i,t-1}$, $\mathrm{S}_{\mathrm{P},i}$, and $\mathrm{y}_{i,t-1}$) are independent of the error term. However, the error term in Equation (11) and Equation (7) can correlate freely. Section 3.3.3 relaxes these assumptions and allows for more general forms of endogeneity by modeling (child-specific) time-invariant heterogeneity in the parental investment function.

Parents can reinforce better-endowed children or compensate for less well-endowed children (Becker & Tomes 1976). Consequently, we can (loosely) interpret the parameter β_1 as the "reinforcing" or "compensating" behavior of the parents, as noted by Agostinelli & Wiswall (2016): Parents can increase future investment because of a desire to reinforce a child's high level of cognitive skill or a desire to compensate for the child's low level of cognitive skill. The parameter associated with past investments, β_2, measures time-dependence. Consider the case in which parents hold a certain belief about the rate of return to spending quality time with the child. A parent may try to make up for time lost by spending more time in the future, causing parental investments to be time-dependent. Another case in point is habit formation. Parents may develop and reinforce habits related to their investments (*e.g.*, reading a bedtime story). In turn, parents' reinforcement of habits may explain why it is challenging to change parenting behaviors and why interventions targeted at parents may not result in long-term effects (Kalil 2015). The parameter β_3 measures

[20] Like in Equation (7), I keep the k ($k = 1, \dots, K$) subscript on the parameters (β_1, β_2, β_3, and β_4) implicit in Equation (11).

[21] I define this auto-regressive process process as follows: $\ln \mathrm{y}_{i,t} = b_0 + b_1 \ln \mathrm{y}_{i,t-1} + \nu_{i,t}$, for child i ($i = 1, \dots, N$) and time t ($t = 0, 1, \dots, T$). The parameter b_0 is an intercept, b_1 measures the (temporal) dependence of family income, and $\nu_{i,t}$ is a mean-zero error term. I assume that the error term is independent across children and over time (for the same child). I also assume that the error term is independent of the lagged family income, the error terms in Equation (7) and Equation (11), and the inputs that determine the child's cognitive skills (Equation 7) and the parental investments (Equation 11).

the extent to which a mother's cognitive skill affects future investment. The last parameter, β_4, measures the relationship between family income and investments.

3.3.3. Accounting for Time-Invariant Unobserved Inputs

Throughout Section 3, I maintained the assumption that the error terms are independent over time (for the same child) and of the other inputs that determine the child's cognitive skill formation and parental investments. For example, Equation (7) and Equation (11) only account for temporal stability through lagged cognitive skills and parental investments. In other words, I assume that there are no (unobserved) stable differences between children that affect cognitive skill formation or parents' investment decisions, which is a strong assumption. For example, genetic material may account for stable differences in cognitive skills. In addition, parenting styles may account for stable differences in investments. Suppose there are differences like these that I do not observe. In that case, the error terms may be serially correlated for the same child, thereby violating the assumption maintained so far.

To separate such within-child and between-child variation, I apply the intuition described in Balestra & Nerlove (1966). I decompose the error terms in Equation (7) and Equation (11) into a time-invariant and time-varying component: $\epsilon_{i,t} = \kappa_{\mathrm{S},i} + \xi_{\mathrm{S},i,t}$ and $\eta_{i,t} = \kappa_{\mathrm{I},i} + \xi_{\mathrm{I},i,t}$, respectively. The variables $\kappa_{\mathrm{S},i}$ and $\kappa_{\mathrm{I},i}$ capture *all* time-invariant inputs that explain child-specific heterogeneity in cognitive skill formation or parental investments, respectively. Furthermore, the variables $\xi_{\mathrm{S},i,t}$ and $\xi_{\mathrm{I},i,t}$ represent the new (conventional) error term.

I rewrite Equation (7) as follows,[22]

$$\mathrm{S}_{i,t} = \gamma_0 + \gamma_1 \mathrm{S}_{i,t-1} + \gamma_2 \mathrm{I}_{i,t-1} + \kappa_{\mathrm{S},i} + \xi_{\mathrm{S},i,t}, \tag{12}$$

for child i ($i = 1, \ldots, N$) and time t ($t = 0, 1, \ldots, T$). I assume that $\kappa_{\mathrm{S},i}$ and $\xi_{\mathrm{S},i,t}$ are mean zero. Second, $\kappa_{\mathrm{S},i}$ is independent across children. Third, conditional on the child-specific time-invariant component, $\kappa_{\mathrm{S},i}$, the error term, $\xi_{\mathrm{S},i,t}$, is independent across children and time (for the same child).

[22] I omitted the mother's cognitive skill, $\mathrm{S}_{\mathrm{P},i}$, from Equation (12) and (13) because the time-invariant components capture this effect.

Fourth, $\xi_{S,i,t}$ is independent of $\kappa_{S,i}$ and the inputs ($S_{i,t-1}$ and $I_{i,t-1}$), but the inputs can freely correlate with $\kappa_{S,i}$. I rewrite Equation (11) as follows:

$$I_{i,t} = \delta_1 S_{i,t-1} + \delta_2 I_{i,t-1} + \delta_3 \ln y_{i,t-1} + \kappa_{I,i} + \xi_{I,i,t}, \tag{13}$$

for child i ($i = 1, \dots, N$) and time t ($t = 0, 1, \dots, T$). The same assumptions concerning the time-invariant, $\kappa_{I,i}$, and time-varying component, $\xi_{I,i,t}$, as in Equation (12) apply in Equation (13). Additionally, the time-invariant components in Equation (12) and Equation (13) can correlate freely. The time-varying components in Equation (12) and Equation (13) can correlate freely as well.

4. Estimation Procedure

In this section I explain how I incorporate unequally spaced intervals using the exact discrete model introduced by Albert R. Bergstrom (Bergstrom 1988). The intuition of the exact discrete model is as follows: Discrete-time model parameters relate exactly to continuous-time model parameters through the exact discrete model for a given interval. By defining this interval as unequally spaced, I obtain the corresponding continuous-time model parameters. Once I obtain those, I can solve the continuous-time model for the interval assumed by the (discrete-time) model, but taking into account the unequally spaced intervals observed in the data. Below, I formalize this intuition using a simplified version of the model described in Section 3. I describe the full model and estimation of the full model more extensively in Appendix A.3. For further details on the exact discrete model, see Oud & Jansen (2000), Hamerle et al. (1991), Hamerle et al. (1993), and Singer (1990, 1993, 1995).

I follow Cunha & Heckman (2008), among others, and assume that the error terms in Section 3 are normally distributed. While normality of the error terms is not required for identification, it facilitates computation. To illustrate the intuition, I focus on the child's cognitive skills and parental investments. I first stack parts of Equation (7) and Equation (11),

$$\begin{bmatrix} S_{i,t} \\ I_{i,t} \end{bmatrix} = \begin{bmatrix} \alpha_1 & \alpha_2 \\ \beta_1 & \beta_2 \end{bmatrix} \begin{bmatrix} S_{i,t-1} \\ I_{i,t-1} \end{bmatrix} + \begin{bmatrix} \epsilon_{i,t} \\ \eta_{i,t} \end{bmatrix}. \tag{14}$$

Let $\boldsymbol{\theta}'_{i,t} = [\mathrm{S}_{i,t}, \mathrm{I}_{i,t}]$, $\boldsymbol{\theta}'_{i,t-1} = [\mathrm{S}_{i,t-1}, \mathrm{I}_{i,t-1}]$, and $\mathbf{w}'_{i,t} = [\epsilon_{i,t}, \eta_{i,t}]$, so that Equation (14) can be represented in matrix form as follows:

$$\boldsymbol{\theta}_{i,t} = \mathbf{B}_{\Delta t}\boldsymbol{\theta}_{i,t-\Delta t} + \mathbf{w}_{i,t}, \quad \mathbf{Cov}(\mathbf{w}_{i,t}) = \mathbf{Q}_{\Delta t}, \tag{15}$$

where $\mathbf{B}_{\Delta t}$ denotes the $(2,2)$-dimensional coefficient matrix and $\mathbf{Q}_{\Delta t}$ the $(2,2)$-dimensional variance-covariance matrix of shocks. Subscript $\Delta t \equiv t-(t-1)$ signifies that the parameters depend on the time interval.

Equation (15) belongs to the stochastic differential equation,

$$\frac{\mathrm{d}\boldsymbol{\theta}_i(t)}{\mathrm{d}t} = \mathbf{B}\boldsymbol{\theta}_i(t) + \mathbf{G}\frac{\mathrm{d}\mathbf{w}_i(t)}{\mathrm{d}t}, \tag{16}$$

where the parameters no longer depend on the time interval.[23] One can show that the solution to Equation (16) is (Arnold 1974, pp. 128–134),

$$\boldsymbol{\theta}_i(t) = \mathbf{e}^{\mathbf{B}(t-t_0)}\boldsymbol{\theta}_i(t_0) + \int_{t_0}^{t} \mathbf{e}^{\mathbf{B}(t-s)}\mathbf{G}\mathrm{d}\mathbf{w}_i(s), \tag{17}$$

for initial value $\boldsymbol{\theta}_i(t_0)$ and observation interval $t-t_0$. The $\mathbf{e}^{\{\cdot\}}$ denotes the matrix exponential operator – a highly nonlinear matrix function (see Appendix A.4 for details on this matrix function). For $\boldsymbol{\theta}_i(t_0) = \boldsymbol{\theta}_{i,t-\Delta t}$ and $t - t_0 = \Delta t$, Equation (15) can be set equal to Equation (17) under the following restrictions (Arnold 1974, Singer 1990):

$$\mathbf{B}_{\Delta t} = \mathbf{e}^{\mathbf{B}\Delta t} \tag{18}$$

$$\mathbf{Q}_{\Delta t} = \mathrm{irow}\{(\mathbf{B}\otimes\mathbf{I} + \mathbf{I}\otimes\mathbf{B})^{-1}(\mathbf{B}_{\Delta t}\otimes\mathbf{B}_{\Delta t} - \mathbf{I}\otimes\mathbf{I})\mathrm{row}\{\mathbf{G}\mathbf{G}'\}\} \tag{19}$$

where row$\{\cdot\}$ denotes the operation that puts the elements of the diffu-

[23]In the beginning of this section, I assumed that the (discrete time) error terms in Section 3 are normally distributed. For this reason, a natural continuous-time process equivalent is the Wiener process, $\mathbf{w}_i(t)$, because increments of the Wiener process, $\Delta\mathbf{w}_i(t) \equiv \mathbf{w}_i(t) - \mathbf{w}_i(t-\Delta t)$, are normally distributed according to $\mathrm{N}(\mathbf{0}, \Delta t\mathbf{I})$. The pre-multiplication with the (lower-triangular) matrix $\mathbf{G}$ allows one to change the error variance-covariance matrix without changing the continuous-time error process. The product $\mathbf{G}\mathbf{G}'$ defines the "diffusion matrix." See Arnold (1974, pp. 45–78) and Hamerle et al. (1991, pp. 204–206) for further details.

sion matrix $\mathbf{GG}'$ row-wise in a column vector, $\otimes$ denotes the Kronecker product, and irow$\{\cdot\}$ is the inverse of row$\{\cdot\}$. By defining the interval in Equation (18) to be unequal (and child-specific), $\Delta t = \Delta t_{i,a} \equiv t_{i,a} - t_{i,a-1}$, as *observed* in the data, I can obtain the (continuous-time) coefficient matrix $\mathbf{B}$ that produced the (discrete-time) coefficient matrix $\mathbf{B}_{\Delta t_{i,a}}$. I can then retrieve the coefficient matrix *assumed* by the model (Equation 15), $\mathbf{B}_{\Delta t}$, by taking the matrix exponential of $\mathbf{B}$ multiplied by $\Delta t = 1$.

After obtaining the (nonlinear) constraints that link the discrete-time model parameters to the continuous-time model parameters in an exact way, the estimation procedure follows standard practices for structural models (Bollen 1989, pp. 104–111). I use the numerical optimization engine for solving nonlinear models evaluated in Zahery et al. (2017) to optimize the likelihood function. I compute the standard errors using the Delta method. The dynamic models defined in Equation (7) and (11) start with initial conditions. I treat these initial conditions as predetermined.

5. Data

I extract a sample of 11,530 children from the NLSY79 Children and Young Adults dataset. I merge those data with data from the female respondents in the NLSY79. Furthermore, following Cunha et al. (2010), I focus only on firstborn children (6,593 children dropped). Lastly, I drop children whose age is not observed on the assessment day (3,028 children dropped). After these restrictions, I have an analytical sample of 1,909 firstborn children. See the Center for Human Resource Research (2009) for more information concerning validity, reliability, and testing procedures.

Since 1986, interviewers have assessed the children of female respondents participating in the NLSY79. Interviewers are experienced and specifically trained to interview the children. I consider assessments that measure several aspects of the child's cognitive achievement and the quality of their home environment. The interviewers have collected data via direct assessment and maternal reports during home visits once every two years (*i.e.*, biennially). I use scores on these measurements from five assessment waves. Lastly, in 1980, interviewers administered several tests to assess the mother's cognitive achievement. I also use these measures.

Appendix A.5 shows an overview of observed and missing data. Following Cunha & Heckman (2008) and Cunha et al. (2010), I assume that the data are missing at random. Missing at random means that the probability a data point is missing does not depend on the value of the missing data point but only on available information. In other words, conditional on the observed data, we assume to have enough information to ignore the missing data mechanism. I can then apply the full-information maximum likelihood methodology to account for missing values (Anderson 1957).

5.1. Unequally Spaced Observation Intervals

Following Cunha et al. (2010) and Agostinelli & Wiswall (2016), I aggregate child age as follows: (T1) age 5 and 6, (T2) age 7 and 8, (T3) age 9 and 10, (T4) age 11 and 12, and (T5) age 13 and 14. By aggregating child age into five assessment waves (*i.e.*, T1, T2, T3, T4, and T5), it would appear that I have constructed equal intervals. While the distance between T1 (age 5–6) and T2 (age 7–8) is similar at an aggregate level to the difference between T2 (age 7–8) and T3 (age 9–10), it ignores age variation on the assessment day.

What is more, such aggregations raise further questions. Should we aggregate five- and six-year-olds as in Cunha et al. (2010) and Agostinelli & Wiswall (2016), or should we follow Cunha & Heckman (2008) and aggregate six- and seven-year-olds? Also, if age is the relevant timing variable and intervals are child-specific, then parameter estimates based on the time distance between T1 (age 5–6) and T2 (age 7–8) cannot be compared with those obtained based on T2 (age 7–8) and T3 (age 9–10).

Figure 1 and Figure 2 reveal the extent of unequally spaced intervals. For each of the five waves, I observe the child's age (in months) on the assessment day. I obtain the time distances by subtracting the age variable in one assessment wave from the age variable in the previous wave. I plot the probability frequency distributions of these distances in Figure 1. If the intervals are equal, there should be a single bar in the plots at two years because the NLSY79 assesses each child biennially. Consider the first plot. The time distance between the first and second measurements was 1.5 years for some and 2.5 years for others. At these early ages, such differences are meaningful (*e.g.*, Crawford et al. 2010).

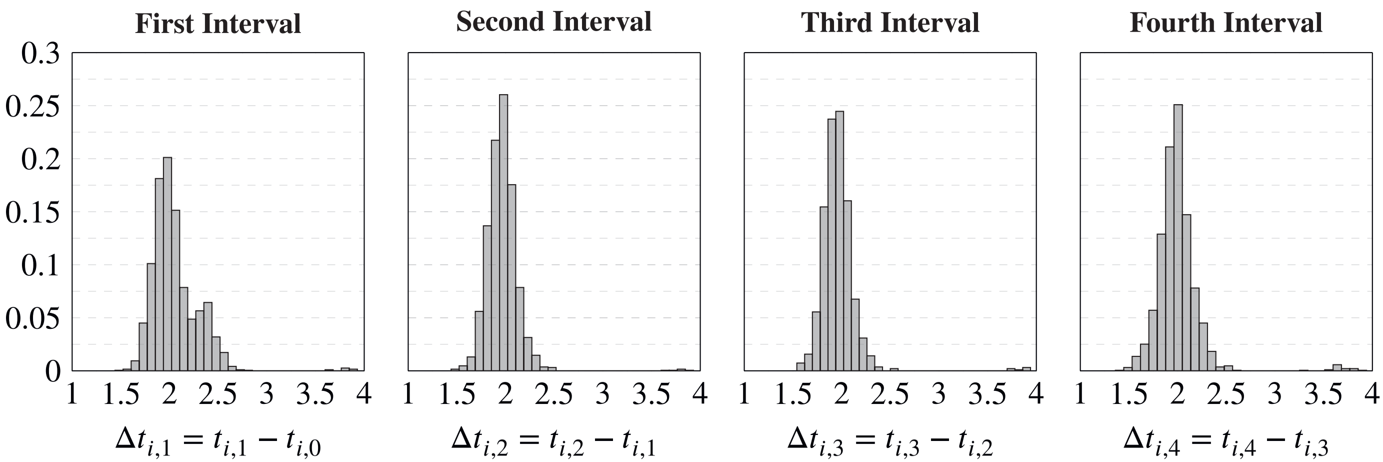

Figure 1. Probability Frequency Distributions for Child-Specific Unequal Observation Intervals

Notes. This figure shows variation in child-specific unequal observation intervals (*i.e.*, $\Delta t_{i,a} \equiv t_{i,a} - t_{i,a-1}$) for the children from the NLSY79 Children and Young Adult dataset. I aggregate child age as follows: (T1) age 5 and 6, (T2) age 7 and 8, (T3) age 9 and 10, (T4) age 11 and 12, and (T5) age 13 and 14. The plots represent probability frequency distributions of the time distance (in years) between the five assessment waves. They y-axis is the same for each plot.

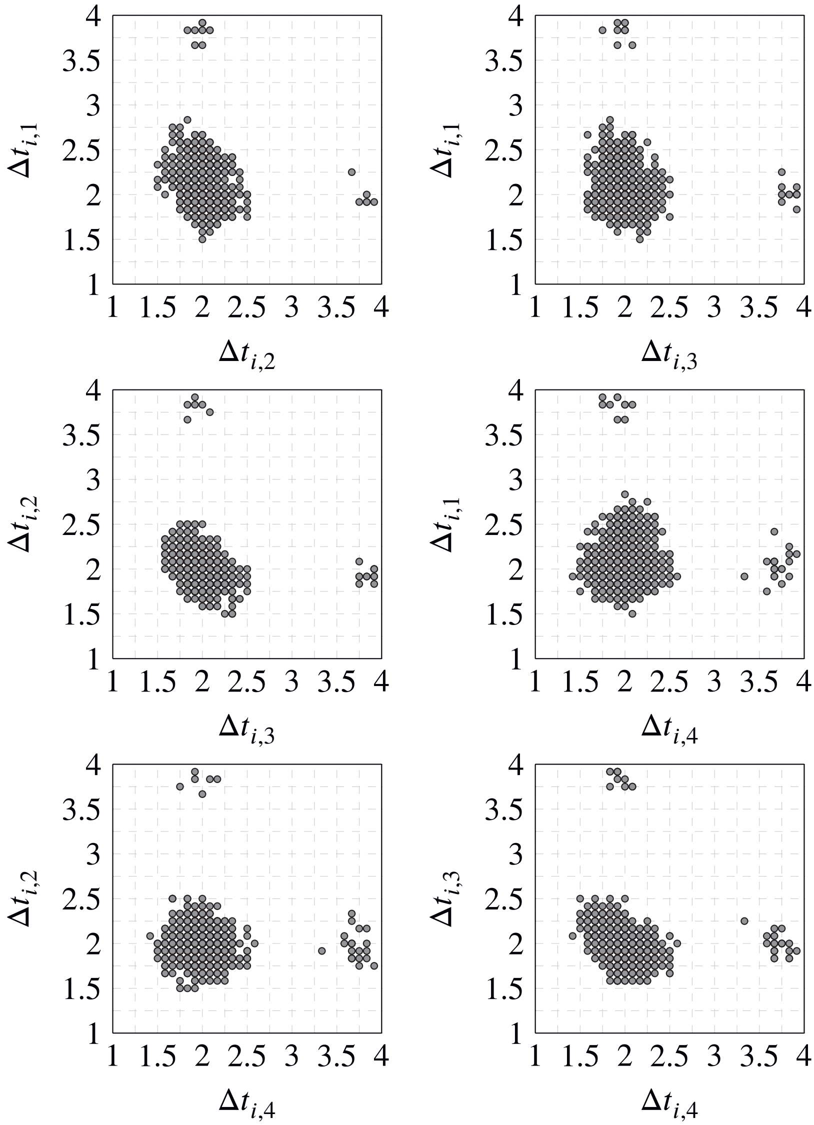

Figure 2. Degree of Correlations for Unequal Observation Intervals

Notes. This figure shows the degree of correlations for child-specific unequal observation intervals (*i.e.*, $\Delta t_{i,a} \equiv t_{i,a} - t_{i,a-1}$) for the children from the NLSY79 Children and Young Adult dataset. I aggregate child age as follows: (T1) age 5 and 6, (T2) age 7 and 8, (T3) age 9 and 10, (T4) age 11 and 12, and (T5) age 13 and 14. The scatter plots show the relationship between child-specific distances between various assessment waves.

The scatter plots in Figure 2 provide insight into the relationship between the temporal distances in Figure 1. It shows whether children for whom the distance between the first and second waves is small are also those for whom the distance between the second and third waves is small. If the intervals are equal, then there should be a single scatter dot. These scatter plots should have a similar pattern if the intervals are child-specific, and children are assessed with regularity. Neither scenario seems to be the case, as evidenced by a lack of consistency in the patterns across the scatter plots.

5.2. Measures of Children's Cognitive Skill

The NLSY79 Children and Young Adult dataset includes several measurements that pertain to cognitive achievement. I use three subtests of the Peabody Individual Achievement Test (PIAT). The PIAT is a wide-ranging measure of cognitive achievement for children aged five and older. I use three subtests: (1) mathematics, (2) reading recognition, and (3) reading comprehension. In all waves, I standardize these subtests by subtracting the mean and dividing by the standard deviation, both calculated over the total sample period to maintain differences over time.

The mathematics subtest measures attainment in mathematics as taught in mainstream education. The mathematics subtest consists of 84 multiple-choice items that increase in difficulty, starting with basic skills such as recognizing numerals and ending with advanced concepts in geometry and trigonometry. The reading recognition subtest (84 items) measures word recognition and pronunciation ability. Children first read a word silently and then say it aloud. Each item has four response alternatives. Like the mathematics subtest, the reading recognition subtest increases in difficulty. Lastly, the reading comprehension subtest (66 items) measures a child's ability to derive meaning from sentences read silently. Children who score below 19 points on the reading recognition subtest receive the score they earned on the reading recognition subtest as their score for the reading comprehension subtest.

Figure 3 shows descriptives calculated over all non-missing observations for each of the three subtests. I use the unstandardized scores and plot the mean and standard deviation across all five assessment waves (T1: age 5–6, T2: age 7–8, T3: age 9–10, T4: age 11–12, and T5: age 13–14).

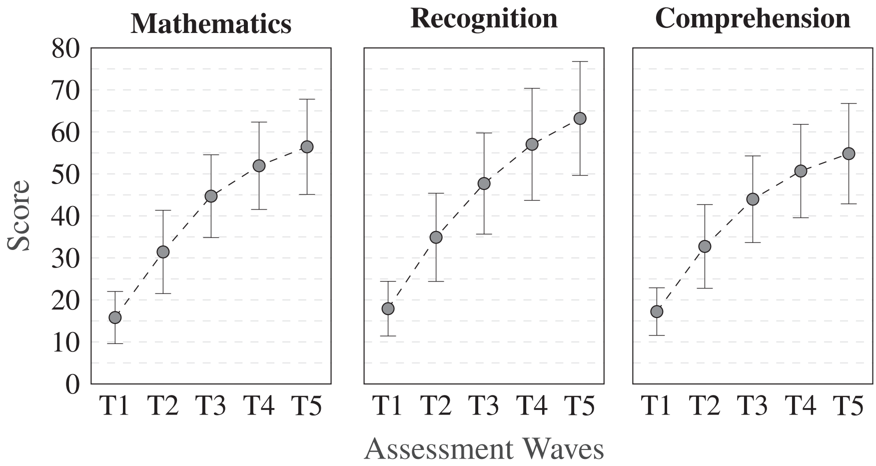

Figure 3. Mean Score by Assessment wave for the PIAT Subtests

Notes. This figure shows the mean score for each PIAT subtests (*i.e.*, mathematics, reading recognition, and reading comprehension) calculated over all non-missing observations by assessment wave. I use the raw scores rather than the standardized scores. Child age in each of the assessment waves is as follows: (T1) age 5 and 6, (T2) age 7 and 8, (T3) age 9 and 10, (T4) age 11 and 12, and (T5) age 13 and 14. Appendix A.5 presents an overview of observed and missing values. The *y*-axis is the same for each plot.

The length of the whiskers signifies one standard deviation above (and below) the mean. Figure 3 shows strong mean development across all three subtests. I use the PIAT reading recognition subtest to set the scale and location (see Section 3.3.1).

5.3. Measures of the Quality of the Home Environment

The NLSY79 Children and Young Adults dataset also includes a measure of the quality of the child's home environment: the Home Observation Measurement of the Environment -– Short Form (HOME-SF). The HOME-SF is a subset of the measures used to construct the HOME scale designed by Bradley & Caldwell (1980, 1984) to assess children's emotional support and their cognitive stimulation in the home environment, planned events, and family surroundings.

The set of items used in this article overlaps to a large extent with Cunha et al. (2010). I use the following items from the HOME-SF: (i) How often mother reads to the child, (ii) Does the child engage in special lessons or do extracurricular activities, (iii) Does the family get a daily newspaper, (iv) How often has the child gone to music/theater performances in the past year, and (v) Whether there is a musical instrument the child can use. I use the same standardization as the PIAT subtests for each item. The following item is used as an anchor: Does the child engage in special lessons or do extracurricular activities (see Section 3.3.1).

When children are around the age of 5 to 6, the median number of times mothers read to their child is three times a week. At later ages, this number drops to once a week. Second, the percentage of children who receive special lessons (or engage in extracurricular activities) increases from 50 percent between the ages of 5 and 8 to 70 percent between the ages of 9 and 14. Third, about half of the sampled families receive a daily newspaper. Fourth, the median number of times children have visited a theater or musical performance in the past year is once or twice. Lastly, about 40 percent of five-to-eight-year-olds have a musical instrument at home they can use. This percentage increases to 50 percent by the age of 9 to 10 and to 60 percent by the age of 11 to 14.

5.4. Measures of Mother's Cognitive Skill and Family Background Characteristics

In 1980, interviewers also administered a battery of tests to the mothers from the Armed Services Vocational Aptitude Battery (ASVAB) (U.S. Department of Defense 1982). Like Cunha et al. (2010), I use the six tests from the ASVAB: (i) arithmetic reasoning (Mean: 14.8, *SD*: 6.7), (ii) word knowledge (Mean: 23.0, *SD*: 8.1), (iii) paragraph composition (Mean: 10.0, *SD*: 3.6), (iv) numerical operations (Mean: 33.0, *SD*: 11.4), (v) coding speed (Mean: 45.2, *SD*: 16.8), and (vi) mathematical knowledge (Mean: 11.7, *SD*: 5.9). I use the word knowledge test as anchor (see Section 3.3.1).

I separately estimate the measurement model (Equation 10) for the mother's cognitive skill to simplify estimation. I then use the estimated measurement model to predict factor scores using regression factor scoring. I treat those factor scores as observed in the model. See Appendix A.6 for further details.

In addition to measures of mothers' cognitive achievement, I also collect information on the child's sex, race (*i.e.*, white, Black, or Hispanic), and birth year. About 51 percent of the sample is female, 29 percent is Black, and about 19 percent is Hispanic. I include these covariates in measurement Equation (8) and Equation (9). Also, I collect data on total (net) family income in the past calendar year. I scale family income in 2000 dollars. Average family income over the ages analyzed is 52,723 dollars (*SD*: 84,392 dollars).

6. Estimating the Technology of Cognitive Skill Formation

I report three sets of results in this section. Section 6.1 discusses the effect of accounting for unequally spaced intervals on the estimated technology and investment function. The reported estimates in this section are based on Equation (7) and Equation (11), respectively. I assume two stages of development. The first stage starts at age 5-6 and ends at age 9-10 (*i.e.*, T1 through T3). The second stage starts at age 9-10 and ends at age 13-14 (*i.e.*, T3 through T5). Section 6.2 reports estimates based on Equation (12) and Equation (13), respectively. Section 6.3 provides insight into the sensitivity of defining developmental stages in a particular manner. Section 6.4 discusses robustness.

I focus on parameters affected by unequally spaced intervals only. For space considerations, I do not report the estimates related to the measurement models, initial conditions, and family income process as these are unaffected by unequally spaced intervals in the current setup. These results are available on request.

6.1. The Effect of Accounting for Unequal Intervals on the Estimated Technology

Table 2 reports the point estimates and standard errors (in parentheses) for the estimated technology equations based on Equation (7). Columns (1) and (3) report estimates based on equal intervals, whereas Columns (2) and (4) report estimates based on unequal intervals. The variables presented in the rows refer to the technology inputs (Equation 7) as described in Section 3.

The findings in Table 2 can be summarized as follows. First, the child's stock of cognitive skills exhibits strong self-productivity in Stage 1 and Stage 2. Strong self-productivity indicates that children who have a high stock of cognitive skills at one point in time are also those with a high stock of cognitive skills later. Cognitive skills thus self-reinforce cognitive skills. In particular, I estimate a self-productivity parameter of about 0.891 in Stage 1 and 1.000 in Stage 2. Note that these estimates are not directly comparable if the intervals that make up these stages differ, which Figure 1 and Figure 2 suggests is the case.

Table 2. Cognitive Skill Production Function Parameters

	Stage 1		Stage 2	
	(1) Equal Intervals	(2) Unequal Intervals	(3) Equal Intervals	(4) Unequal Intervals
Cognitive Skills	0.891	0.960	1.000	0.990
	(0.019)	(0.010)	(0.016)	(0.008)
Parental Investments	0.086	0.045	0.017	0.013
	(0.022)	(0.012)	(0.018)	(0.009)
Mother's Cognitive Skills	0.024	0.008	0.012	−0.001
	(0.012)	(0.006)	(0.011)	(0.005)
Total Factor Productivity	0.789	0.349	0.346	0.177
	(0.032)	(0.015)	(0.029)	(0.014)
Random Shock	0.098	0.048	0.044	0.015
	(0.008)	(0.004)	(0.009)	(0.003)

Notes. This table reports the parameter estimates and standard errors in parentheses. The standard errors are computed using the Delta method. The estimated technology equations are:

$$\mathrm{S}_{i,t} = \alpha_{0,k} + \alpha_{1,k}\mathrm{S}_{i,t-1} + \alpha_{2,k}\mathrm{I}_{i,t-1} + \alpha_{3,k}\mathrm{S}_{\mathrm{P},i} + \epsilon_{i,t}.$$

The model is estimated using full-information maximum likelihood (1,909 observations). Columns (1) and (2) relate to developmental stage 1 (starts at age 5-6 and ends at age 9-10). Columns (3) and (4) relate to developmental stage 2 (starts at age 9-10 and ends at age 13-14). Columns (1) and (3) are based on equal intervals. Columns (2) and (4) are based on unequal intervals.

When I account for unequal intervals, these parameter estimates become 0.960 and 0.990, respectively. Accounting for unequal spacing suggests that the temporal stability of children's cognitive skills changes less in the first developmental and slightly more in the second developmental stage than when we assume equal intervals. More strikingly, accounting for unequally spaced intervals improves precision substantially, as evidenced by the smaller standard errors in Columns (2) and (4), compared with Columns (1) and (3). The improved precision is consistent with the Monte Carlo simulation results in Voelkle & Oud (2013). In sum, when I do not account for unequally spaced intervals, self-productivity appears stronger at earlier ages and weaker at later ages. In all cases, the precision is greater when I account for unequal intervals.

Second, I find that parental investments are particularly effective in Stage 1, indicating that a sensitive period for cognitive skill investments occurs earlier in the child's life cycle (*cf.* Cunha & Heckman 2008, Cunha et al. 2010). As in the case of self-productivity, the conclusion that parental investments are particularly effective in Stage 1 cannot technically be based on the comparison of Columns (1) and (3) if the intervals maintained in these stages differ. I find that a one-standard-deviation increase in parental investments in Stage 1 increases children's cognitive skills by 0.086 percent of a standard deviation. When I account for unequal intervals, this effect drops to 0.045 percent of a standard deviation, suggesting a percentage difference of 62.6-percent between estimates that assume equal intervals and estimates that incorporate unequal spacing. While parental investments appear to be no longer effective in Stage 2, I do find differences in the point estimates depending on whether unequally spaced intervals are incorporated. In sum, when I do not account for unequally spaced intervals, investments appear to be more effective than they are, given that intervals are in fact unequally spaced.

Third, I find that children exhibit strong (residual) cognitive development during Stage 1, as evidenced by the total factor productivity of 0.789 (Column 1). However, when I account for unequally spaced intervals, this point estimate drops to 0.349, representing a percentage difference of 77.3-percent. In Stage 2, children continue to exhibit (residual) cognitive development, though to a lesser extent as the total factor productivity drops from 0.789 in Stage 1 to 0.346 in Stage 2 (Column 3). When I account for unequal spacing, I estimate a parameter of 0.177, representing a percentage difference of 64.6-percent relative to the 0.346 that did not account for unequal intervals. In sum, when I do not account for unequally spaced intervals, mean (residual) cognitive growth appears larger than it is, given that intervals are unequally spaced.

Lastly, I find that the variance of the random shock is also affected. Consider Stage 1: The variance of the random shock is 0.098 (0.313 *SD*) in Column (1) and 0.048 (0.219 *SD*) in Column (2), representing a percentage difference of 68.5-percent (35.3-percent based on *SD*). For Stage 2, the variance of the random shock is 0.044 (0.210 *SD*) in Column (3) without accounting for unequally spaced intervals and 0.015 (0.122 *SD*) in Column (4), representing a percentage difference of 98.3-percent (53.0-percent based on *SD*). In sum, when I do not account for unequally spaced intervals, the variance of the random shock appears larger than it is, given that intervals are unequally spaced.

Table 3 reports the results of the investment function (Equation 11), which can be summarized as follows: First, there seems to be strong persistence in the investment behavior of parents. The variation in parental investment in the current period seems to determine the variation in parental investment in the next period to a large extent. This persistence is comparable across stages and reflects temporal stability (*e.g.*, habits) with respect to how parents invest in their children. In Stage 1, the temporal stability is 0.731, and in Stage 2, the temporal stability is 0.741. Like the self-reinforcement of cognitive skills in Table 2, these estimates of temporal stability in parental investment are not directly comparable if the intervals that make up these stages differ. When I account for unequally spaced intervals, the degree of temporal persistence increases from 0.731 and 0.741 to 0.861 and 0.856, respectively. Also, the degree of precision changes significantly. In sum, accounting for the fact that intervals are unequally spaced increases the temporal stability of parental investments.

Mothers' cognitive skills positively affect the quality of the home environment in Stage 2. Assuming equal intervals, I estimate an effect of 0.057 of a standard deviation. When I consider the unequal spacing, this effect drops to 0.028, a percentage difference of 68.2-percent. When I do not account for unequal intervals, the effect of mothers' cognitive skills appears more important than it is given unequally spaced intervals.

Table 3. Parental Investment Function Parameters

	Stage 1		Stage 2	
	(1) Equal Intervals	(2) Unequal Intervals	(3) Equal Intervals	(4) Unequal Intervals
Cognitive Skills	0.026 (0.027)	0.015 (0.015)	−0.007 (0.022)	0.001 (0.014)
Parental Investments	0.731 (0.045)	0.861 (0.026)	0.741 (0.048)	0.856 (0.028)
Mother's Cognitive Skills	0.021 (0.017)	0.012 (0.009)	0.056 (0.020)	0.028 (0.010)
Family Income (log)	0.090 (0.046)	0.097 (0.049)	0.143 (0.057)	0.140 (0.056)
Random Shock	0.059 (0.023)	0.034 (0.014)	0.095 (0.034)	0.054 (0.020)

Notes. This table reports the parameter estimates and standard errors in parentheses. The standard errors are computed using the Delta method. The estimated technology equations are:

$$\mathrm{I}_{i,t} = \beta_{1,k}\mathrm{S}_{i,t-1} + \beta_{2,k}\mathrm{I}_{i,t-1} + \beta_{3,k}\mathrm{S}_{\mathrm{P},i} + \beta_{4,k}\ln \mathrm{y}_{i,t-1} + \eta_{i,t}.$$

The model is estimated using full-information maximum likelihood (1,909 observations). Columns (1) and (2) relate to developmental stage 1 (starts at age 5-6 and ends at age 9-10). Columns (3) and (4) relate to developmental stage 2 (starts at age 9-10 and ends at age 13-14). Columns (1) and (3) are based on equal intervals. Columns (2) and (4) are based on unequal intervals.

Like the technology estimates in Table 2, the variance of the random shock is also affected by unequal spacing. Consider Stage 1: The variance of the random shock is 0.059 (0.243 *SD*) in Column (1) and 0.034 (0.184 *SD*) in Column (2), representing a percentage difference of 53.8-percent. For Stage 2, the variance is 0.095 (0.308 *SD*) in Column (3) without accounting for unequally spaced intervals and 0.054 (0.232 *SD*) in Column (4), representing a percentage difference of 55.0-percent. When I do not account for unequal intervals, the variance of the random shock appears larger than it is, given that intervals are unequally spaced.

6.2. The Effect of Unequal Intervals on Technology Estimated Allowing for Time-invariant Heterogeneity

Table 4 reports the point estimates and standard errors (in parentheses) for the estimated technology equations that allow for time-invariant heterogeneity (Equation 12). Like Table 2, Columns (1) and (3) report estimates based on equal intervals, whereas Columns (2) and (4) report estimates based on unequal intervals. The variables presented in the rows refer to the technology inputs (Equation 12). Note that the parameters have a different interpretation than those estimated in Table 2 because they measure *within-child* effects.

The findings in Table 4 can be summarized as follows: In Stage 1, I still find that the child's stock of cognitive skills exhibits strong self-productivity. In particular, I estimate a self-productivity parameter of about 0.776 in Stage 1. When I account for unequal intervals, this parameter becomes 0.821. In Stage 2, I find a reduction in self-productivity, suggesting less within-child self-reinforcement of cognitive skills, conditional on child-specific time-invariant heterogeneity. I estimate self-productivity parameters of about 0.417 (Column 3) and 0.414 (Column 4). The drop likely occurs because the time-invariant component is capturing a lot of the temporal stability (*cf.* Table 2). In sum, the contrast between parameters based on equal and unequally spaced intervals is consistent with Table 2. First, when I do not account for unequally spaced intervals, the self-reinforcement of cognitive skills appears to be less than it is, given that intervals are unequally spaced. Second, the precision increases.

Furthermore, I find that parental investments are most effective in Stage 1, particularly when I account for stable differences between children, indicating that this stage is a sensitive period for cognitive skill investments. When I account for unequal intervals, the effect of parental investments drops from 0.168 to 0.144 in Stage 1 and from 0.049 to 0.020 in Stage 2, representing percentage differences of 15.4-percent and 84.1-percent. The findings on investments in Table 4 are largely consistent with those reported in Table 2. When I do not account for unequally spaced intervals, investments appear to be more effective than they are, given that intervals are unequally spaced. I also find large differences in mean (residual) cognitive development and the variance of the shock between Stage 1 and Stage 2 when I compare estimates based on equal and unequal intervals.

In Appendix A.7, I report the results concerning the investment function with time-invariant heterogeneity (Equation 13). The findings can be summarized as follows: First, the child's cognitive skills affect the parent's investment decisions in Stage 1. In particular, parents appear to reinforce cognitive skills with further investments. While accounting for unequally spaced intervals has little impact on the point estimate, it does affect precision. Second, there seems to be strong persistence in the investment behavior of parents, especially in Stage 1. In Stage 2, parental investment's temporal stability drops, much like children's cognitive skills in Table 4 .

Table 4. Cognitive Skill Production Function Parameters with Time-Invariant Unobserved Heterogeneity

	Stage 1		Stage 2	
	(1) Equal Intervals	(2) Unequal Intervals	(3) Equal Intervals	(4) Unequal Intervals
Cognitive Skills	0.776	0.821	0.417	0.414
	(0.044)	(0.027)	(0.019)	(0.013)
Parental Investments	0.168	0.144	0.049	0.020
	(0.038)	(0.017)	(0.024)	(0.016)
Total Factor Productivity	0.758	0.373	0.454	0.362
	(0.060)	(0.027)	(0.037)	(0.029)
Time-Invariant Shock	0.020	0.016	0.267	0.275
	(0.006)	(0.004)	(0.017)	(0.014)
Random Shock	0.086	0.036	0.024	0.014
	(0.013)	(0.006)	(0.010)	(0.007)

Notes. This table reports the parameter estimates and standard errors in parentheses. The standard errors are computed using the Delta method. The estimated technology equations are:

$$\mathrm{S}_{i,t} = \gamma_{0,k} + \gamma_{1,k}\mathrm{S}_{i,t-1} + \gamma_{2,k}\mathrm{I}_{i,t-1} + \kappa_{\mathrm{S},i} + \xi_{\mathrm{S},i,t}.$$

The model is estimated using full-information maximum likelihood (1,909 observations). Columns (1) and (2) relate to developmental stage 1 (starts at age 5-6 and ends at age 9-10). Columns (3) and (4) relate to developmental stage 2 (starts at age 9-10 and ends at age 13-14). Columns (1) and (3) are based on equal intervals. Columns (2) and (4) are based on unequal intervals.

The temporal stability of parental investments appears smaller than it is, given that intervals are unequally spaced. Lastly, when I do not account for unequally spaced intervals, the variance of the random shock is larger than when I do account for unequally spaced intervals.

6.3. The Effect of Unequal Intervals on Period-by-Period Technology Estimates

Table 5 shows the impact of accounting for unequally spaced intervals by estimating technology equations period by period. The findings further support the findings reported in Table 2 and Table 4, suggesting that the estimates are not the result of how I defined the developmental stages. When we account for unequally spaced intervals, a number of things hold true: First, self-productivity estimates are larger. Second, the precision of self-productivity estimates is larger. Third, the effect of investments is lower. Fourth, the effect of a mother's cognitive skill is (generally) lower. Fifth, the effect of total factor productivity is always lower. Lastly, the error variance of the shock is always lower.

6.4. Robustness Analyses

I conduct several robustness analyses (see Appendix A.7). First, I estimate the cognitive skill production function and investment equation without measurement error correction (see Table A.7.9 through Table A.7.12). Second, Figure 1 and Figure 2 show that the time distance between assessment waves is particularly large (or small) for some children (i.e., $\Delta t_{i,a} > 2.5$ and $\Delta t_{i,a} < 1.5$). To investigate the extent to which these children drive my results, I recode these intervals to be minimum 1.5 and maximum 2.5 (see Table A.7.13 and Table A.7.14). Third, I dropped children for whom I did not observe their age at the time of assessment in each wave. Therefore, the same children are included in each stage. Alternatively, I can construct stage-specific samples and only drop children for whom I do not observe their age at the time of assessment in that stage (see Table A.7.15 and Table A.7.16). Finally, I examine if changing the "anchors" changes the results (see Table A.7.17 through Table A.7.20). None of these analyses change the conclusions, though the magnitude of estimates varies.

Table 5. Cognitive Skill Production Function Parameters: Period-by-Period

	Stage 1		Stage 2		Stage 3		Stage 4	
	(1) Equal	(2) Unequal	(3) Equal	(4) Unequal	(5) Equal	(6) Unequal	(7) Equal	(8) Unequal
Cognitive Skills	0.888	0.957	0.923	0.957	0.942	0.969	0.994	1.001
	(0.036)	(0.019)	(0.024)	(0.013)	(0.028)	(0.015)	(0.033)	(0.017)
Parental Investments	0.178	0.100	0.003	0.007	0.048	0.027	0.011	0.004
	(0.039)	(0.021)	(0.028)	(0.016)	(0.027)	(0.016)	(0.030)	(0.016)
Mother's Cognitive Skills	0.049	0.021	-0.011	-0.001	-0.002	-0.002	0.022	0.009
	(0.019)	(0.009)	(0.016)	(0.009)	(0.019)	(0.009)	(0.019)	(0.009)
Total Factor Productivity	0.871	0.336	0.580	0.314	0.409	0.206	0.248	0.127
	(0.050)	(0.021)	(0.043)	(0.021)	(0.049)	(0.024)	(0.056)	(0.027)
Random Shock	0.116	0.058	0.055	0.032	0.055	0.028	0.022	0.011
	(0.012)	(0.006)	(0.009)	(0.005)	(0.013)	(0.007)	(0.010)	(0.005)

Notes. This table reports the parameter estimates and standard errors in parentheses. The standard errors are computed using the Delta method. The estimated technology equations are:

$$\mathrm{S}_{i,t} = \alpha_{0,k} + \alpha_{1,k}\mathrm{S}_{i,t-1} + \alpha_{2,k}\mathrm{I}_{i,t-1} + \alpha_{3,k}\mathrm{S}_{\mathrm{P},i} + \epsilon_{i,t}.$$

The model is estimated using full-information maximum likelihood (1,909 observations). Columns (1) and (2) relate to developmental stage 1 (starts at age 5-6 and ends at age 7-8). Columns (3) and (4) relate to developmental stage 2 (starts at age 7-8 and ends at age 9-10). Columns (5) and (6) relate to developmental stage 3 (starts at age 9-10 and ends at age 11-12). Columns (7) and (8) relate to developmental stage 4 (starts at age 11-12 and ends at age 13-14). Columns (1), (3), (5), and (7) are based on equal intervals. Columns (2), (4), (6), and (8) are based on unequal intervals.

7. Conclusion

I have evaluated the extent to which unequal intervals affect estimates of the technology parameters that govern cognitive development, using an often-used dataset for studying cognitive development. Despite a consensus that it is important to account for unequal intervals, little empirical work has studied the effect such intervals have on the parameter estimates that are crucial to understanding children's cognitive skill formation. By applying insights from the continuous-time modeling literature, I evaluated this impact. I found that failing to account for unequal intervals affects the strength and precision of estimated parameters. Except in the case of (temporal) persistence (*i.e.*, self-productivity of cognitive skills and temporal dependence of parental investment), effects are generally smaller when unequal intervals are considered. These findings have implications for the econometrics of early human capital development (*cf.* Cunha et al. 2021).

Consistent with a growing literature on econometric methods, my findings imply that failing to account for unequal intervals can be problematic. As explained at the beginning of this article, (child-specific) unequal intervals affect all parameters. Such unequal intervals are probable in most settings and might arise as a result of survey-design decisions or age differences at the time of assessment. The extent to which this is problematic will likely depend on the data, and in particular on the type of skill being modeled, age differences, and the regularity of sampling. Fortunately, most experimental studies are likely to have information on child assessment dates and age. Such information can then be usefully combined with the insights from the continuous-time modeling literature applied in this paper (see, *e.g.*, Thijssen et al. 2022). When one does not account for (child-specific) unequal intervals, it becomes impossible to compare the strength of parameter estimates within the same study and across different studies, when estimation is based on samples that vary in the degree of unequally spaced intervals. Such inability to compare findings severely hinders consilience.

References

Agostinelli, F. & Wiswall, M. (2016), Estimating the technology of children's skill formation, Working Paper 22442, National Bureau of Economic Research, Cambridge, Massachusetts.

Alwin, D. F. (1989), 'Problems in the estimation and interpretation of the reliability of survey data', *Quality and Quantity* **23**(3), 277–331.

Anderson, T. W. (1957), 'Maximum likelihood estimates for a multivariate normal distribution when some observations are missing', *Journal of the American Statistical Association* **52**(278), 200–203.

Anderson, T. W. & Rubin, H. (1956), *Statistical Inference in Factor Analysis*, Vol. 5 of *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 5: Contributions to Econometrics, Industrial Research, and Psychometry*, University of California Press, Berkeley, California.

Arnold, L. (1974), *Stochastic Differential Equations: Theory and Applications*, John Wiley & Sons, Inc., New York City, New York.

Attanasio, O., Cattan, S., Fitzsimons, E., Meghir, C. & Rubio-Codina, M. (2020), 'Estimating the production function for human capital: Results from a randomized controlled trial in Colombia', *American Economic Review* **110**(1), 48–85.

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

Balestra, P. & Nerlove, M. (1966), 'Pooling cross section and time series data in the estimation of a dynamic model: The demand for natural gas', *Econometrica* **34**(3), 585–612.

Baltagi, B. H. & Wu, P. X. (1999), 'Unequally spaced panel data regressions with AR(1) disturbances', *Econometric Theory* **15**(6), 814–823.

Becker, G. S. & Tomes, N. (1976), 'Child endowments and the quantity and quality of children', *Journal of Political Economy* **84**(4), S143–S162.

Ben-Moshe, D. (2012), Essays on Nonparametric Identification, Doctoral thesis.

Ben-Moshe, D. (2018), ‘Identification of joint distributions in dependent factor models’, *Econometric Theory* **34**(1), 134–165.

Bergstrom, A. R. (1988), ‘The history of continuous-time econometric models’, *Econometric Theory* **4**(3), 365–383.

Bollen, K. A. (1989), *Structural Equations with Latent Variables*, John Wiley & Sons, Inc., New York City, New York.

Borghans, L., Duckworth, A. L., Heckman, J. J. & ter Weel, B. (2008), ‘The economics and psychology of personality traits’, *Journal of Human Resources* **43**(4), 972–1059.

Bradley, R. H. & Caldwell, B. M. (1980), ‘The relation of home environment, cognitive competence, and IQ among males and females’, *Child Development* **51**(4), 1140–1148.

Bradley, R. H. & Caldwell, B. M. (1984), ‘The relation of infants’ home environments to achievement test performance in first grade: A follow-up study’, *Child Development* **55**(3), 803–809.

Cawley, J., Heckman, J. & Vytlacil, E. (2001), ‘Three observations on wages and measured cognitive ability’, *Labour Economics* **8**(4), 419–442.

Center for Human Resource Research (2009), *NLSY79 Child and Young Adult Data User’s Guide*, Center for Human Resource Research, Ohio State University, Columbus, Ohio.
URL: *https://search.library.wisc.edu/catalog/9910094201202121*

Crawford, C., Dearden, L. & Meghir, C. (2010), When you are born matters: The impact of date of birth on educational outcomes in England, Working Paper 10,06, Institute for Fiscal Studies, London, England.

Cunha, F. & Heckman, J. (2007), ‘The technology of skill formation’, *American Economic Review* **97**(2), 31–47.

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

Cunha, F., Nielsen, E. & Williams, B. (2021), 'The econometrics of early childhood human capital and investments', *Annual Review of Economics* **13**, 487–513.

Del Boca, D., Flinn, C. & Wiswall, M. (2014), 'Household choices and child development', *Review of Economic Studies* **81**(1), 137–185.

Driver, C. C. Voelkle, M. C. (2018), 'Hierarchical bayesian continuous time dynamic modeling.', *Psychological Methods* **23**(4), 774.

Elder, T. E. & Lubotsky, D. H. (2009), 'Kindergarten entrance age and children's achievement: Impacts of state policies, family background, peers', *Journal of Human Resources* **44**(3), 641–683.

Hamerle, A., Nagl, W. & Singer, H. (1991), 'Problems with the estimation of stochastic differential equations using structural equations models', *The Journal of Mathematical Sociology* **16**(3), 201–220.

Hamerle, A., Singer, H. & Nagl, W. (1993), 'Identification and estimation of continuous time dynamic systems with exogenous variables using panel data', *Econometric Theory* **9**(2), 283–295.

Heckman, J. J. & Mosso, S. (2014), 'The economics of human development and social mobility', *Annual Review of Economics* **6**, 689–733.

Heckman, J., Pinto, R. & Savelyev, P. (2013), 'Understanding the mechanisms through which an influential early childhood program boosted adult outcomes', *American Economic Review* **103**(6), 2052–2086.

Holzinger, K. J. & Swineford, F. (1937), 'The bi-factor method', *Psychometrika* **2**(1), 41–54.

Hu, Y. & Schennach, S. M. (2008), 'Instrumental variable treatment of nonclassical measurement error models', *Econometrica* **76**(1), 195–216.

Jones, R. H. & Boadi-Boateng, F. (1991), 'Unequally spaced longitudinal data with AR(1) serial correlation', *Biometrics* **47**(1), 161–175.

Kalil, A. (2015), *Inequaliy Begins at Home: The Role of Parenting in the Diverging Destinies of Rich and Poor Children*, Vol. 5 of *National Symposium on Family Issues*, 1 edn, Springer International Publishing, Cham, Germany, book section 5, pp. 63–82.

Kotlarski, I. (1967), 'On characterizing the gamma and the normal distribution', *Pacific Journal of Mathematics* **20**(1), 69–76.

Li, T. & Vuong, Q. (1998), 'Nonparametric estimation of the measurement error model using multiple indicators', *Journal of Multivariate Analysis* **65**(2), 139–165.

Little, R. J. A. & Rubin, D. B. (2002), *Statistical Analysis with Missing Data*, Wiley Series in Probability and Statistics, 2 edn, John Wiley & Sons, Inc., Hoboken, New Jersey.

McArdle, J. J. & McDonald, R. P. (1984), 'Some algebraic properties of the reticular action model for moment structures', *British Journal of Mathematical and Statistical Psychology* **37**(2), 234–251.

McKenzie, D. J. (2001), 'Estimation of AR(1) models with unequally spaced pseudo-panels', *The Econometrics Journal* **4**(1), 89–108.

Meredith, W. (1993), 'Measurement invariance, factor analysis and factorial invariance', *Psychometrika* **58**(4), 525–543.

Millimet, D. L. & McDonough, I. K. (2017), 'Dynamic panel data models with irregular spacing: With an application to early childhood development', *Journal of Applied Econometrics* **32**(4), 725–743.

Moler, C. & Van Loan, C. (2003), 'Nineteen dubious ways to compute the exponential of a matrix, twenty-five years later', *SIAM review* **45**(1), 3–49.

National Research Council Institute of Medicine (2000), *From neurons to neighborhoods: The science of early childhood development*, National Academy Press, Washington, D.C.

Neale, M. C., Hunter, M. D., Pritikin, J. N., Zahery, M., Brick, T. R., Kirkpatrick, R. M., Estabrook, R., Bates, T. C., Maes, H. H. & Boker, S. M. (2016), ‘OpenMx 2.0: Extended structural equation and statistical modeling’, *Psychometrika* **81**(2), 535–549.

Oud, J. H. L. & Jansen, R. A. R. G. (2000), ‘Continuous time state space modeling of panel data by means of SEM’, *Psychometrika* **65**(2), 199–215.

Rege, M., Størksen, I., Solli, I. F., Kalil, A., McClelland, M. M., ten Braak, D., Lenes, R., Lunde, S., Breive, S., Carlsen, M., Erfjord, I. & Hundeland, P. S. (2021), ‘The effects of a structured curriculum on preschool effectiveness: A field experiment’, *Journal of Human Resources* .

Rosner, B. & Muňoz, A. (1988), ‘Autoregressive modelling for the analysis of longitudinal data with unequally spaced examinations’, *Statistics in Medicine* **7**(1–2), 59–71.

Sasaki, Y. & Xin, Y. (2017), ‘Unequal spacing in dynamic panel data: Identification and estimation’, *Journal of Econometrics* **196**(2), 320–330.

Schennach, S. M. (2004*a*), ‘Estimation of nonlinear models with measurement error’, *Econometrica* **72**(1), 33–75.

Schennach, S. M. (2004*b*), ‘Nonparametric regression in the presence of measurement error’, *Econometric Theory* **20**(6), 1046–1093.

Singer, H. (1990), Parameterschätzung in Zeitkontinuierlichen Dynamischen Systemen, Doctoral thesis.

Singer, H. (1993), ‘Continuous-time dynamical systems with sampled data, errors of measurement and unobserved components’, *Journal of Time Series Analysis* **14**(5), 527–545.

Singer, H. (1995), ‘Analytical score function for irregularly sampled continuous time stochastic processes with control variables and missing values’, *Econometric Theory* **11**(4), 721–735.

Skrondal, A. & Laake, P. (2001), ‘Regression among factor scores’, *Psychometrika* **66**(4), 563–575.

Skrondal, A. & Rabe-Hesketh, S. (2004), *Generalized Latent Variable Modeling: Multilevel, Longitudinal, and Structural Equation Models*, Interdisciplinary Statistics, 1 edn, Chapman & Hall/CRC, Boca Raton, Florida.

Sylvia, S., Warrinnier, N., Luo, R., Yue, A., Attanasio, O., Medina, A. & Rozelle, S. (2020), ‘From quantity to quality: Delivering a home-based parenting intervention through China’s family planning cadres’, *Economic Journal* **131**(635), 1365–1400.

Thijssen, M. W. P., Rege, M., Solli, I. F. & Størksen, I. (2022), Cross-productivities of executive functions.

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

Todd, P. E. & 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.

Torgerson, W. S. (1958), *Theory and Methods of Scaling*, John Wiley & Sons, Inc., New York City, New York.

U.S. Department of Defense (1982), *Profile of American Youth: 1980 Nationwide Administration of the Armed Services Vocational Aptitude Battery*, Office of the Assistant Secretary of Defense for Manpower, Reserve Affairs, and Logistics, Washington, D.C.

Voelkle, M. C. & Oud, J. H. L. (2013), ‘Continuous time modelling with individually varying time intervals for oscillating and non-oscillating processes’, *British Journal of Mathematical and Statistical Psychology* **66**(1), 103–126.

Zahery, M., Maes, H. H. & Neale, M. C. (2017), 'CSOLNP: Numerical optimization engine for solving non-linearly constrained problems', *Twin Research and Human Genetics* **20**(4), 290–297.

Supplementary Material

Supplement to "Cognitive Skill Production and Unequally Spaced Panel Data": Appendices

By

Maximiliaan W. P. Thijssen

A.1. Further Notes on Unequally Spaced Intervals

The following explication draws heavily from Millimet & McDonough (2017). Consider Equation (A.1.1), which is a repetition of Equation (1) in Section 2,

$$\mathrm{S}_{i,t} = \alpha_1 \mathrm{S}_{i,t-1} + \alpha_2 \mathrm{I}_{i,t-1} + \epsilon_{i,t}, \tag{A.1.1}$$

for child i ($i = 1, \dots, N$) and time t ($t = 0, 1, \dots, T$), where I assume that the error term, $\epsilon_{i,t}$, is independent of the inputs (*i.e.*, $\mathrm{S}_{i,t-1}$ and $\mathrm{I}_{i,t-1}$). To derive the general case, I first consider the case of $T = 2$. With $T = 2$ and applying repeated substitution, I can write,

$$\begin{aligned}
\mathrm{S}_{i,2} &= \alpha_1 \mathrm{S}_{i,1} + \alpha_2 \mathrm{I}_{i,1} + \epsilon_{i,2} \\
&= \alpha_1(\alpha_1 \mathrm{S}_{i,0} + \alpha_2 \mathrm{I}_{i,0} + \epsilon_{i,1}) + \alpha_2 \mathrm{I}_{i,1} + \epsilon_{i,2} \\
&= \alpha_1^2 \mathrm{S}_{i,0} + \alpha_1 \alpha_2 \mathrm{I}_{i,0} + \alpha_1 \epsilon_{i,1} + \alpha_2 \mathrm{I}_{i,1} + \epsilon_{i,2} \\
&= \alpha_1^2 \mathrm{S}_{i,0} + (\alpha_2 \mathrm{I}_{i,1} + \alpha_1 \alpha_2 \mathrm{I}_{i,0}) + (\epsilon_{i,2} + \alpha_1 \epsilon_{i,1}),
\end{aligned}$$

which can be written as,

$$\mathrm{S}_{i,2} = \alpha_1^2 \mathrm{S}_{i,0} + \sum_{r=0}^{1} \alpha_1^r \alpha_2 \mathrm{I}_{i,1-r} + \sum_{r=0}^{1} \alpha_1^r \epsilon_{i,1-r}. \tag{A.1.2}$$

The second power on the parameter α_1 refers to the two steps along the time dimension; from $\mathrm{S}_{i,0}$ to $\mathrm{S}_{i,1}$ and from $\mathrm{S}_{i,1}$ to $\mathrm{S}_{i,2}$. I can write the general case as follows:

$$\mathrm{S}_{i,t} = \alpha_1^{\Delta t} \mathrm{S}_{i,t-\Delta t} + \sum_{r=0}^{\Delta t-1} \alpha_1^r \alpha_2 \mathrm{I}_{i,t-1-r} + \sum_{r=0}^{\Delta t-1} \alpha_1^r \epsilon_{i,t-r}, \tag{A.1.3}$$

where $\Delta t \geq 1$ denotes the time distance.

In Equation (A.1.3), t and Δt represent, respectively, the points in time and time intervals *assumed* by the model in Equation (A.1.1). Let a ($a = 0, 1, \dots, A$) denote the *observed* points in time so that t_a denotes the actual point in time reflected by point a and $\Delta t_a \equiv t_a - t_{a-1}$ denotes the interval observed in the data. I can write Equation (A.1.3) in terms of observed periods as follows:

$$\begin{aligned}
S_{i,a} &= \alpha_1^{\Delta t_a} S_{i,a-1} + \sum_{r=0}^{\Delta t_a - 1} \alpha_1^r \alpha_2 I_{i,t_a-1-r} + \sum_{r=0}^{\Delta t_a - 1} \alpha_1^r \epsilon_{i,t_a-r}, \\
&= \alpha_1^{\Delta t_a} S_{i,a-1} + \alpha_2 I_{i,a-1} + \sum_{r=1}^{\Delta t_a - 1} \alpha_1^r \alpha_2 I_{i,t_a-r} + \sum_{r=0}^{\Delta t_a - 1} \alpha_1^r \epsilon_{i,t_a-r}, \\
&= \alpha_1 S_{i,a-1} + \alpha_2 I_{i,a-1} + \tilde{\epsilon}_{i,a} \qquad \text{(A.1.4)}
\end{aligned}$$

with

$$\tilde{\epsilon}_{i,a} \equiv (\alpha_1^{\Delta t_a} - \alpha_1) S_{i,a-1} + \sum_{r=1}^{\Delta t_a - 1} \alpha_1^r \alpha_2 I_{i,t_a-r} + \sum_{r=0}^{\Delta t_a - 1} \alpha_1^r \epsilon_{i,t_a-r}$$

Equation (A.1.4) reduces to Equation (A.1.1) if (i) observed intervals are equally spaced (*i.e.*, $\Delta t_a = 1$ for all $a = 0, 1, \ldots, A$) and (ii) if the assumed time points correspond to the observed time points. For instance, if we observe children at $t_0 = 0$ and $t_1 = 1$, so that $\Delta t_a = t_1 - t_0 = 1 - 0 = 1$, Equation (A.1.4) reduces to,

$$S_{i,1} = \alpha_1 S_{i,0} + \alpha_2 I_{i,0} + \epsilon_{i,1}. \qquad \text{(A.1.5)}$$

However, when the observed intervals are unequally spaced (or not unit interval), Equation (A.1.4) has several features. To show these features, consider the case in which we observe children at $t_0 = 0$ and $t_1 = 3$, so that $\Delta t_1 = t_1 - t_0 = 3 - 0 = 3$. Under this scenario, Equation (A.1.4) reduces to,

$$\begin{aligned}
S_{i,1} = \alpha_1 S_{i,0} + \alpha_2 I_{i,0} + [(\alpha_1^3 - \alpha_1) S_{i,0} + \alpha_1 \alpha_2 I_{i,2} + \alpha_1^2 \alpha_2 I_{i,1} \\
+ \epsilon_{i,3} + \alpha_1 \epsilon_{i,2} + \alpha_1^2 \epsilon_{i,1}], \qquad \text{(A.1.6)}
\end{aligned}$$

where $I_{i,2}$ and $I_{i,1}$ are the parental investments from missing periods.

Equation (A.1.6) has several features: (i) all parameters depend on the observed time interval, (ii) the effect of parental investments in unobserved periods depends on the self-productivity of cognitive skills, (iii) self-productivity is a weighted average of polynomials of α_1, and (iv) despite the independence assumption, inputs (*i.e.*, $S_{i,a-1}$ and $I_{i,a-1}$) may correlate with the error term because of the inputs from unobserved periods.

A.2. Further Notes on the Identification

I first show the identification of the factor loadings and intercepts in each measurement equation. Next, I show the identification of the joint distribution of factors. Specifically, I present an application of Kotlarski's lemma (see Lemma 1 in Kotlarski 1967) using the first-order partial derivative of the log characteristic function.

A.2.1. Identification of the Factor Loadings and Intercepts

While identifying the factor loadings and intercepts is sufficient for applying Kotlarski's lemma, I also write other "unknowns" as a function of "knowns" for completeness. Consistent with the data (see Section 5), I assume a minimum of three *valid* manifest variables for each common factor. Manifest variables are valid if they have nonzero factor loadings.

A.2.1.1. The Child's Cognitive Skills

I assumed the following linear-in-parameters measurement equation to measure the child's cognitive skill:

$$\mathrm{M}_{1,l,i,t} = \mu_{1,l,t} + \lambda_{1,l,t}\mathrm{S}_{i,t} + \zeta_{1,l,i,t}, \tag{A.2.7}$$

for manifest variable l ($l = 1, \dots, L_{1,t}$), child i ($i = 1, \dots, N$), and time t ($t = 0, 1, \dots, T$). In Equation (A.2.7), $\mathrm{M}_{1,l,i,t}$ denotes the lth manifest variable for child i at time t, $\mu_{1,l,t}$ is the intercept, $\lambda_{1,l,t}$, is a factor loading, $\mathrm{S}_{i,t}$ is the child's (unobserved) cognitive skill, and $\zeta_{1,l,i,t}$ is an error term. To set the location, I normalized $\mathbb{E}(\mathrm{S}_{i,0}) = 0$, where $\mathbb{E}(\cdot)$ denotes the expectation operator, and $\mu_{1,1,t} = \mu_{1,1,t'}$ for all $t, t' = 0, 1, \dots, T$. To set the scale, I normalized $\lambda_{1,1,0} = 1$ and $\lambda_{1,1,t} = \lambda_{1,1,t'}$ for all $t, t' = 0, 1, \dots, T$. Below, I write the unknowns (right-hand side) as a function of knowns (or identified) parameters (left-hand side).

I start with the variance-covariance structure. First, I can identify the factor loadings from the ratio of covariances as follows:

$$\frac{\mathrm{Cov}(\mathrm{M}_{1,l,i,t}, \mathrm{M}_{1,l',i,t})}{\mathrm{Cov}(\mathrm{M}_{1,1,i,t}, \mathrm{M}_{1,l',i,t})} = \frac{\lambda_{1,l,t}\lambda_{1,l',t}\mathrm{Var}(\mathrm{S}_{i,t})}{\lambda_{1,l',t}\mathrm{Var}(\mathrm{S}_{i,t})} = \lambda_{1,l,t},$$

for time $t = 0, 1, \dots, T$; and manifest variables l, l'=2,3, where $l \neq l'$. With the factor loadings identified, I can identify the common factor variance,

$$\frac{\text{Cov}(\text{M}_{1,1,i,t}, \text{M}_{1,l,i,t})}{\lambda_{1,l,t}} = \text{Var}(\text{S}_{i,t}),$$

for time $t = 0, 1, \dots, T$; and manifest variables $l = 2, 3$. With the factor loadings identified, I can also identify the covariance between the common factors (over time) from,

$$\frac{\text{Cov}(\text{M}_{1,1,i,t}, \text{M}_{1,l,i,t'})}{\lambda_{1,l,t'}} = \text{Cov}(\text{S}_{i,t}, \text{S}_{i,t'}),$$

for time $t, t' = 0, 1, \dots, T$, $t \neq t'$; and manifest variables $l = 2, 3$. With the common factor variance and factor loadings identified, I can identify the unique factor variance from,

$$\text{Var}(\text{M}_{1,l,i,t}) - (\lambda_{1,l,t})^2 \text{Var}(\text{S}_{i,t}) = \text{Var}(\zeta_{1,l,i,t}),$$

for $t = 0, 1, \dots, T$; and manifest variables $l = 1, 2, 3$. With the factor loadings and the covariance between the common factors (over time) identified, I can identify the covariance between the unique factors (over time) from,

$$\text{Cov}(\text{M}_{1,l,i,t}, \text{M}_{1,l,i,t'}) - \lambda_{1,l,t}\lambda_{1,l,t'}\text{Cov}(\text{S}_{i,t}, \text{S}_{i,t'}) = \text{Cov}(\zeta_{1,l,i,t}, \zeta_{1,l,i,t'}),$$

for $t, t' = 0, 1, \dots, T$, $t \neq t'$; and manifest variables $l = 1, 2, 3$. With the factor loadings identified, I can identify the intercepts in the initial period from the expectations,

$$\mathbb{E}(\text{M}_{1,l,i,0}) = \mu_{1,l,0} + \lambda_{1,l,0}\mathbb{E}(\text{S}_{i,0}) = \mu_{1,l,0},$$

for manifest variables $l = 1, 2, 3$. Next, since I can identify the intercepts and factor loadings, and because of the time-invariance normalizations, I can identify the mean of the common factor for $t > 0$ from,

$$\frac{\mathbb{E}(\text{M}_{1,1,i,t}) - \mu_{1,1,0}}{\lambda_{1,1,0}} = \mathbb{E}(\text{S}_{i,t})$$

for time $t = 1, \ldots, T$. Lastly, with the factor loadings and common factor mean in each period identified, I can identify the other intercepts for $t > 0$ from,

$$\mathbb{E}(\mathrm{M}_{1,l,i,t}) - \lambda_{1,l,t}\mathbb{E}(\mathrm{S}_{i,t}) = \mu_{1,l,t},$$

for time $t = 1, \ldots, T$; and manifest variables $l = 2, 3$.

A.2.1.2. The Parental Investments

For the parental investments, I assumed the following linear-in-parameters measurement equation:

$$\mathrm{M}_{2,l,i,t} = \mu_{2,l,t} + \lambda_{2,l,t}\mathrm{I}_{i,t} + \zeta_{2,l,i,t}, \tag{A.2.8}$$

for manifest variable l ($l = 1, \ldots, L_{2,t}$), child i ($i = 1, \ldots, N$), and time t ($t = 0, 1, \ldots, T$). In Equation (A.2.8), $\mathrm{M}_{2,l,i,t}$ denotes the lth manifest variable for child i at time t, $\mu_{2,l,t}$ is the intercept, $\lambda_{2,l,t}$ is a factor loading, $\mathrm{I}_{i,t}$ is the parent's (unobserved) investment, and $\zeta_{2,l,i,t}$ is an error term. To set the location, I normalized $\mathbb{E}(\mathrm{I}_{i,t}) = 0$ for all $t = 0, 1, \ldots, T$. Furthermore, to set the scale, I normalized $\lambda_{2,1,t} = 1$ for all $t = 0, 1, \ldots, T$. I follow the same procedure as in Section A.2.1.1 and write the unknowns (right-hand side) as a function of known (or identified) parameters (left-hand side).

I start again with the variance-covariance structure. I can identify the factor loadings for the parental investments from the ratio of covariances as follows:

$$\frac{\mathrm{Cov}(\mathrm{M}_{2,l,i,t}, \mathrm{M}_{2,l',i,t})}{\mathrm{Cov}(\mathrm{M}_{2,1,i,t}, \mathrm{M}_{2,l',i,t})} = \frac{\lambda_{2,l,t}\lambda_{2,l',t}\mathrm{Var}(\mathrm{I}_{i,t})}{\lambda_{2,l',t}\mathrm{Var}(\mathrm{I}_{i,t})} = \lambda_{2,l,t},$$

for time $t = 0, 1, \ldots, T$; and manifest variables $l, l' = 2, 3$, where $l \neq l'$. With the factor loadings identified, I can identify the common factor variance,

$$\frac{\mathrm{Cov}(\mathrm{M}_{2,1,i,t}, \mathrm{M}_{2,l,i,t})}{\lambda_{2,l,t}} = \mathrm{Var}(\mathrm{I}_{i,t}),$$

for time $t = 0, 1, \ldots, T$; and manifest variables $l = 2, 3$. Also, I can identify the covariance between the common factors (over time) from,

$$\frac{\mathrm{Cov}(\mathrm{M}_{2,1,i,t}, \mathrm{M}_{2,l,i,t'})}{\lambda_{2,l,t}} = \mathrm{Cov}(\mathrm{I}_{i,t}, \mathrm{I}_{i,t'}),$$

for time $t, t' = 0, 1, \dots, T$; and manifest variables $l = 2, 3$. With the factor loadings and common factor variance of parental investments identified, I can identify the unique factor variance from,

$$\mathrm{Var}(\mathrm{M}_{2,l,i,t}) - (\lambda_{2,l,t})^2 \mathrm{Var}(\mathrm{I}_{i,t}) = \mathrm{Var}(\zeta_{2,l,i,t}),$$

for time $t, t' = 0, 1, \dots, T$; and manifest variable $l = 1, 2, 3$. With the factor loadings and the covariance between the common factors identified, I can identify the covariance between the unique factors (over time) from,

$$\mathrm{Cov}(\mathrm{M}_{2,l,i,t}, \mathrm{M}_{2,l,i,t'}) - \lambda_{2,l,t}\lambda_{2,l,t'}\mathrm{Cov}(\mathrm{I}_{i,t}, \mathrm{I}_{i,t'}) = \mathrm{Cov}(\zeta_{2,l,i,t}, \zeta_{2,l,i,t'}),$$

for time $t, t' = 0, 1, \dots, T$; and manifest variables $l = 1, 2, 3$. Lastly, with the factor loadings identified, I can identify the intercepts from the expectations,

$$\mathbb{E}(\mathrm{M}_{2,l,i,t}) = \mu_{2,l,t} + \lambda_{2,l,t}\mathbb{E}(\mathrm{I}_{i,t}) = \mu_{2,l,t},$$

for manifest variables $l = 1, 2, 3$.

A.2.1.3. The Mother's Cognitive Skills

For the mother's cognitive skill, I assumed the following linear-in-parameters measurement equation:

$$\mathrm{M}_{3,l,i} = \mu_{3,l} + \lambda_{3,l}\mathrm{S}_{\mathrm{P},i} + \zeta_{3,l,i}, \tag{A.2.9}$$

for manifest variable l ($l = 1, \dots, L_3$) and child i ($i = 1, \dots, N$). In Equation (A.2.9), $\mathrm{M}_{3,l,i}$ denotes the lth manifest variable for the mother of child i, $\mu_{3,l}$ is the measurement intercept, $\lambda_{3,l}$ is a factor loading, $\mathrm{S}_{\mathrm{P},i}$ is the mother's (unobserved) cognitive skill, and $\zeta_{3,l,i}$ is an error term. To set the location, I normalized $\mathbb{E}(\mathrm{S}_{\mathrm{P},i}) = 0$. Furthermore, to set the scale, I normalized $\lambda_{3,1} = 1$. I follow the same procedure as in Section A.2.1.1 and

Section A.2.1.2 and write the unknowns (right-hand side) as a function of known (or identified) parameters (left-hand side).

I start again with the variance-covariance structure. I can identify the factor loadings for the mother's cognitive skill from the ratio of covariances as follows:

$$\frac{\mathrm{Cov}(\mathrm{M}_{3,l,i}, \mathrm{M}_{3,l',i})}{\mathrm{Cov}(\mathrm{M}_{3,1,i}, \mathrm{M}_{3,l',i})} = \frac{\lambda_{3,l}\lambda_{3,l'}\mathrm{Var}(\mathrm{S}_{\mathrm{P},i})}{\lambda_{3,l'}\mathrm{Var}(\mathrm{S}_{\mathrm{P},i})} = \lambda_{3,l},$$

for manifest variables $l, l' = 2, 3$, where $l \neq l'$. With the factor loadings identified, I can identify the common factor variance from,

$$\frac{\mathrm{Cov}(\mathrm{M}_{3,1,i}, \mathrm{M}_{3,l,i})}{\lambda_{3,l}} = \mathrm{Var}(\mathrm{S}_{\mathrm{P},i}).$$

With the factor loadings and common factor variance identified, I can identify the unique factor variance from,

$$\mathrm{Var}(\mathrm{M}_{3,l,i}) - (\lambda_{3,l})^2\mathrm{Var}(\mathrm{S}_{\mathrm{P},i}) = \mathrm{Var}(\zeta_{3,l,i}),$$

for manifest variables $l = 1, 2, 3$. With the factor loadings identified, I can identify the measurement intercepts from the expectations,

$$\mathbb{E}(\mathrm{M}_{3,l,i}) = \mu_{3,l} + \lambda_{3,l}\mathbb{E}(\mathrm{S}_{\mathrm{P},i}) = \mu_{3,l},$$

for manifest variables $l = 1, 2, 3$.

A.2.2. Identifying the Joint Distribution of Factors

This section identifies the joint distribution of the factors by applying Kotlarski's lemma (see Lemma 1, Remark 4, and Remark 5 in Kotlarski 1967, pp. 70–73). See also Li & Vuong (1998, pp. 140–145), Ben-Moshe (2012, pp. 7–8), and Ben-Moshe (2018, pp. 138–139). Cunha et al. (2010) build on the analyses in Schennach (2004*a*,*b*) and Hu & Schennach (2008) and provide a generalization of Kotlarski's lemma. Cunha et al. (2010) also prove that one can nonparametrically identify a nonseparable measurement model.

Consider the measurement model for the child's cognitive skills,

$$\mathrm{M}_{1,l,i,t} = \mu_{1,l,t} + \lambda_{1,l,t}\mathrm{S}_{i,t} + \zeta_{1,l,i,t}, \tag{A.2.10}$$

for manifest variable l ($l = 1, 2, 3$), child i ($i = 1, \dots, N$), and time t ($t = 0, 1, \dots, T$), where the intercepts and factor loadings are identified (see Section A.2.1.1). The idea is to write Equation (A.2.10) as in Cunha & Heckman (2008),

$$\frac{\mathrm{M}_{1,l,i,t} - \mu_{1,l,t}}{\lambda_{1,l,t}} = \mathrm{S}_{i,t} + \frac{\zeta_{1,l,i,t}}{\lambda_{1,l,t}}. \tag{A.2.11}$$

Next, let (i) $\mathbf{M}_{1,l,i}$ denote a T-dimensional vector, the elements of which are $(\mathbf{M}_{1,l,i,0} - \mu_{1,l,0})/\lambda_{1,l,0}$, $(\mathbf{M}_{1,l,i,1} - \mu_{1,l,1})/\lambda_{1,l,1}$, …, $(\mathbf{M}_{1,l,i,T} - \mu_{1,l,T})/\lambda_{1,l,T}$; (ii) $\mathbf{S}_i$ denote a T-dimensional vector, the elements of which are $\mathrm{S}_{i,0}$, $\mathrm{S}_{i,1}$, …, $\mathrm{S}_{i,T}$; and (iii) $\boldsymbol{\zeta}_{1,l,i}$ denote a T-dimensional vector, the elements of which are $\zeta_{1,l,i,0}/\lambda_{1,l,0}, \zeta_{1,l,i,1}/\lambda_{1,l,1}, \dots, \zeta_{1,l,i,T}/\lambda_{1,l,T}$, such that,

$$\mathbf{M}_{1,l} = \mathbf{S} + \boldsymbol{\zeta}_{1,l}, \tag{A.2.12}$$

for manifest variables $l = 1, 2, 3$. I dropped subscript i to avoid notational clutter.

I can nonparametrically identify the distribution of the factors by applying Kotlarski's lemma (Kotlarski 1967) so that the distribution of the observed vectors ($\mathbf{M}_{1,1}$, $\mathbf{M}_{1,2}$, and $\mathbf{M}_{1,3}$) determines, up to a change in sign, the densities of the factors ($\mathbf{S}$, $\boldsymbol{\zeta}_{1,1}$, $\boldsymbol{\zeta}_{1,2}$, and $\boldsymbol{\zeta}_{1,3}$), provided that the characteristic function does not vanish. We can use the first-order partial derivatives of the log characteristic function to show the identification of the distribution. In particular, we can write the log characteristic function of the observed vectors ($\mathbf{M}_{1,1}$, $\mathbf{M}_{1,2}$, and $\mathbf{M}_{1,3}$) as

$$\begin{aligned}\ln \mathbb{E}(e^{\mathbb{i}t_1\mathbf{M}_{1,1}+\mathbb{i}t_2\mathbf{M}_{1,2}+\mathbb{i}t_3\mathbf{M}_{1,3}}) &= \ln \mathbb{E}(e^{\mathbb{i}\mathbf{S}(t_1+t_2+t_3)}) + \ln \mathbb{E}(e^{\mathbb{i}\boldsymbol{\zeta}_{1,1}t_1}) \\ &+ \ln \mathbb{E}(e^{\mathbb{i}\boldsymbol{\zeta}_{1,2}t_2}) + \ln \mathbb{E}(e^{\mathbb{i}\boldsymbol{\zeta}_{1,3}t_3})\end{aligned} \tag{A.2.13}$$

where $\mathbb{i}$ is the imaginary unit, and t is the argument of the characteristic function. To get at Equation (A.2.13), I first substitute Equation (A.2.12). Next, I reorder terms. Lastly, I rewrite (A.2.13) in concordance with the independence assumption. Define,

$$\phi_{\mathbf{S}}(t_1 + t_2 + t_3) \equiv \ln \mathbb{E}(e^{\mathbb{i}\mathbf{S}(t_1+t_2+t_3)}), \tag{A.2.14}$$

$$\phi_{\zeta_{1,1}}(t_1) \equiv \ln \mathbb{E}(e^{\mathbb{i}\boldsymbol{\zeta}_{1,1}t_1}) \tag{A.2.15}$$

$$\phi_{\zeta_{1,2}}(t_2) \equiv \ln \mathbb{E}(e^{\mathbb{i}\boldsymbol{\zeta}_{1,2}t_2}) \tag{A.2.16}$$

$$\phi_{\zeta_{1,3}}(t_3) \equiv \ln \mathbb{E}(e^{\mathbb{i}\boldsymbol{\zeta}_{1,3}t_3}) \tag{A.2.17}$$

The partial derivative of Equation (A.2.13) with respect to t_1 is,

$$\begin{aligned}\frac{\partial \ln \mathbb{E}(e^{\mathbb{i}t_1\mathbf{M}_{1,1}+\mathbb{i}t_2\mathbf{M}_{1,2}+\mathbb{i}t_3\mathbf{M}_{1,3}})}{\partial t_1} &= \frac{\mathbb{i}\mathbb{E}(\mathbf{M}_{1,1}e^{\mathbb{i}t_1\mathbf{M}_{1,1}+\mathbb{i}t_2\mathbf{M}_{1,2}+\mathbb{i}\mathbf{M}_{1,3}t_3})}{\mathbb{E}(e^{\mathbb{i}t_1\mathbf{M}_{1,1}+\mathbb{i}t_2\mathbf{M}_{1,2}+\mathbb{i}t_3\mathbf{M}_{1,3}})} \\ &= \phi'_{\mathbf{S}}(t_1 + t_2 + t_3) + \phi'_{\boldsymbol{\zeta}_{1,1}}(t_1). \end{aligned} \tag{A.2.18}$$

Substitute $(t_1, t_2, t_3) = (0, 0, u)$,

$$\frac{\mathbb{i}\mathbb{E}(\mathbf{M}_{1,1}e^{\mathbb{i}u\mathbf{M}_{1,3}})}{\mathbb{E}(e^{\mathbb{i}u\mathbf{M}_{1,3}})} = \phi'_{\mathbf{S}}(u) + \phi'_{\boldsymbol{\zeta}_{1,1}}(0) = \phi'_{\mathbf{S}}(u), \tag{A.2.19}$$

where the last equality follows from $\phi'_{\boldsymbol{\zeta}_{1,1}}(0) = \mathbb{i}\mathbb{E}(\boldsymbol{\zeta}_{1,1})$ and the assumption that $\mathbb{E}(\boldsymbol{\zeta}_{1,1}) = 0$. Equation (A.2.19) thus identifies the distribution of the common factor. Identifying the distribution of unique factors follows a similar procedure. In conclusion, using Kotlarski's lemma, I can identify the distribution of the common and unique factors from the knowledge of the joint distribution of observed vectors ($\mathbf{M}_{1,1}$, $\mathbf{M}_{1,2}$, and $\mathbf{M}_{1,3}$).

A.3. Further Notes on the Estimation Procedure

This appendix illustrates the estimation procedure for the full model. I illustrate the case in which there are two periods (*i.e.*, $t = 0, 1$). The *total* number of manifest variables at time t is $L_t = L_{1,t} + L_{2,t}$. Parenthetically, as explained in the main text, I estimate the measurement model for the mother's cognitive skill separately and treat it as an observed covariate below. The total number of covariates (including the mother's cognitive skill) is P. For the system of measurement equations, we have the following matrices of unknowns:

$$\underbrace{\boldsymbol{\mu}_t}_{(L_t,1)} = \begin{bmatrix} \mu_{1,1,t} \\ \vdots \\ \mu_{1,L_{1,t},t} \\ \mu_{2,1,t} \\ \vdots \\ \mu_{2,L_{2,t},t} \end{bmatrix},$$

$$\underbrace{\boldsymbol{\beta}_t}_{(L_t,P)} = \begin{bmatrix} \beta_{1,1,t,1} & \cdots & \beta_{1,1,t,P} \\ \vdots & \ddots & \vdots \\ \beta_{1,L_{1,t},t,1} & \cdots & \beta_{1,L_{1,t},t,P} \\ \beta_{2,1,t,1} & \cdots & \beta_{2,1,t,P} \\ \vdots & \ddots & \vdots \\ \beta_{2,L_{2,t},t,1} & \cdots & \beta_{2,L_{2,t},t,P} \end{bmatrix},$$

$$\underbrace{\boldsymbol{\Lambda}_t}_{(L_t,2)} = \begin{bmatrix} \lambda_{1,1,t} & 0 \\ \vdots & \vdots \\ \lambda_{1,L_{1,t},t} & 0 \\ 0 & \lambda_{2,1,t} \\ \vdots & \vdots \\ 0 & \lambda_{2,L_{2,t},t} \end{bmatrix},$$

and the following variance-covariance matrix for the unique factors,

$$\underbrace{\mathbf{Z}_t}_{(L_t,L_t)} = \begin{bmatrix} \mathrm{Var}(\zeta_{1,1,i,t}) & \dots & 0 & 0 & \dots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & \mathrm{Var}(\zeta_{1,L_{1,t},i,t}) & 0 & \dots & 0 \\ 0 & \dots & 0 & \mathrm{Var}(\zeta_{2,1,i,t}) & \dots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & 0 & 0 & \dots & \mathrm{Var}(\zeta_{2,L_{2,t},i,t}) \end{bmatrix}.$$

For the technology, we have the following matrices of unknowns:

$$\mathbf{a}_{\Delta t} = \begin{bmatrix} \alpha_{0,\Delta t} \\ 0 \end{bmatrix},$$

$$\mathbf{B}_{\Delta t} = \begin{bmatrix} \alpha_{1,\Delta t} & \alpha_{2,\Delta t} \\ \beta_{1,\Delta t} & \beta_{2,\Delta t} \end{bmatrix},$$

$$\mathbf{C}_{\Delta t} = \begin{bmatrix} \alpha_{3,\Delta t} \\ \beta_{3,\Delta t} \end{bmatrix},$$

$$\mathbf{d}_{\Delta t} = \begin{bmatrix} 0 \\ \beta_{4,\Delta t} \end{bmatrix},$$

for $\Delta t \equiv t - (t-1)$, and the following variance-covariance for the errors:

$$\mathbf{Q}_t = \begin{bmatrix} \text{Var}(\epsilon_{i,t}) & \text{Cov}(\epsilon_{i,t}, \eta_{i,t}) \\ \text{Cov}(\epsilon_{i,t}, \eta_{i,t}) & \text{Var}(\eta_{i,t}) \end{bmatrix}.$$

We can write the system of measurement equations and the technology at time t in matrix form as follows:

$$\boldsymbol{\theta}_{i,t} = \mathbf{a}_{\Delta t} + \mathbf{B}_{\Delta t}\boldsymbol{\theta}_{i,t-\Delta t} + \mathbf{C}_{\Delta t}\text{S}_{\text{P},i} + \mathbf{d}_{\Delta t}\ln y_{i,t-1} + \mathbf{w}_{i,t}, \quad \mathbf{w}_{i,t} \sim \text{N}(\mathbf{0}, \mathbf{Q}_t), \tag{A.3.20}$$

$$\mathbf{M}_{i,t} = \boldsymbol{\mu}_t + \mathbf{X}_i\boldsymbol{\beta}_t + (\boldsymbol{\Lambda}_t\boldsymbol{\theta}_{i,t} + \boldsymbol{\zeta}_{i,t}), \quad \boldsymbol{\zeta}_{i,t} \sim \text{N}(\mathbf{0}, \mathbf{Z}_t), \tag{A.3.21}$$

where $\boldsymbol{\theta}'_{i,t} = [\text{S}_{i,t}, \text{I}_{i,t}]$ and $\mathbf{w}'_{i,t} = [\epsilon_{i,t}, \eta_{i,t}]$.

Next, define a filter matrix, $\mathscr{F}$. The filter matrix "filters" observed and unobserved (or latent) variables. Let $\mathbf{0}$ denote a (p,q)-dimensional null sub-matrix and let $\mathbf{I}$ denote a (p,p)-dimensional identity matrix. The filter matrix then becomes,

$$\underbrace{\mathscr{F}}_{(p,q+p)} = [\underbrace{\mathbf{0}}_{(p,q)} : \underbrace{\mathbf{I}}_{(p,p)}]. \tag{A.3.22}$$

For example, if is one unobserved variable ($q = 1$) and three observed variables ($p = 3$), then the filter matrix is,

$$\underbrace{\mathscr{F}}_{(3,4)} = [\underbrace{\mathbf{0}}_{(3,1)} : \underbrace{\mathbf{I}}_{(3,3)}] = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

Next, define a $(q+p, q+p)$-dimensional *asymmetric* coefficient matrix,

$\mathscr{A}$, and a $(q+p,q+p)$-dimensional *symmetric* coefficient matrix, $\mathscr{S}$,

$$
\underbrace{\mathscr{A}}_{(q+p,q+p)} = \begin{bmatrix}
\underbrace{\mathbf{0}}_{(2,2)} & \underbrace{\mathbf{0}}_{(2,2)} & \underbrace{\mathbf{0}}_{(2,L_0)} & \underbrace{\mathbf{0}}_{(2,L_1)} & \underbrace{\mathbf{d}_0}_{(2,2)} & \underbrace{\mathbf{0}}_{(2,1)} & \underbrace{\mathbf{0}}_{(2,1)} & \underbrace{\mathbf{0}}_{(2,P)} \\
\underbrace{\mathbf{B}_1}_{(2,2)} & \underbrace{\mathbf{0}}_{(2,2)} & \underbrace{\mathbf{0}}_{(2,L_0)} & \underbrace{\mathbf{0}}_{(2,L_1)} & \underbrace{\mathbf{d}_1}_{(2,1)} & \underbrace{\mathbf{0}}_{(2,1)} & \underbrace{\mathbf{C}_1}_{(2,1)} & \underbrace{\mathbf{0}}_{(2,P)} \\
\underbrace{\mathbf{\Lambda}_0}_{(L_0,2)} & \underbrace{\mathbf{0}}_{(L_0,2)} & \underbrace{\mathbf{0}}_{(L_0,L_0)} & \underbrace{\mathbf{0}}_{(L_0,L_1)} & \underbrace{\mathbf{0}}_{(L_0,1)} & \underbrace{\mathbf{0}}_{(L_0,1)} & \underbrace{\mathbf{0}}_{(L_0,1)} & \underbrace{\boldsymbol{\beta}_0}_{(L_0,P)} \\
\underbrace{\mathbf{0}}_{(L_1,2)} & \underbrace{\mathbf{\Lambda}_1}_{(L_1,2)} & \underbrace{\mathbf{0}}_{(L_1,L_0)} & \underbrace{\mathbf{0}}_{(L_1,L_1)} & \underbrace{\mathbf{0}}_{(L_1,1)} & \underbrace{\mathbf{0}}_{(L_1,1)} & \underbrace{\mathbf{0}}_{(L_1,1)} & \underbrace{\boldsymbol{\beta}_1}_{(L_1,P)} \\
\underbrace{\mathbf{0}}_{(1,2)} & \underbrace{\mathbf{0}}_{(1,2)} & \underbrace{\mathbf{0}}_{(1,L_0)} & \underbrace{\mathbf{0}}_{(1,L_1)} & \underbrace{0}_{(1,1)} & \underbrace{0}_{(1,1)} & \underbrace{0}_{(1,1)} & \underbrace{\mathbf{0}}_{(1,P)} \\
\underbrace{\mathbf{0}}_{(1,2)} & \underbrace{\mathbf{0}}_{(1,2)} & \underbrace{\mathbf{0}}_{(1,L_0)} & \underbrace{\mathbf{0}}_{(1,L_1)} & \underbrace{b_1}_{(1,1)} & \underbrace{0}_{(1,1)} & \underbrace{0}_{(1,1)} & \underbrace{\mathbf{0}}_{(1,P)} \\
\underbrace{\mathbf{0}}_{(1,2)} & \underbrace{\mathbf{0}}_{(1,2)} & \underbrace{\mathbf{0}}_{(1,L_0)} & \underbrace{\mathbf{0}}_{(1,L_1)} & \underbrace{0}_{(1,1)} & \underbrace{0}_{(1,1)} & \underbrace{0}_{(1,1)} & \underbrace{\mathbf{0}}_{(1,P)} \\
\underbrace{\mathbf{0}}_{(P,2)} & \underbrace{\mathbf{0}}_{(P,2)} & \underbrace{\mathbf{0}}_{(P,L_0)} & \underbrace{\mathbf{0}}_{(P,L_1)} & \underbrace{\mathbf{0}}_{(P,1)} & \underbrace{\mathbf{0}}_{(P,1)} & \underbrace{\mathbf{0}}_{(P,1)} & \underbrace{\mathbf{0}}_{(P,P)}
\end{bmatrix},
$$

and

$$
\underbrace{\mathscr{S}}_{(q+p,q+p)} = \begin{bmatrix}
\underbrace{\mathbf{Q}_0}_{(2,2)} & \underbrace{\mathbf{0}}_{(2,2)} & \underbrace{\mathbf{0}}_{(2,L_0)} & \underbrace{\mathbf{0}}_{(2,L_1)} & \underbrace{\mathbf{0}}_{(2,1)} & \underbrace{\mathbf{0}}_{(2,2)} & \underbrace{\mathbf{0}}_{(2,1} & \underbrace{\mathbf{0}}_{(2,P)} \\
\underbrace{\mathbf{0}}_{(2,2)} & \underbrace{\mathbf{Q}_1}_{(2,2)} & \underbrace{\mathbf{0}}_{(2,L_0)} & \underbrace{\mathbf{0}}_{(2,L_1)} & \underbrace{\mathbf{0}}_{(2,1)} & \underbrace{\mathbf{0}}_{(2,1)} & \underbrace{\mathbf{0}}_{(2,1)} & \underbrace{\mathbf{0}}_{(2,P)} \\
\underbrace{\mathbf{0}}_{(L_0,2)} & \underbrace{\mathbf{0}}_{(L_0,2)} & \underbrace{\mathbf{Z}_{0,0}}_{(L_0,L_0)} & \underbrace{\mathbf{Z}_{0,1}}_{(L_0,L_1)} & \underbrace{\mathbf{0}}_{(L_0,1)} & \underbrace{\mathbf{0}}_{(L_0,1)} & \underbrace{\mathbf{0}}_{(L_0,1)} & \underbrace{\mathbf{0}}_{(L_0,P)} \\
\underbrace{\mathbf{0}}_{(L_1,2)} & \underbrace{\mathbf{0}}_{(L_1,2)} & \underbrace{\mathbf{Z}_{1,0}}_{(L_1,L_0)} & \underbrace{\mathbf{Z}_{1,1}}_{(L_1,L_1)} & \underbrace{\mathbf{0}}_{(L_1,1)} & \underbrace{\mathbf{0}}_{(L_1,1)} & \underbrace{\mathbf{0}}_{(L_1,1)} & \underbrace{\mathbf{0}}_{(L_1,P)} \\
\underbrace{\mathbf{0}}_{(1,2)} & \underbrace{\mathbf{0}}_{(1,2)} & \underbrace{\mathbf{0}}_{(1,L_0)} & \underbrace{\mathbf{0}}_{(1,L_1)} & \underbrace{\Sigma_{\mathrm{y}_0}}_{(1,1)} & \underbrace{0}_{(1,1)} & \underbrace{\Sigma_{\mathrm{y}_0\mathrm{S}_\mathrm{P}}}_{(1,1)} & \underbrace{\mathbf{\Sigma}_{\mathrm{y}_0\mathbf{X}}}_{(1,P)} \\
\underbrace{\mathbf{0}}_{(1,2)} & \underbrace{\mathbf{0}}_{(1,2)} & \underbrace{\mathbf{0}}_{(1,L_0)} & \underbrace{\mathbf{0}}_{(1,L_1)} & \underbrace{0}_{(1,1)} & \underbrace{\Sigma_{\mathrm{y}_1}}_{(1,1)} & \underbrace{\Sigma_{\mathrm{y}_1\mathrm{S}_\mathrm{P}}}_{(1,1)} & \underbrace{\mathbf{\Sigma}_{\mathrm{y}_1\mathbf{X}}}_{(1,P)} \\
\underbrace{\mathbf{0}}_{(1,2)} & \underbrace{\mathbf{0}}_{(1,2)} & \underbrace{\mathbf{0}}_{(1,L_0)} & \underbrace{\mathbf{0}}_{(1,L_1)} & \underbrace{\Sigma_{\mathrm{y}_0\mathrm{S}_\mathrm{P}}}_{(1,1)} & \underbrace{\Sigma_{\mathrm{y}_1\mathrm{S}_\mathrm{P}}}_{(1,1)} & \underbrace{\Sigma_{\mathrm{S}_\mathrm{P}}}_{(1,1)} & \underbrace{\Sigma_{\mathbf{X},\mathrm{S}_\mathrm{P}}}_{(1,P)} \\
\underbrace{\mathbf{0}}_{(P,2)} & \underbrace{\mathbf{0}}_{(P,2)} & \underbrace{\mathbf{0}}_{(P,L_0)} & \underbrace{\mathbf{0}}_{(P,L_1)} & \underbrace{\Sigma_{\mathrm{y}_0\mathbf{X}}}_{(P,1)} & \underbrace{\Sigma_{\mathrm{y}_1\mathbf{X}}}_{(P,1)} & \underbrace{\Sigma_{\mathbf{X}\mathrm{S}_\mathrm{P}}}_{(P,1)} & \underbrace{\Sigma_{\mathbf{X}}}_{(P,P)}
\end{bmatrix}.
$$

Together, these matrices represent the Reticular Action Model (McArdle & McDonald 1984). Using these three matrices, I can write the model-implied variance-covariance as,

$$\boldsymbol{\Sigma}_{\mathrm{M}} = \mathscr{F}(\mathbf{I} - \mathscr{A})^{-1}\mathscr{S}(\mathbf{I} - \mathscr{A})^{-1\prime}\mathscr{F}', \tag{A.3.23}$$

and the model-implied mean structure is,

$$\boldsymbol{\mu}_{\mathrm{M}} = \mathscr{F}(\mathbf{I} - \mathscr{A})^{-1}\mathscr{M}, \tag{A.3.24}$$

where $\mathscr{M}$ is a vector with mean components such as $\mathbf{a}_{\Delta t}$ and $\boldsymbol{\mu}_t$.

With the model-implied variance-covariance and mean structure defined, the next step involves deriving the nonlinear constraints that link the discrete-time model parameters exactly to the continuous-time parameters. One can show that Equation (A.3.20) belongs to the following stochastic differential equation:

$$\frac{\mathrm{d}\boldsymbol{\theta}_i(t)}{\mathrm{d}t} = \mathbf{a} + \mathbf{B}\boldsymbol{\theta}_i(t) + \mathbf{C}\mathrm{S}_{\mathrm{P},i} + \mathbf{d}\ \ln\,\mathrm{y}_i(t) + \mathbf{G}\frac{\mathrm{d}\mathbf{w}_i(t)}{\mathrm{d}t}, \tag{A.3.25}$$

where the parameters are no longer dependent on the observation interval. I assume that the continuous-time error term, $\mathbf{w}_i(t)$, follows a Wiener process. A central property of this process is independent and normally distributed increments, $\Delta\mathbf{w}_i(t) \equiv \mathbf{w}_i(t) - \mathbf{w}_i(t - \Delta t)$, with mean zero and covariance matrix $\Delta t\mathbf{I}$ (Arnold 1974, p. 46). The matrix $\mathbf{G}$ allows the continuous-time error process variance to be lower or higher than one. The associated variance-covariance matrix, $\mathbf{GG}'$, is the diffusion matrix associated with the stochastic process. Arnold (1974, 45–56) explains that one cannot interpret the integral of the Wiener process as an ordinary Riemann-Stieltjes integral owing to unbounded variation (see also Hamerle et al. 1991). One can interpret the integral alternatively, which gives the following solution (Arnold 1974, pp. 128–135):

$$\begin{aligned}\boldsymbol{\theta}_i(t) = {} & \mathbf{e}^{\mathbf{B}(t-t_0)}\boldsymbol{\theta}_i(t_0) + \mathbf{B}^{-1}(\mathbf{e}^{\mathbf{B}(t-t_0)} - \mathbf{I})(\mathbf{a} + \mathbf{C}\mathrm{S}_{\mathrm{P},i}) \\ & + \mathbf{d}\sum_{r\in R}\ln\,\mathrm{y}_{i,r}\Delta(t - t_r) + \int_{t_0}^{t}\mathbf{e}^{\mathbf{B}(t-p)}\mathbf{G}\mathrm{d}\mathbf{w}_i(p),\end{aligned} \tag{A.3.26}$$

for initial value $\boldsymbol{\theta}_i(t_0)$ and observation interval $t - t_0$, where $\mathbf{e}^{\{\cdot\}}$ denotes the matrix exponential (see Appendix A.4 for more details on the matrix exponential).

I assume a simple impulse form for (log) family income in Equation (A.3.26) (see, *e.g.*, Driver 2018). As a result, (log) family income only affects parental investments at the time of observation $r \in R$. At the time of observation r, (log) family income causes an upward (or downward) spike by $\mathbf{d} \ln \mathrm{y}_{i,r}$. The Dirac delta function $\Delta(t - t_r)$, which is a generalized function that is ∞ at zero and zero elsewhere, yet has an integral of one when zero is in the range of integration, causes an instantaneous impulse so that the discrete-time and continuous-time forms are equivalent at the time of observation.

Choosing for $t - t_0$ the observation interval $\Delta t_{i,a} \equiv t_{i,a} - t_{i,a-1}$, we observe that Equation (A.3.20) equates to Equation (A.3.26) under the following restrictions,

$$\mathbf{B}_{\Delta t_{i,a}} = \mathbf{e}^{\mathbf{B}\Delta t_{i,a}}, \tag{A.3.27}$$

$$\mathbf{a}_{\Delta t_{i,a}} = \mathbf{B}^{-1}(\mathbf{e}^{\mathbf{B}\Delta t_{i,a}} - \mathbf{I})\mathbf{a}, \tag{A.3.28}$$

$$\mathbf{C}_{\Delta t_{i,a}} = \mathbf{B}^{-1}(\mathbf{e}^{\mathbf{B}\Delta t_{i,a}} - \mathbf{I})\mathbf{C}, \tag{A.3.29}$$

and

$$\mathbf{Q}_t = \mathrm{irow}\{(\mathbf{B} \otimes \mathbf{I} + \mathbf{I} \otimes \mathbf{B})^{-1} (\mathbf{B}_{\Delta t} \otimes \mathbf{B}_{\Delta t} - \mathbf{I} \otimes \mathbf{I})\mathrm{row}\{\mathbf{G}\mathbf{G}'\}\}, \tag{A.3.30}$$

where $\otimes$ denotes the Kronecker product, row$\{\cdot\}$ denotes the "rowvec" operation. This operation puts the elements of a matrix row-wise in a column vector. The irow$\{\cdot\}$ is the inverse of the rowvec operation.

In sum, using the restrictions in Equation (A.3.27) through Equation (A.3.30), I obtain the continuous-time parameter matrices that gave rise to the (child-specific) unequally spaced intervals, $\Delta t_{i,a}$. Once I know the continuous-time parameters, I can solve the model for the unit interval and obtain the discrete-time parameters assumed by the model in Equation (A.3.20).

With the matrices and nonlinear constraints defined, the final part of the estimation procedure is to define the likelihood function. Under the normality assumption, and given the matrix specifications, the parameter estimates are obtained by optimizing the following maximum likelihood function, given the restrictions in Equation (A.3.27) through Equation (A.3.30),

$$\mathscr{L} = \sum_{i=1}^{N} \{ l_i \ln(2\pi) + \ln(|\boldsymbol{\Sigma}_{\mathrm{M},i}|) + (\mathbf{Y}_i - \boldsymbol{\mu}_{\mathrm{M},i})\boldsymbol{\Sigma}_{\mathrm{M},i}^{-1}(\mathbf{Y}_i - \boldsymbol{\mu}_{\mathrm{M},i})' \}, \quad \text{(A.3.31)}$$

with $\mathbf{Y}_i$ representing the vector of all observed variables and l_i being the number of non-missing observed variables in row i.[24] I use the C++ based optimizer for solving nonlinear programs (see Zahery et al. 2017), which is available in the "`OpenMx`" package (Neale et al. 2016).

A.4. Further Notes on the Matrix Exponential

This appendix provides further notes on the matrix exponential, a *highly* nonlinear matrix function. The notation used below does not relate to the notation in the main text. Consider a matrix **A**. The matrix exponential, denoted by $\mathbf{e}^{\{\cdot\}}$, can be defined as an infinite sum,

$$\mathbf{e}^{\mathbf{A}} = \sum_{p=0}^{\infty} \frac{1}{p!}\mathbf{A}^p = \mathbf{I} + \mathbf{A} + \frac{1}{2}\mathbf{A}^2 + \frac{1}{6}\mathbf{A}^3 + \frac{1}{24}\mathbf{A}^4 + \ldots \quad \text{(A.4.32)}$$

If $p = 0$, $\mathbf{A}^0$, then one obtains the identity matrix with dimensions like **A**. In what follows, I assume **A** is diagonalizable.

One way of finding the matrix exponential is through an eigenvalue decomposition (Moler & Van Loan 2003, pp. 23–25),

$$\mathbf{A} = \mathbf{M}\mathbf{V}\mathbf{M}^{-1}, \quad \text{(A.4.33)}$$

where **M** denotes the eigenvalue matrix of **A** and **V** denotes a diagonal eigenvalue matrix of **A**. The matrix exponential is then given by

[24] In Equation (A.3.31), $|\cdot|$ denotes the matrix determinant.

$$\mathbf{e}^{\mathbf{A}} = \mathbf{M}\mathbf{e}^{\mathbf{V}}\mathbf{M}^{-1}, \tag{A.4.34}$$

where $\mathbf{e}^{\mathbf{V}}$ is a diagonal matrix whose entries are the scalar exponential of the eigenvalues of **A**. Programs such as Mathematica (`MatrixExp[·]`), Matlab (`expm(·)`), Stata (`mataexpsym(·)`), and the package OpenMx (Neale et al. 2016) (`expm(·)`, implemented in R), can (directly) compute the matrix exponential. In programs that do not have a matrix exponential function, one can follow the steps below.

Consider the matrix

$$\mathbf{A} = \begin{bmatrix} a_{1,1} & a_{1,2} \\ a_{2,1} & a_{2,2} \end{bmatrix}.$$

The first step involves computing the eigenvalue of **A**

$$\frac{1}{2}(a_{1,1} - \sqrt{4a_{1,2}a_{2,1} + (a_{1,1} - a_{2,2})^2} + a_{2,2}), \text{ and}$$

$$\frac{1}{2}(a_{1,1} + \sqrt{4a_{1,2}a_{2,1} + (a_{1,1} - a_{2,2})^2} + a_{2,2}).$$

The scalar exponentials of the eigenvalues are, respectively,

$$e^{\frac{1}{2}(a_{1,1} - \sqrt{4a_{1,2}a_{2,1} + (a_{1,1} - a_{2,2})^2} a_{2,2})}, \text{ and}$$

$$e^{\frac{1}{2}(a_{1,1} + \sqrt{4a_{1,2}a_{2,1} + (a_{1,1} - a_{2,2})^2} + a_{2,2})}.$$

Define a new matrix, $\mathbf{e}^{\mathbf{V}}$; the diagonal elements are the scalar exponentials of the eigenvalues and the off-diagonal elements are zero. The second step involves computing the eigenvectors of matrix A. These eigenvectors are as follows:

$$\mathbf{M} = \begin{bmatrix} \frac{-a_{1,1} + \sqrt{4a_{1,2}a_{2,1} + (a_{1,1} - a_{2,2})^2} + a_{2,2}}{2a_{2,1}} & 1 \\ \frac{a_{1,1} + \sqrt{4a_{1,2}a_{2,1} + (a_{1,1} - a_{2,2})^2} + a_{2,2}}{2a_{2,1}} & 1 \end{bmatrix}$$

with matrix inverse,

$$\mathbf{M}^{-1} = \begin{bmatrix} -\frac{a_{2,1}}{\sqrt{4a_{1,2}a_{2,1}+(a_{1,1}-a_{2,2})^2}} & \frac{a_{2,1}}{\sqrt{4a_{1,2}a_{2,1}+(a_{1,1}-a_{2,2})^2}} \\ \frac{a_{1,1}+\sqrt{4a_{1,2}a_{2,1}+(a_{1,1}-a_{2,2})^2}-a_{2,2}}{2\sqrt{4a_{1,2}a_{2,1}+(a_{1,1}-a_{2,2})^2}} & \frac{-a_{1,1}+\sqrt{4a_{1,2}a_{2,1}+(a_{1,1}-a_{2,2})^2}-a_{2,2}}{2\sqrt{4a_{1,2}a_{2,1}+(a_{1,1}-a_{2,2})^2}} \end{bmatrix}$$

The last step involves pre-multiplying $\mathbf{e}^{\mathbf{V}}$ with $\mathbf{M}$ and post-multiplying $\mathbf{e}^{\mathbf{V}}$ with the inverse $\mathbf{M}^{-1}$,

$$\mathbf{e}^{\mathbf{A}} = \mathbf{M}\mathbf{e}^{\mathbf{V}}\mathbf{M}^{-1}$$

Let $\mathbf{e}^{\mathbf{A}(i,j)}$ denote the ith row and the jth column in matrix $\mathbf{A}$. The elements are then

$$\mathbf{e}^{\mathbf{A}(1,1)} = \frac{e^{\frac{1}{2(a_{1,1}+a_{2,2}-K)}}((a_{1,1} - a_{2,2})(e^K - 1) + (e^K + 1)K)}{2K}$$

$$\mathbf{e}^{\mathbf{A}(1,2)} = \frac{a_{1,2}(e^K - 1)e^{\frac{1}{2(a_{1,1}+a_{2,2}-K)}}}{2K}$$

$$\mathbf{e}^{\mathbf{A}(2,1)} = \frac{a_{2,1}(e^K - 1)e^{\frac{1}{2(a_{1,1}+a_{2,2}-K)}}}{2K}$$

$$\mathbf{e}^{\mathbf{A}(2,2)} = \frac{e^{\frac{1}{2(a_{1,1}+a_{2,2}-K)}}((-a_{1,1} - a_{2,2})(e^K - 1) + (e^K + 1)K)}{2K}$$

where $K = \sqrt{a_{1,1}^2 + 2a_{2,2}a_{1,1} + a_{2,2}^2 + 4a_{1,2}a_{2,1}}$.

A.5. Further Notes on Missing Data

Table A.5.1, Table A.5.2, Table A.5.3, and Table A.5.4 present an overview of missingness. Table A.5.1 reports the observed and missing values for the PIAT subtests. Table A.5.2 reports the observed and missing values for the HOME-SF items. Table A.5.3 reports the observed and missing values for the ASVAB items. Table A.5.4 reports the observed and missing values for (net) family income in the past calendar year.

Typically, researchers apply listwise deletion to address missing values. That is, the children who have missing values for one or more variables are removed. This approach will lead to biased point estimates unless the data are missing completely at random. Instead, I follow Cunha & Heckman (2008) and Cunha et al. (2010). I assume the data are missing at random and apply full information (parametric) maximum likelihood estimation (Anderson 1957). Full information maximum likelihood estimation uses all available data and provides valid point estimates even when the data are only missing at random. Missing at random implies that values are not missing randomly, but that missingness depends only on observed variables. See Little & Rubin (2002) for a thorough treatment of data that are missing at random.

Table A.5.1. Overview of Missing Values PIAT Subtests

		Observed		Missing	
	(1) Period	(2) Obs.	(3) Pct.	(4) Obs.	(5) Pct.
PIAT Reading Comprehension	T5	1,240	65.0	669	35.0
PIAT Mathematics	T5	1,247	65.3	662	34.7
PIAT Reading Recognition	T5	1,250	65.6	659	34.5
PIAT Reading Comprehension	T4	1,570	82.2	339	17.8
PIAT Reading Comprehension	T1	1,581	82.8	328	17.2
PIAT Mathematics	T4	1,583	82.9	326	17.1
PIAT Reading Recognition	T4	1,584	83.0	325	17.0
PIAT Reading Comprehension	T3	1,614	84.5	295	15.5
PIAT Reading Comprehension	T2	1,628	85.3	281	14.7
PIAT Reading Recognition	T3	1,629	85.3	281	14.7
PIAT Mathematics	T3	1,633	85.5	276	14.5
PIAT Reading Recognition	T1	1,678	87.9	231	12.1
PIAT Reading Recognition	T2	1,682	88.1	227	11.9
PIAT Mathematics	T2	1,683	88.2	226	11.8
PIAT Mathematics	T1	1,713	89.7	196	10.3

Notes. This table reports the descriptive frequencies of observed and missing observations for the PIAT subtests (Mathematics, Reading Recognition, Reading Comprehension) from the NLSY79 Children and Young Adult dataset.

Table A.5.2. Overview of Missing Values HOME-SF Items

		Observed		Missing	
	(1) Period	(2) Obs.	(3) Pct.	(4) Obs.	(5) Pct.
Mother reads to the child	T3	797	41.7	1,112	58.3
Lessons/extracurricular activities	T1	846	44.3	1,063	55.7
Daily newspaper	T1	847	44.4	1,062	55.6
Musical instrument	T1	847	44.4	1,062	55.6
Music/theater performances	T1	849	44.5	1,060	55.5
Music/theater performances	T5	1,091	57.2	818	42.8
Musical instrument	T5	1,331	69.7	578	30.3
Daily newspaper	T5	1,332	69.8	577	30.2
Lessons/extracurricular activities	T5	1,333	69.8	576	30.2
Music/theater performances	T4	1,449	75.9	460	24.1
Music/theater performances	T3	1,606	84.1	303	15.9
Daily newspaper	T4	1,669	87.4	240	12.6
Musical instrument	T4	1,671	87.5	238	12.5
Lessons/extracurricular activities	T4	1,672	87.6	237	12.4
Lessons/extracurricular activities	T3	1,682	88.1	227	11.9
Daily newspaper	T3	1,685	88.3	224	11.7
Musical instrument	T3	1,685	88.3	224	11.7
Musical instrument	T2	1,754	91.9	155	8.1
Daily newspaper	T2	1,755	91.9	154	8.1
Lessons/extracurricular activities	T2	1,756	92.0	153	8.0
Music/theater performances	T2	1,756	92.0	153	8.0
Mother reads to the child	T2	1,761	92.2	148	7.8
Mother reads to the child	T1	1,771	92.8	138	7.2

Notes. This table reports the descriptive frequencies of observed and missing observations for the HOME-SF items from the NLSY79 Children and Young Adult dataset. The HOME-SF is a subset of the HOME scale designed by Bradley & Caldwell (1980, 1984) to assess children's emotional support and cognitive stimulation in the home environment, planned events, and family surroundings. The items presented in the rows are shortened: (i) How often mother reads to the child, (ii) Does the child get special lessons or do extra-curricular activities, (iii) Does the family get a daily newspaper, (iv) How often has the child gone to music/theater performances in the past year, and (v) Whether there is a musical instrument the child can use.

Table A.5.3. Overview of Missing Values ASVAB Items

	Observed		Missing	
	(1) Obs.	(2) Pct.	(3) Obs.	(4) Pct.
Arithmetic Reasoning	1,852	97.0	57	3.0
Word Knowledge	1,852	97.0	57	3.0
Paragraph Comprehension	1,852	97.0	57	3.0
Numerical Operations	1,852	97.0	57	3.0
Coding Speed	1,852	97.0	57	3.0
Mathematics Knowledge	1,852	97.0	57	3.0

Notes. This table reports the descriptive frequencies of observed and missing observations for the ASVAB tests (arithmetic reasoning, word knowledge, paragraph comprehension, numerical operations, coding speed, and mathematics knowledge) from the NLSY79 Children and Young Adult dataset (U.S. Department of Defense 1982).

Table A.5.4. Overview of Missing Values Family Income

		Observed		Missing	
	(1) Period	(2) Obs.	(3) Pct.	(4) Obs.	(5) Pct.
Net Family Income past year	T4	1,586	83.1	323	16.9
Net Family Income past year	T2	1,588	83.2	321	16.8
Net Family Income past year	T5	1,595	83.6	314	16.4
Net Family Income past year	T2	1,610	84.3	299	15.7
Net Family Income past year	T1	1,612	84.4	297	15.6

Notes. This table reports the descriptive frequencies of observed and missing observations for the total net family income from the NLSY79 Children and Young Adult dataset.

A.6. Further Notes on the Mother's Cognitive Skills

I separately estimate the measurement model for the mother's cognitive skill (Equation 10). See Appendix A.2 for the identification. I report Pearson correlation coefficients in Table A.6.5 to quantify the degree to which test scores move together. Before estimating the measurement model, I standardize the variables so that the mean is zero and the standard deviation is one. I anchor the common factor in "word knowledge." I report the estimates of the measurement model in Table A.6.6. Using these estimates, I then calculate factor scores for the mother's cognitive skill. I use the method of "regression scoring," which produces consistent estimates when the subsequent variable is used as a regressor (Skrondal & Laake 2001).

Table A.6.5. Pearson Correlation Matrix ASVAB Items

	(1)	(2)	(3)	(4)	(5)	(6)
(1) Arithmetic Reasoning	1	0.71	0.69	0.61	0.55	0.81
(2) Word Knowledge	0.71	1	0.80	0.62	0.58	0.70
(3) Paragraph Composition	0.69	0.80	1	0.63	0.59	0.66
(4) Numerical Operations	0.61	0.62	0.63	1	0.70	0.60
(5) Coding Speed	0.55	0.58	0.59	0.70	1	0.51
(6) Mathematical Knowledge	0.81	0.70	0.66	0.60	0.51	1

Notes. This table reports the Pearson correlations for the manifest variables in the measurement model for the mother's cognitive skill. See U.S. Department of Defense (1982) for the test battery.

Table A.6.6. Measurement Equation for the Mother's Skills

	Factor Loadings		Error Variance	
	(1) Est.	(2) SE	(3) Est.	(4) SE
Arithmetic Reasoning	0.98	0.02	0.28	0.01
Word Knowledge	1	–	0.25	0.01
Paragraph Comprehension	0.98	0.01	0.28	0.01
Numerical Operations	0.86	0.02	0.44	0.02
Coding Speed	0.79	0.02	0.53	0.02
Mathematics Knowledge	0.96	0.02	0.31	0.01

Notes. This table reports the estimates for the measurement model for the mother's cognitive skills. I present Huber-White robust standard errors in parentheses. See U.S. Department of Defense (1982) for details on the test battery.

A.7. Supplementary Tables

I conduct several robustness analyses in this appendix. First, I estimate the cognitive skill production function and investment equation without measurement error correction (see Table A.7.9 through Table A.7.12). To measure cognitive skills, I first standardize the PIAT mathematics, reading recognition, and reading comprehension subtests to be mean zero and standard deviation one. Next, I construct an arithmetic average. Lastly, I re-standardize the arithmetic average to be mean zero and standard deviation one. For parental investments, I use the same procedure.

Second, Figure 1 and Figure 2 show that the time distance between assessment waves is particularly large (or small) for some children (i.e., $\Delta t_{i,a} > 2.5$ and $\Delta t_{i,a} < 1.5$). To investigate the extent to which these children drive my results, I recode these intervals to be minimum 1.5 and maximum 2.5 (see Table A.7.13 and Table A.7.14).

Third, I dropped children for whom I did not observe their age at the time of assessment in each wave. Consequently, I include the same children in each assessment wave. Alternatively, I can construct stage-specific samples and only drop children for whom I do not observe their age at the time of assessment in that stage (see Table A.7.15 and Table A.7.16). I do not report the results of the investment function as the results are comparable and do not alter the conclusion.

Finally, I examine if changing the "anchors" changes the results (see Table A.7.17 through Table A.7.20). In Section 3, I picked PIAT reading recognition as an anchor for the child's cognitive skills and whether the child gets special lessons or does extracurricular activities as an anchor for parental investments. I investigate whether changing these anchors, changes the conclusions. I focus on self-productivity of cognitive skills, the effect of parental investments on the child's cognitive skills, and total factor productivity, as these appeared to be most affected. I do not report the results of the investment function as the results are comparable and do not alter the conclusion.

Table A.7.7. Parental Investment Function Parameters with Time-Invariant Unobserved Heterogeneity

	Stage 1		Stage 2	
	(1) Equal Intervals	(2) Unequal Intervals	(3) Equal Intervals	(4) Unequal Intervals
Cognitive Skills	0.140	0.141	−0.018	−0.039
	(0.072)	(0.034)	(0.026)	(0.020)
Parental Investments	0.655	0.791	0.385	0.420
	(0.116)	(0.043)	(0.072)	(0.047)
Family Income (log)	0.092	0.112	0.184	0.213
	(0.069)	(0.055)	(0.072)	(0.085)
Time-Invariant Shock	0.026	0.012	0.166	0.176
	(0.032)	(0.007)	(0.050)	(0.033)
Random Shock	0.057	0.034	0.074	0.045
	(0.040)	(0.021)	(0.032)	(0.025)

Notes. This table reports the parameter estimates and standard errors in parentheses. The standard errors are computed using the Delta method. The estimated technology equations are:

$$\mathrm{I}_{i,t} = \delta_{1,k}\mathrm{S}_{i,t-1} + \delta_{2,k}\mathrm{I}_{i,t-1} + \delta_{3,k}\ln \mathrm{y}_{i,t-1} + \kappa_{\mathrm{I},i} + \xi_{\mathrm{I},i,t}.$$

The model is estimated using full-information maximum likelihood (1,909 observations). Columns (1) and (2) relate to developmental stage 1 (starts at age 5-6 and ends at age 9-10). Columns (3) and (4) relate to developmental stage 2 (starts at age 9-10 and ends at age 13-14). Columns (1) and (3) are based on equal intervals. Columns (2) and (4) are based on unequal intervals.

Table A.7.8. Parental Investment Function Parameters: Period-by-Period

	Stage 1		Stage 2		Stage 3		Stage 4	
	(1) Equal	(2) Unequal	(3) Equal	(4) Unequal	(5) Equal	(6) Unequal	(7) Equal	(8) Unequal
Cognitive Skills	-0.038	-0.030	0.093	0.050	-0.018	-0.013	-0.001	-0.001
	(0.051)	(0.028)	(0.040)	(0.019)	(0.035)	(0.020)	(0.045)	(0.024)
Parental Investments	0.698	0.863	0.759	0.876	0.644	0.799	0.731	0.839
	(0.091)	(0.055)	(0.080)	(0.046)	(0.073)	(0.046)	(0.093)	(0.054)
Mother's Cognitive Skills	0.029	0.015	0.023	0.017	0.024	0.012	0.108	0.054
	(0.026)	(0.013)	(0.029)	(0.015)	(0.035)	(0.019)	(0.034)	(0.018)
Family Income (log)	-0.004	-0.005	0.200	0.200	0.187	0.184	0.143	0.138
	(0.062)	(0.064)	(0.077)	(0.078)	(0.103)	(0.103)	(0.081)	(0.081)
Random Shock	0.041	0.025	0.076	0.043	0.138	0.084	0.067	0.038
	(0.030)	(0.018)	(0.039)	(0.023)	(0.057)	(0.035)	(0.041)	(0.024)

Notes. This table reports the parameter estimates and standard errors in parentheses. The standard errors are computed using the Delta method. The estimated technology equations are:

$$\mathrm{I}_{i,t} = \beta_{1,k}\mathrm{S}_{i,t-1} + \beta_{2,k}\mathrm{I}_{i,t-1} + \beta_{3,k}\mathrm{S}_{\mathrm{P},i} + \beta_{4,k}\ln \mathrm{y}_{i,t-1} + \eta_{i,t}.$$

The model is estimated using full-information maximum likelihood (1,909 observations). Columns (1) and (2) relate to developmental stage 1 (starts at age 5-6 and ends at age 7-8). Columns (3) and (4) relate to developmental stage 2 (starts at age 7-8 and ends at age 9-10). Columns (5) and (6) relate to developmental stage 3 (starts at age 9-10 and ends at age 11-12). Columns (7) and (8) relate to developmental stage 4 (starts at age 11-12 and ends at age 13-14). Columns (1), (3), (5), and (7) are based on equal intervals. Columns (2), (4), (6), and (8) are based on unequal intervals.

Table A.7.9. Cognitive Skill Production Function Parameters (No Measurement Error Correction): Period-by-Period

	Stage 1		Stage 2		Stage 3		Stage 4	
	(1) Equal	(2) Unequal	(3) Equal	(4) Unequal	(5) Equal	(6) Unequal	(7) Equal	(8) Unequal
Cognitive Skills	0.814	0.907	0.850	0.924	0.839	0.915	0.748	0.864
	(0.034)	(0.019)	(0.019)	(0.010)	(0.018)	(0.010)	(0.019)	(0.011)
Parental Investments	0.048	0.036	0.012	0.009	0.038	0.025	0.020	0.013
	(0.015)	(0.009)	(0.010)	(0.006)	(0.011)	(0.008)	(0.013)	(0.009)
Mother's Cognitive Skills	0.080	0.045	0.084	0.043	0.052	0.027	0.068	0.036
	(0.012)	(0.006)	(0.011)	(0.005)	(0.012)	(0.006)	(0.013)	(0.007)
Total Factor Productivity	0.584	0.296	0.512	0.270	0.478	0.252	0.481	0.260
	(0.047)	(0.021)	(0.017)	(0.008)	(0.016)	(0.009)	(0.021)	(0.012)
Random Shock	0.175	0.092	0.144	0.076	0.165	0.090	0.175	0.101
	(0.013)	(0.008)	(0.010)	(0.006)	(0.012)	(0.007)	(0.014)	(0.009)

Notes. This table reports the parameter estimates and standard errors in parentheses. The standard errors are computed using the Delta method. The estimated technology equations are:

$$\mathrm{S}_{i,t} = \alpha_{0,k} + \alpha_{1,k}\mathrm{S}_{i,t-1} + \alpha_{2,k}\mathrm{I}_{i,t-1} + \alpha_{3,k}\mathrm{S}_{\mathrm{P},i} + \epsilon_{i,t}.$$

The model is estimated using full-information maximum likelihood (1,909 observations). Columns (1) and (2) relate to developmental stage 1 (starts at age 5-6 and ends at age 7-8). Columns (3) and (4) relate to developmental stage 2 (starts at age 7-8 and ends at age 9-10). Columns (5) and (6) relate to developmental stage 3 (starts at age 9-10 and ends at age 11-12). Columns (7) and (8) relate to developmental stage 4 (starts at age 11-12 and ends at age 13-14). Columns (1), (3), (5), and (7) are based on equal intervals. Columns (2), (4), (6), and (8) are based on unequal intervals.

Table A.7.10. Parental Investment Function Parameters (No Measurement Error Correction): Period-by-Period

	Stage 1		Stage 2		Stage 3		Stage 4	
	(1) Equal	(2) Unequal	(3) Equal	(4) Unequal	(5) Equal	(6) Unequal	(7) Equal	(8) Unequal
Cognitive Skills	0.186	0.122	0.258	0.171	0.028	0.018	0.015	0.011
	(0.057)	(0.038)	(0.042)	(0.028)	(0.037)	(0.024)	(0.013)	(0.025)
Parental Investments	0.370	0.611	0.362	0.598	0.375	0.606	0.354	0.586
	(0.030)	(0.025)	(0.022)	(0.018)	(0.023)	(0.019)	(0.027)	(0.023)
Mother's Cognitive Skills	0.127	0.072	0.057	0.032	0.084	0.050	0.122	0.079
	(0.025)	(0.015)	(0.023)	(0.015)	(0.024)	(0.015)	(0.027)	(0.017)
Family Income (log)	0.097	0.101	0.066	0.063	0.054	0.060	−0.057	−0.059
	(0.054)	(0.054)	(0.031)	(0.031)	(0.034)	(0.034)	(0.044)	(0.044)
Random Shock	0.831	0.606	0.742	0.546	0.723	0.532	0.766	0.575
	(0.063)	(0.060)	(0.051)	(0.041)	(0.049)	(0.041)	(0.059)	(0.050)

Notes. This table reports the parameter estimates and standard errors in parentheses. The standard errors are computed using the Delta method. The estimated technology equations are:

$$I_{i,t} = \beta_{1,k} S_{i,t-1} + \beta_{2,k} I_{i,t-1} + \beta_{3,k} S_{P,i} + \beta_{4,k} \ln y_{i,t-1} + \eta_{i,t}.$$

The model is estimated using full-information maximum likelihood (1,909 observations). Columns (1) and (2) relate to developmental stage 1 (starts at age 5-6 and ends at age 7-8). Columns (3) and (4) relate to developmental stage 2 (starts at age 7-8 and ends at age 9-10). Columns (5) and (6) relate to developmental stage 3 (starts at age 9-10 and ends at age 11-12). Columns (7) and (8) relate to developmental stage 4 (starts at age 11-12 and ends at age 13-14). Columns (1), (3), (5), and (7) are based on equal intervals. Columns (2), (4), (6), and (8) are based on unequal intervals.

Table A.7.11. Technology Equations (No Measurement Error Correction)

	Stage 1		Stage 2	
	(1) Equal Intervals	(2) Unequal Intervals	(3) Equal Intervals	(4) Unequal Intervals
Cognitive Skills	0.790	0.899	0.797	0.891
	(0.012)	(0.007)	(0.012)	(0.007)
Parental Investments	0.025	0.020	0.031	0.020
	(0.008)	(0.005)	(0.009)	(0.006)
Mother's Cognitive Skills	0.086	0.043	0.054	0.028
	(0.008)	(0.004)	(0.008)	(0.005)
Total Factor Productivity	0.516	0.276	0.469	0.250
	(0.015)	(0.007)	(0.012)	(0.007)
Random Shock	0.162	0.086	0.171	0.096
	(0.008)	(0.004)	(0.009)	(0.005)

Notes. This table reports the parameter estimates and standard errors in parentheses. The standard errors are computed using the Delta method. The estimated technology equations are:

$$\mathrm{S}_{i,t} = \alpha_{0,k} + \alpha_{1,k}\mathrm{S}_{i,t-1} + \alpha_{2,k}\mathrm{I}_{i,t-1} + \alpha_{3,k}\mathrm{S}_{\mathrm{P},i} + \epsilon_{i,t}.$$

The model is estimated using full-information maximum likelihood (1,909 observations). Columns (1) and (2) relate to developmental stage 1 (starts at age 5-6 and ends at age 9-10). Columns (3) and (4) relate to developmental stage 2 (starts at age 9-10 and ends at age 13-14). Columns (1) and (3) are based on equal intervals. Columns (2) and (4) are based on unequal intervals.

Table A.7.12. Investment Equation (No Measurement Error Correction)

	Stage 1		Stage 2	
	(1) Equal Intervals	(2) Unequal Intervals	(3) Equal Intervals	(4) Unequal Intervals
Cognitive Skills	0.148	0.100	0.010	0.007
	(0.026)	(0.017)	(0.024)	(0.016)
Parental Investments	0.370	0.607	0.366	0.598
	(0.017)	(0.014)	(0.017)	(0.014)
Mother's Cognitive Skills	0.099	0.059	0.103	0.066
	(0.017)	(0.011)	(0.017)	(0.011)
Family Income (log)	0.062	0.059	0.008	0.007
	(0.025)	(0.025)	(0.027)	(0.027)
Random Shock	0.788	0.576	0.744	0.552
	(0.040)	(0.034)	(0.038)	(0.032)

Notes. This table reports the parameter estimates and standard errors in parentheses. The standard errors are computed using the Delta method. The estimated technology equations are:

$$\mathrm{I}_{i,t} = \beta_{1,k}\mathrm{S}_{i,t-1} + \beta_{2,k}\mathrm{I}_{i,t-1} + \beta_{3,k}\mathrm{S}_{\mathrm{P},i} + \beta_{4,k}\ln \mathrm{y}_{i,t-1} + \eta_{i,t}.$$

The model is estimated using full-information maximum likelihood (1,909 observations). Columns (1) and (2) relate to developmental stage 1 (starts at age 5-6 and ends at age 9-10). Columns (3) and (4) relate to developmental stage 2 (starts at age 9-10 and ends at age 13-14). Columns (1) and (3) are based on equal intervals. Columns (2) and (4) are based on unequal intervals.

Table A.7.13. Cognitive Skill Production Function Parameters (No Outliers): Period-by-Period

	Stage 1		Stage 2		Stage 3		Stage 4	
	(1) Equal	(2) Unequal	(3) Equal	(4) Unequal	(5) Equal	(6) Unequal	(7) Equal	(8) Unequal
Cognitive Skills	0.888	0.954	0.829	0.914	0.952	0.969	0.994	1.001
	(0.036)	(0.019)	(0.019)	(0.010)	(0.027)	(0.015)	(0.033)	(0.017)
Parental Investments	0.178	0.102	0.025	0.017	0.038	0.027	0.011	0.004
	(0.039)	(0.022)	(0.029)	(0.016)	(0.025)	(0.016)	(0.030)	(0.016)
Mother's Cognitive Skills	0.049	0.020	−0.004	0.001	−0.009	−0.002	0.022	0.009
	(0.019)	(0.009)	(0.017)	(0.009)	(0.024)	(0.010)	(0.019)	(0.009)
Total Factor Productivity	0.871	0.342	0.573	0.317	0.417	0.212	0.248	0.127
	(0.050)	(0.021)	(0.044)	(0.022)	(0.048)	(0.025)	(0.056)	(0.027)
Random Shock	0.116	0.058	0.073	0.038	0.039	0.029	0.022	0.011
	(0.012)	(0.006)	(0.010)	(0.005)	(0.009)	(0.007)	(0.010)	(0.005)

Notes. This table reports the parameter estimates and standard errors in parentheses. The standard errors are computed using the Delta method. The estimated technology equations are:

$$\mathrm{S}_{i,t} = \alpha_{0,k} + \alpha_{1,k}\mathrm{S}_{i,t-1} + \alpha_{2,k}\mathrm{I}_{i,t-1} + \alpha_{3,k}\mathrm{S}_{\mathrm{P},i} + \epsilon_{i,t}.$$

The model is estimated using full-information maximum likelihood (1,909 observations). Columns (1) and (2) relate to developmental stage 1 (starts at age 5-6 and ends at age 7-8). Columns (3) and (4) relate to developmental stage 2 (starts at age 7-8 and ends at age 9-10). Columns (5) and (6) relate to developmental stage 3 (starts at age 9-10 and ends at age 11-12). Columns (7) and (8) relate to developmental stage 4 (starts at age 11-12 and ends at age 13-14). Columns (1), (3), (5), and (7) are based on equal intervals. Columns (2), (4), (6), and (8) are based on unequal intervals.

Table A.7.14. Investment Function Parameters (No Outliers): Period-by-Period

	Stage 1		Stage 2		Stage 3		Stage 4	
	(1) Equal	(2) Unequal	(3) Equal	(4) Unequal	(5) Equal	(6) Unequal	(7) Equal	(8) Unequal
Cognitive Skills	−0.038	−0.029	0.090	0.052	0.011	−0.011	−0.001	−0.001
	(0.051)	(0.029)	(0.035)	(0.019)	(0.045)	(0.020)	(0.045)	(0.024)
Parental Investments	0.698	0.864	0.768	0.877	0.635	0.800	0.731	0.839
	(0.097)	(0.056)	(0.083)	(0.046)	(0.073)	(0.046)	(0.094)	(0.054)
Mother's Cognitive Skills	0.029	0.015	0.024	0.017	0.024	0.011	0.108	0.054
	(0.025)	(0.013)	(0.029)	(0.016)	(0.034)	(0.019)	(0.034)	(0.018)
Family Income (log)	−0.004	−0.004	0.200	0.201	0.179	0.185	0.143	0.138
	(0.062)	(0.063)	(0.086)	(0.078)	(0.102)	(0.104)	(0.082)	(0.081)
Random Shock	0.041	0.025	0.076	0.043	0.134	0.086	0.067	0.038
	(0.030)	(0.018)	(0.042)	(0.023)	(0.054)	(0.035)	(0.041)	(0.024)

Notes. This table reports the parameter estimates and standard errors in parentheses. The standard errors are computed using the Delta method. The estimated technology equations are:

$$\mathrm{I}_{i,t} = \beta_{1,k}\mathrm{S}_{i,t-1} + \beta_{2,k}\mathrm{I}_{i,t-1} + \beta_{3,k}\mathrm{S}_{\mathrm{P},i} + \beta_{4,k}\ln \mathrm{y}_{i,t-1} + \eta_{i,t}.$$

The model is estimated using full-information maximum likelihood (1,909 observations). Columns (1) and (2) relate to developmental stage 1 (starts at age 5-6 and ends at age 9-10). Columns (3) and (4) relate to developmental stage 2 (starts at age 9-10 and ends at age 13-14). Columns (1) and (3) are based on equal intervals. Columns (2) and (4) are based on unequal intervals.

Table A.7.15. Cognitive Skill Production Function Parameters (Missingness per Stage): Period-by-Period

	Stage 1 (2,694 Obs.)		Stage 2 (2,981 Obs.)		Stage 3 (3,038 Obs.)		Stage 4 (2,975 Obs.)	
	(1) Equal	(2) Unequal	(3) Equal	(4) Unequal	(5) Equal	(6) Unequal	(7) Equal	(8) Unequal
Cognitive Skills	0.868	0.961	0.856	0.955	0.915	0.964	0.994	1.002
	(0.029)	(0.015)	(0.022)	(0.010)	(0.021)	(0.011)	(0.033)	(0.012)
Parental Investments	0.172	0.093	0.045	0.024	0.039	0.017	0.011	−0.013
	(0.037)	(0.021)	(0.032)	(0.015)	(0.025)	(0.014)	(0.030)	(0.018)
Mother's Cognitive Skills	0.030	0.011	0.023	0.005	0.003	−0.006	0.022	0.019
	(0.016)	(0.008)	(0.015)	(0.007)	(0.014)	(0.007)	(0.019)	(0.009)
Total Factor Productivity	0.867	0.377	0.673	0.322	0.468	0.212	0.314	0.149
	(0.035)	(0.017)	(0.030)	(0.007)	(0.031)	(0.014)	(0.034)	(0.016)
Random Shock	0.133	0.064	0.094	0.034	0.058	0.028	0.034	0.024
	(0.012)	(0.006)	(0.010)	(0.005)	(0.009)	(0.004)	(0.009)	(0.007)

Notes. This table reports the parameter estimates and standard errors in parentheses. The standard errors are computed using the Delta method. The estimated technology equations are:

$$S_{i,t} = \alpha_{0,k} + \alpha_{1,k} S_{i,t-1} + \alpha_{2,k} I_{i,t-1} + \alpha_{3,k} S_{P,i} + \epsilon_{i,t}.$$

The model is estimated using full-information maximum likelihood (1,909 observations). Columns (1) and (2) relate to developmental stage 1 (starts at age 5-6 and ends at age 7-8). Columns (3) and (4) relate to developmental stage 2 (starts at age 7-8 and ends at age 9-10). Columns (5) and (6) relate to developmental stage 3 (starts at age 9-10 and ends at age 11-12). Columns (7) and (8) relate to developmental stage 4 (starts at age 11-12 and ends at age 13-14). Columns (1), (3), (5), and (7) are based on equal intervals. Columns (2), (4), (6), and (8) are based on unequal intervals.

Table A.7.16. Investment Function Parameters (Missingness per Stage): Period-by-Period

	Stage 1 (2,694 Obs.)		Stage 2 (2,981 Obs.)		Stage 3 (3,038 Obs.)		Stage 4 (2,975 Obs.)	
	(1) Equal	(2) Unequal	(3) Equal	(4) Unequal	(5) Equal	(6) Unequal	(7) Equal	(8) Unequal
Cognitive Skills	−0.007	−0.032	0.054	0.039	0.020	0.009	−0.001	−0.002
	(0.038)	(0.021)	(0.024)	(0.016)	(0.032)	(0.018)	(0.045)	(0.013)
Parental Investments	0.721	0.904	0.783	0.879	0.716	0.848	0.731	0.826
	(0.079)	(0.051)	(0.073)	(0.040)	(0.025)	(0.039)	(0.094)	(0.046)
Mother's Cognitive Skills	0.027	0.013	0.033	0.020	0.031	0.018	0.108	0.044
	(0.020)	(0.011)	(0.021)	(0.011)	(0.025)	(0.013)	(0.034)	(0.013)
Family Income (log)	0.008	−0.017	0.190	0.211	0.057	0.077	0.167	0.187
	(0.049)	(0.057)	(0.061)	(0.066)	(0.071)	(0.076)	(0.059)	(0.063)
Random Shock	0.039	0.024	0.063	0.037	0.108	0.064	0.041	0.025
	(0.025)	(0.016)	(0.028)	(0.023)	(0.038)	(0.023)	(0.025)	(0.016)

Notes. This table reports the parameter estimates and standard errors in parentheses. The standard errors are computed using the Delta method. The estimated technology equations are:

$$I_{i,t} = \beta_{1,k} S_{i,t-1} + \beta_{2,k} I_{i,t-1} + \beta_{3,k} S_{P,i} + \beta_{4,k} \ln y_{i,t-1} + \eta_{i,t}.$$

The model is estimated using full-information maximum likelihood (1,909 observations). Columns (1) and (2) relate to developmental stage 1 (starts at age 5-6 and ends at age 9-10). Columns (3) and (4) relate to developmental stage 2 (starts at age 9-10 and ends at age 13-14). Columns (1) and (3) are based on equal intervals. Columns (2) and (4) are based on unequal intervals.

Table A.7.17. Cognitive Skill Production Function Parameters Stage 1 (Age 5-6 through Age 7-8): Different Anchors

	Cognitive Skills		Parental Investments		Total Factor Productivity	
	(1) Equal	(2) Unequal	(3) Equal	(4) Unequal	(5) Equal	(6) Unequal
PIAT Reading Comprehension	0.906	0.974	0.193	0.103	0.929	0.349
	(0.039)	(0.020)	(0.042)	(0.022)	(0.057)	(0.023)
PIAT Mathematics	0.559	0.767	0.138	0.089	0.950	0.418
	(0.031)	(0.020)	(0.030)	(0.019)	(0.054)	(0.025)
How Often Mother Reads to Child	0.888	0.957	0.264	0.141	0.871	0.336
	(0.036)	(0.019)	(0.058)	(0.031)	(0.049)	(0.021)
Family Receives Daily Newspaper	0.888	0.957	0.270	0.145	0.871	0.335
	(0.036)	(0.019)	(0.066)	(0.035)	(0.049)	(0.021)
Visiting Music/Theater Performances	0.896	0.957	0.175	0.095	0.872	0.336
	(0.037)	(0.019)	(0.040)	(0.021)	(0.049)	(0.021)
Has Musical Instrument Child Can Use	0.888	0.960	0.285	0.158	0.871	0.334
	(0.036)	(0.019)	(0.071)	(0.037)	(0.050)	(0.021)

Notes. This table reports the parameter estimates and standard errors in parentheses. The standard errors are computed using the Delta method. The estimated technology equations are:

$$\mathrm{S}_{i,t} = \alpha_{0,k} + \alpha_{1,k}\mathrm{S}_{i,t-1} + \alpha_{2,k}\mathrm{I}_{i,t-1} + \alpha_{3,k}\mathrm{S}_{\mathrm{P},i} + \epsilon_{i,t}.$$

The model is estimated using full-information maximum likelihood (1,909 observations). Columns (1), (3), and (5) are based on equal intervals. Columns (2), (4), and (6) are based on unequal intervals.

Table A.7.18. Cognitive Skill Production Function Parameters Stage 2 (Age 7-8 through Age 9-10): Different Anchors

	Cognitive Skills		Parental Investments		Total Factor Productivity	
	(1) Equal	(2) Unequal	(3) Equal	(4) Unequal	(5) Equal	(6) Unequal
PIAT Reading Comprehension	0.805	0.907	0.002	0.005	0.669	0.368
	(0.027)	(0.015)	(0.030)	(0.015)	(0.058)	(0.010)
PIAT Mathematics	0.876	0.743	0.002	0.015	0.390	0.359
	(0.036)	(0.014)	(0.019)	(0.016)	(0.033)	(0.029)
Family Receives Daily Newspaper	0.906	0.966	0.012	0.008	0.577	0.312
	(0.025)	(0.013)	(0.051)	(0.026)	(0.048)	(0.021)
Visiting Music/Theater Performances	0.905	0.966	0.008	0.006	0.577	0.312
	(0.025)	(0.013)	(0.033)	(0.017)	(0.044)	(0.021)
Has Musical Instrument Child Can Use	0.923	0.956	0.006	0.012	0.580	0.314
	(0.024)	(0.013)	(0.051)	(0.027)	(0.043)	(0.021)

Notes. This table reports the parameter estimates and standard errors in parentheses. The standard errors are computed using the Delta method. The estimated technology equations are:

$$S_{i,t} = \alpha_{0,k} + \alpha_{1,k} S_{i,t-1} + \alpha_{2,k} I_{i,t-1} + \alpha_{3,k} S_{P,i} + \epsilon_{i,t}.$$

The model is estimated using full-information maximum likelihood (1,909 observations). Columns (1), (3), and (5) are based on equal intervals. Columns (2), (4), and (6) are based on unequal intervals.

Table A.7.19. Cognitive Skill Production Function Parameters Stage 3 (Age 9-10 through Age 11-12): Different Anchors

	Cognitive Skills		Parental Investments		Total Factor Productivity	
	(1) Equal	(2) Unequal	(3) Equal	(4) Unequal	(5) Equal	(6) Unequal
PIAT Reading Comprehension	0.869	0.932	0.035	0.020	0.388	0.196
	(0.032)	(0.017)	(0.023)	(0.013)	(0.047)	(0.023)
PIAT Mathematics	0.908	0.967	0.026	0.009	0.438	0.220
	(0.036)	(0.021)	(0.017)	(0.015)	(0.050)	(0.025)
Family Receives Daily Newspaper	0.942	0.976	0.114	0.045	0.409	0.204
	(0.028)	(0.014)	(0.066)	(0.031)	(0.049)	(0.023)
Visiting Music/Theater Performances	0.942	0.901	0.068	0.067	0.409	0.210
	(0.028)	(0.011)	(0.038)	(0.022)	(0.049)	(0.025)
Has Musical Instrument Child Can Use	0.942	0.976	0.082	0.033	0.409	0.204
	(0.028)	(0.014)	(0.046)	(0.023)	(0.049)	(0.023)

Notes. This table reports the parameter estimates and standard errors in parentheses. The standard errors are computed using the Delta method. The estimated technology equations are:

$$\mathrm{S}_{i,t} = \alpha_{0,k} + \alpha_{1,k}\mathrm{S}_{i,t-1} + \alpha_{2,k}\mathrm{I}_{i,t-1} + \alpha_{3,k}\mathrm{S}_{\mathrm{P},i} + \epsilon_{i,t}.$$

The model is estimated using full-information maximum likelihood (1,909 observations). Columns (1), (3), and (5) are based on equal intervals. Columns (2), (4), and (6) are based on unequal intervals.

Table A.7.20. Cognitive Skill Production Function Parameters Stage 4 (Age 11-12 through Age 13-14): Different Anchors

	Cognitive Skills		Parental Investments		Total Factor Productivity	
	(1) Equal	(2) Unequal	(3) Equal	(4) Unequal	(5) Equal	(6) Unequal
PIAT Reading Comprehension	0.931	0.963	0.017	0.004	0.316	0.118
	(0.040)	(0.021)	(0.034)	(0.015)	(0.074)	(0.025)
PIAT Mathematics	1.060	1.037	−0.005	−0.006	0.387	0.188
	(0.043)	(0.021)	(0.032)	(0.017)	(0.083)	(0.037)
Family Receives Daily Newspaper	0.994	1.001	0.021	0.007	0.248	0.127
	(0.033)	(0.017)	(0.056)	(0.031)	(0.056)	(0.027)
Visiting Music/Theater Performances	0.994	1.001	0.015	0.005	0.248	0.127
	(0.033)	(0.017)	(0.038)	(0.021)	(0.056)	(0.027)
Has Musical Instrument Child Can Use	1.032	1.001	−0.008	0.005	0.250	0.127
	(0.033)	(0.017)	(0.052)	(0.022)	(0.057)	(0.027)

Notes. This table reports the parameter estimates and standard errors in parentheses. The standard errors are computed using the Delta method. The estimated technology equations are:

$$\mathrm{S}_{i,t} = \alpha_{0,k} + \alpha_{1,k}\mathrm{S}_{i,t-1} + \alpha_{2,k}\mathrm{I}_{i,t-1} + \alpha_{3,k}\mathrm{S}_{\mathrm{P},i} + \epsilon_{i,t}.$$

The model is estimated using full-information maximum likelihood (1,909 observations). Columns (1), (3), and (5) are based on equal intervals. Columns (2), (4), and (6) are based on unequal intervals.

Chapter 3

Cross-Productivities of Executive Functions

By

Maximiliaan W. P. Thijssen
Mari Rege
Ingeborg F. Solli
Ingunn Størksen

Several studies suggest that executive functions are a key skill to target in early childhood education programs because they are strong predictors of academic development. However, there is little evidence on how preschool program-induced improvements in a child's executive functions promote other skills in primary school. We combine experimental data with a model of skill formation to provide such evidence, using an econometric decomposition framework. We find that preschool program-induced improvements in a child's executive functions led to improvements in mathematical skills and language skills in primary school. Our findings have implications for policies regarding school readiness, as they suggest a dynamic complementarity. In particular, primary school investments are more effective at promoting mathematical and language skills because prior investments in preschool have led to improved executive functions by the start of primary school.

Keywords. Executive functions, school readiness, self-productivity, cross-productivity, dynamic complementarity, randomized controlled trial.

JEL codes. I21, I24, J24, H75.

Author note. We received funding from Grant 270703 from the Norwegian Research Council. We thank Torberg Falch and participants in the Quantitative Forum at the University of Stavanger for helpful comments.

1. Introduction

Children start school with different skill levels. As these skill disparities are predictive of success in school, there has been growing interest in early childhood education programs that boost "school readiness" (*e.g.*, Clements & Sarama 2011, Diamond & Lee 2011, Dillon et al. 2017, Rege et al. 2021). One challenge in designing these programs is deciding which skills to target, as not all skills are equally important in school success (Duncan et al. 2007, Lewit & Baker 1995). Evidence showing which skills are most beneficial could inform public policy in the design of preschool education programs, with the goal of ensuring that every child is ready to learn at the start of formal schooling.

Several studies suggest that executive function (EF) skills are key skills to target. EF skills are the cognitive processes that control behavior, and are thought to provide a critical foundation for school readiness (Blair 2002, Blair & Razza 2007). The foundational importance of EF skills is reflected in the fact that they correlate with academic performance (*e.g.*, Blair 2002, Blair & Razza 2007), social-emotional skills (*e.g.*, Broidy et al. 2003), criminal activity (*e.g.*, Moffitt et al. 2011, Nagin & Tremblay 1999), (risky) health behaviors (*e.g.*, Miller et al. 2011, Moffitt et al. 2011), and other socioeconomic outcomes (see, *e.g.*, Diamond & Ling 2020, pp. 157–160). It is therefore important to investigate (i) whether preschool education programs can improve EF skills and (ii) whether these preschool program-induced improvements have a favorable impact on other skills in the course of formal schooling.

While there is evidence that early childhood education programs can improve EF skills (Diamond & Lee 2011, Diamond & Ling 2020), there is only limited evidence on whether program-induced improvements in EF skills lead to improvements in other skills. The child development literature (*e.g.*, Bierman et al. 2008, Raver et al. 2011) has identified EF skills as important mediators of intervention effects on school readiness. However, these studies measured EF and school readiness skills simultaneously, which sheds no light on the cross-productivity of program-induced EF skills.[1]

[1] McCoy et al. (2019) is an exception. They investigate the long-term effects of treatment-induced EF skills and academic skills on self-reported high school performance.

We investigate cross-productivities of EF skills by combining high-quality experimental data from the Agder Project (Rege et al. 2021) with an economic model of skill formation (Cunha & Heckman 2007). The Agder Project is a nine-month-long intervention in Norway targeting 701 five-year-old preschool children. Through a randomized controlled trial, Norway's (relatively) unstructured pedagogical tradition was compared with a comprehensive, structured curriculum. Rege et al. (2021) report a sizable, positive treatment effect on children's EF skills post-intervention. We exploit the experimental design to investigate how these preschool program-induced improvements in EF skills led to improvements in mathematical and language skills in the first grade of primary school.

Without additional assumptions, data from the experiment do not allow us to identify the degree to which EF skills promote mathematical and language skills, since the preschool program may improve unmeasured skills. Changes in EF skills may simply be proxies for changes in these unmeasured skills that affect mathematical and language skills. To address this issue, we apply the framework in Heckman et al. (2013) to identify causal effects and account for measurement error. We also assume that measured skills are independent of unmeasured skills, as do Berger et al. (2020), Conti et al. (2016), and Kosse et al. (2020). The framework in Heckman et al. (2013) can be applied if the intervals assumed by the model correspond to the intervals observed in the data. However, we generally observe children at different ages and intervals. Research suggests that failing to account for such child-specific, unequally spaced intervals can be problematic (Thijssen 2022, Voelkle & Oud 2013).[2] Moreover, as it takes time for causes to show their effects, and as the size of an effect depends on the observation interval (Gollob & Reichardt 1987, Kraft 2020), the results of our decomposition (and, by extension, its causal interpretation) would change when the observation interval changes. We therefore supplement the framework in Heckman et al. (2013) with insights from the continuous-time modeling literature (for a review, see Bergstrom 1988) to account for unequally spaced intervals.

[2]Voelkle & Oud (2013) conduct Monte Carlo simulations to study the effects of individual-specific unequal observation intervals for oscillating and non-oscillating processes. They conclude that it is important to account for individually varying time intervals.

We find that preschool program-induced improvements in EF skills led to improvements in mathematical skills and language skills in primary school. Furthermore, by interpolating the preschool and primary school periods, we are able to examine the time course of the treatment effect and the sources to which it can be attributed. Our findings suggest that the proportion of gains in mathematical and language skills attributable to program-induced improvements in EF skills increases over time, implying that EF skills at school entry become more important as children progress through school. Lastly, we find that accounting for (child-specific) unequally spaced intervals affect the estimated parameters that govern skill formation dynamics

We provide causal evidence on the role of EF skills in promoting other skills during a significant period in early childhood (*i.e.*, the transition from preschool to primary school), which contributes to our understanding of childhood skill formation. Moreover, our findings may indicate a dynamic complementarity between the preschool curriculum (*i.e.*, an investment in preschool process quality) and the investments made in primary school (*e.g.*, teacher quality, instructional practices). Indeed, primary school investments seem to be more productive in promoting mathematical and language skills for children who have experienced the structured preschool curriculum, relative to the preschool curriculum they would otherwise have experienced. The mechanism by which this occurs is improved EF skills. We also contribute by augmenting the framework outlined in Heckman et al. (2013) with insights from the continuous-time modeling literature. By incorporating the timing of observations, we produce more evidence on the importance of accounting for (child-specific) unequally spaced intervals when estimating skill formation models, building on Thijssen (2022). Such evidence has implications for the econometrics of human capital formation in early childhood (Cunha et al. 2021).

Our article contributes not only to the child development literature discussed above, but also to the literature concerning the economics of human development (see Heckman & Mosso 2014, for a recent survey). This strand of literature regards preschool education as an investment in skill formation (Blau & Currie 2006). Moreover, this literature considers cross-productivity and dynamic complementarity to be among the salient features of skill formation models (Cunha & Heckman 2007). Typically, researchers study either the effect of early childhood education programs (*e.g.*, Chor et al. 2016, Currie 2001, Gormley Jr & Gayer 2005, Jenkins et al. 2018) or cross-production of skills (*e.g.*, Cunha & Heckman 2008, Cunha et al. 2010). Few studies investigate those relationships jointly, which is unfortunate because doing so can provide valuable insights into the mechanisms by which differences subsequently emerge (or disappear) as a result of variations in preschool education.

Some notable exceptions include the following[3]: Heckman et al. (2013) investigate the mechanisms by which an early childhood education program conducted in the United States, the HighScope Perry Preschool Program, affected a variety of education, labor market, crime, and health outcomes. They find that persistent personality changes played a substantial role in the program's success. Second, Attanasio et al. (2020) investigate the mechanisms of an early childhood education program in Colombia. Their findings show that program-induced gains in cognitive and social-emotional skills are mainly the result of parents changing their investments. Additionally, they find that cognitive skills cross-produce social-emotional skills. While Heckman et al. (2013) and Attanasio et al. (2020) investigate the mechanisms of early childhood interventions, they do not focus on EF skills. Moreover, evidence on cross-productivities based on a skill dichotomy (see Attanasio et al. 2020) does not lead to a deeper understanding of skill formation because it fails to identify more fundamental abilities.

2. Conceptual Model

This section describes how the "*Technology of Skill Formation*" (Cunha & Heckman 2007) provides a framework for considering our research hypothesis. In particular, we explain the concepts of self-productivity and dynamic complementarity and how those concepts form the basis of our hypothesis. Before we explain our hypothesis, let us briefly review what is meant by the term "executive functions."

[3]Campbell et al. (2014) and Conti et al. (2016) are also exceptions, but they focus primarily on health behaviors. These studies examine the HighScope Perry Preschool Program and the Carolina Abecedarian Project; both are early childhood programs targeted at disadvantaged children in the United States.

2.1. What are Executive Functions?

The term executive function (EF, or cognitive control) refers to a set of interrelated, top-down processes needed for concentration and thinking (Diamond & Lee 2011). It is generally agreed that there are three core EF skills (Diamond 2013): working memory, inhibitory control (which overlaps substantially with self-regulation), and cognitive flexibility. The three core EF skills give rise to the higher-order EF skills of reasoning, problem-solving, and planning. Inhibition, working memory, and cognitive flexibility are central to learning, reasoning, problem-solving, and planning (Blair 2002, Diamond & Lee 2011), as well as to the regulation of attention, emotion, and behavior (Rueda et al. 2005).

The American Psychological Association (APA) dictionary (VandenBos 2007) defines "working memory" as "[...] *the short-term maintenance and manipulation of information necessary for performing complex cognitive tasks such as learning, reasoning, and comprehension,*" in keeping with the working model proposed by Baddeley and Hitch. The APA defines "inhibitory control" as "*the process of restraining one's impulses or behavior, either consciously or unconsciously* [...]." EF skills thus enable a child to "block" habitual behaviors and execute less familiar behaviors (Matsumoto & Tanaka 2004). Furthermore, working memory is key to knowing what to inhibit, while inhibition enables us to focus on specific content (Diamond 2016); this suggests that there is a reciprocal relationship between working memory and inhibitory control. Lastly, the APA defines "cognitive flexibility" as "*the capacity for objective appraisal and appropriately flexible action*." Cognitive flexibility, which requires working memory and inhibitory control, thus refers to the ability to view things from different perspectives and to "think outside the box." For an overview of the components that comprise EF skills and relationships to other concepts, see Figure 1 in Diamond (2016, p. 16). For an extensive review of the evidence on EF skills, see Diamond & Ling (2020).

2.2. A Model of Skill Formation

We start by introducing some notation and terminology. First, let $\mathbf{S}_{i,t}$ denote a vector the elements of which represent a child's skills, where t ($t = 0, 1, \dots, T$) indexes age over the T periods of childhood. We assume that these T periods cluster in $K \leq T$ stages of development ($k = 1, \ldots, K$). Let $\mathbf{I}_{i,t}$ denote investments. These investments can take various forms, such as good (*e.g.*, pedagogical toys), time spent with the child (*e.g.*, Del Boca et al. 2014), or measures to improve ECEC. We classify the Agder Project's structured curriculum as an ECEC type of investment (see Størksen et al. 2018, for details about the curriculum), and specifically as an investment in the process quality of preschool education. Process quality is defined by the interactions between children and teachers, the environment, and interactions among the children (Blau & Currie 2006, pp. 1183–1188).

A child develops skills when environmental influences (as experienced by the child) interact with skills previously acquired,[4]

$$\mathbf{S}_{i,t+1} = \mathbf{f}_k(\mathbf{S}_{i,t}, \mathbf{I}_{i,t}). \tag{1}$$

If these environmental influences are enriched, then it is said that an "investment" is made. Equation (1) thus provides a mathematical representation of human capital formation, defined as a dynamic and continuous interaction between a child's biology (*e.g.*, genes) and experiences in the environment (*cf.* National Research Council Institute of Medicine 2000).

One central assumption in the technology of skill formation is "self-productivity." Self-productivity encompasses the idea that skills are self-reinforcing and cross-producing. Self-reinforcement involves skills that are "alike," whereas cross-production involves skills that are "unlike." The following positive partial derivative formally defines self-productivity:

$$\frac{\partial \mathbf{S}_{i,t+1}}{\partial \mathbf{S}_{i,t}} > 0. \tag{2}$$

The assumed positive relationship between skills at age t and age $t + 1$ has three implications. First, if skills are self-reinforcing, then investments will not fully depreciate over a given length of time (all else being equal). Second, if skills are cross-productive, then investments will produce synergistic effects (all else being equal). Third, the stronger the self-reinforcement, the more stable the rank order of children from one age to the next will be (all else being equal).

[4]The technology of skill formation in Equation (1) is twice continuously differentiable, increasing and concave in investments (Cunha & Heckman 2007).

The reported evidence in Rege et al. (2021) suggests that the Agder Project's structured curriculum improved children's EF skills by the end of preschool, which is consistent with the first implication of self-productivity (Equation 2). While the development of EF skills has a biological basis, it also has a social basis, as the development of these skills is influenced by experiences in the environment (Blair 2006). This social basis explains how the structured curriculum, which focused on play-based learning, boosted children's EF skills by the end of preschool. While Rege et al. (2021) did not observe differences in EF skills between the treated and non-treated children in primary school, differences in mathematical skills started to emerge. A limitation of these findings is that the treatment effects in primary school capture the total difference between treated and non-treated, it is not clear *why* these differences emerged (*i.e.*, what mechanisms underlie these treatment effects).

Considering the second implication of self-productivity – if skills are cross-productive, then investments will produce synergistic effects – we might expect treated children who started primary school with better EF skills to develop more mathematical and language skills. Since EF skills are thought to be foundational in supporting other skills, it stands to reason that they may play a crucial role in boosting early mathematical and language skills. This expectation is consistent with empirical studies that document correlations between EF skills in preschool and academic development in school (Blair 2002, Blair & Raver 2015, Blair & Razza 2007). We are interested in testing this second expectation. Specifically, what is the effect of preschool program-induced improvements in EF skills on the formation of mathematical and language skills in primary school, relative to the EF skills children would have developed under a (relatively) unstructured preschool curriculum? To answer this question, we test the following hypothesis:

Hypothesis. *Mathematical and language skills in the first grade of primary school are more advanced for children who started primary school with higher levels of preschool program-induced executive functioning.*

While our hypothesis concerns the cross-productivity of EF skills, support for it might be indicative of a dynamic complementarity. Assume that EF skills are cross-productive of other skills in primary school. In the context of this study, the higher levels of EF skills observed at the start of primary school can be assumed to have resulted from an investment in preschool. Investments made in primary school may complement EF skills in producing other skills. If so, then a dynamic complementarity exists because current investments (*i.e.*, those made in primary school) are becoming more effective at producing other future skills, thanks to investments made in the past (*i.e.*, those made in preschool). The mechanism underlying such a dynamic complementarity would then be the more advanced EF skills at the start of primary school.[5] Evidence on such dynamic complementarities can inform public policy and the design of preschool education programs to ensure that all children are ready to learn at school entry.

While the fact that the Agder Project's structured curriculum improved EF skills, and that these skills are predictive of mathematical and language skills, is suggestive, it is not sufficient for concluding that targeting EF skills in preschool education programs will improve school readiness. For example, the structured curriculum may have improved skills we did not measure, such as social competence. Suppose that treated children improved their social competence and that social competence correlates with EF and early mathematical (or language) skills. In this case, the cross-productivity of EF skills may proxy the effect of social competence, resulting in erroneous conclusions concerning which skills early education programs should target. Section 5 addresses this concern using the framework described in Heckman et al. (2013).

3. The Agder Project

We will investigate the cross-productivity of EF skills using high-quality experimental data from the Agder Project. This section presents some background information on that project (see Rege et al. 2021, for further details and an analysis of the mean treatment effect). For an exposition and evaluation of the quality of the project's implementation, see Størksen et al. (2021). The high implementation quality of the Agder Project is important because inferences from compromised experiments can be incorrect (see, *e.g.*, Heckman & Karapakula 2021).

[5] See Heckman & Mosso (2014, pp. 696–698) for further details.

The welfare system in Norway includes generous social security and family policies. All one-to-five-year-old children are entitled to receive publicly regulated and subsidized preschool education and care. Preschool uptake amounts to about 98 percent among five-year-olds, the Agder Project's target population. Preschool centers in Norway typically group three-to-five-year-olds, with at least one teacher and two assistants per 14 to 18 children. These teachers generally have a bachelor's degree in early education and care. Children start school the year they turn six.

The social pedagogical tradition characterizes preschool education in Norway. This tradition emphasizes free play and natural curiosity. As such, it contrasts with school readiness approaches commonly used in English-speaking countries (OECD 2006). Research in psychology and education suggests that preschool curricula aimed at school readiness are more effective (Clements & Sarama 2011, Dillon et al. 2017); consequently, returns on investments in terms of skill formation may be sub-optimal in Norway's preschool centers.

This situation motivated the Agder Project, which aimed to foster school readiness and human potential through playful learning in preschool (Rege et al. 2021). The intervention consisted of a comprehensive curriculum with various age-appropriate activities aimed at stimulating EF skills, social competence, mathematical skills, and language skills (Størksen et al. 2018). The structured curriculum included games that challenge children to memorize and follow instructions (or, in some cases, to do the opposite) (Best & Miller 2010). These games targeted the three central EF skills: inhibitory control, working memory, and cognitive flexibility. Second, it included games that stimulate self-control, assertiveness, responsibility, cooperation, and empathy, all of which affect social competence. Because of a lack of relevant assessment tools for Norway, we were unable to measure social competence. Third, it engaged children in activities to familiarize them with numbers, measurement, geometry, and statistics, in an effort to foster early mathematical skills. Fourth, it stimulated early literacy through interactive book reading (Mol et al. 2009) and games that targeted letter and sound recognition to promote early language skills. The preschool teachers received training through a credit-based university class held during the academic year prior to the intervention and were offered coaching during the implementation phase.

Figure 1 shows the Agder Project's experimental design. In the preschool year 2015/2016, preschool teachers in the treatment group attended the credit-based university class and, as part of this training, provided extensive feedback on preschool activities, resulting in revisions of the curriculum. In August 2016, we assessed children's EF, mathematical, and language skills. This assessment is the baseline. The trained preschool teachers subsequently implemented the structured curriculum with the five-year-olds in the preschool center (in preschool year 2016/2017). Centers in the control group continued per usual, according to the social pedagogical tradition of free play and natural curiosity. Immediately following the intervention in June 2017, we assessed the children for the second time (post-intervention). The follow-up assessment in primary school took place in March 2018, after the children had started formal schooling. The preschool teachers in the control group participated in the credit-based university class and received the intervention materials in the preschool year 2017/2018, after the participating children had left preschool.

We received parental consent for 701 children (92 percent) from a total of 71 preschool centers with varying group sizes.[6] Seventy-two preschool centers initially signed up, but one center from the control group subsequently withdrew from the study. The largest preschool center included 25 participating children, while the smallest center had only two children for whom parental consent had been obtained). We grouped the centers into 15 (randomization) blocks based on location and the number of children in the preschool center. The number of children per block ranges from 29 to 92. We then randomized the centers into treatment and control groups within each block. This procedure resulted in 35 control centers (318 children) and 36 treated centers (383 children). There has been relatively little attrition, and very few observations are missing.[7]

[6]We test differences in baseline characteristics by regressing premeasured and predetermined variables, alternately, on a treatment indicator and randomization-block fixed effects. All baseline characteristics are balanced, including baseline skills of children. We also allow for multiple hypothesis testing using the Romano & Wolf (2005) procedure. None of the differences are significant. See Table A.1.3 in Appendix A.1. for the results of the balance test.

[7]See Table A.1.2 and Table A.1.3 in Appendix A.1 for an overview of observed and missing data in the Agder Project.

Playful Learning:
Towards a More Intentional Practice in Norwegian Preschool Groups

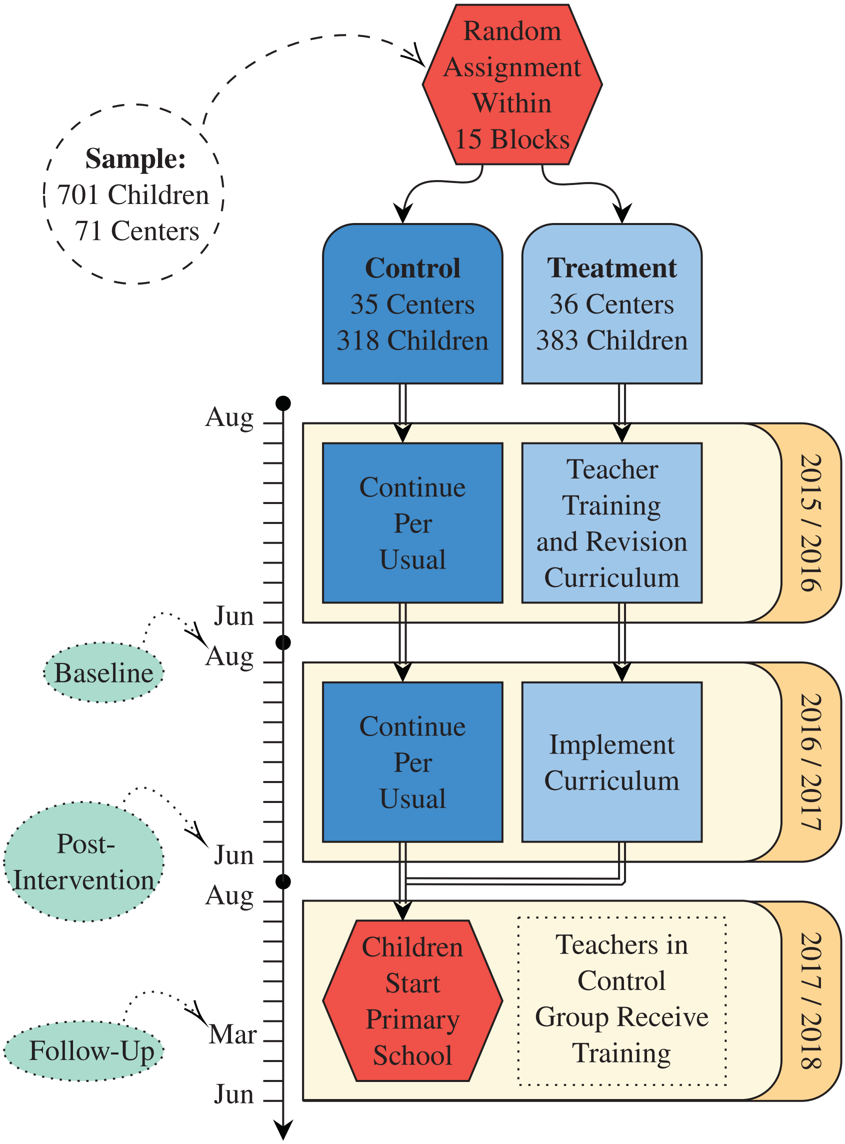

Figure 1. Experimental Design of the Agder Project

4. Data

During the first two assessment waves, the preschool centers' classes met in a science museum, where the children engaged in museum activities throughout the day. Each teacher would bring children to an assessment station at a scheduled time. The third measurement wave took place during the first grade of primary school. Testers had to travel to the primary schools to administer the tests. During each assessment wave, testers, who were trained, certified, and blind to treatment status, administered six tests: (1) the Ani Banani Math Test; (2) the Norwegian Vocabulary Test; (3) a Blending Test measuring phonological awareness; (4) the Hearts and Flowers Test; (5) the Head-Toes-Knees-Shoulders Test; and (6) the Forward and Backward Digit Span Test. We assume that the Ani Banani Math Test results reflect mathematical skills, while performance on the Norwegian Vocabulary Test and the Blending Test reflect language skills. Finally, performance on the Forward and Backward Digit Span Test, the Hearts and Flowers Test, and the Head-Toes-Knees-Shoulders Test provides an indication of the children's EF skills. We provide further details about each test below.

Figure 2 shows descriptives calculated over all non-missing observations for each of the six assessments. We plot the mean and standard deviation across all three assessment waves (baseline 2016, post-intervention 2017, and follow-up 2018). The length of the whiskers signifies one standard deviation above (and below) the mean. Figure 2 shows a strong mean development across all six tests from the start of the last year in preschool (*i.e.*, August 2016) to midway through primary school (*i.e.*, March 2018). In Appendix A.2, we use the baseline skills test balance (see Table A.2.3).

4.1. Mathematical Skills

We used the Ani Banani Math Test to assess children's mathematical skills (Størksen & Mosvold 2013), selecting 11 out of the 18 items. We dropped two items because almost all children answered them correctly (*i.e.*, filler items). During the third assessment wave, technical problems with the tablet computer application caused another five items to become unusable. We omit these five items in each assessment wave to maintain consistency with Rege et al. (2021).

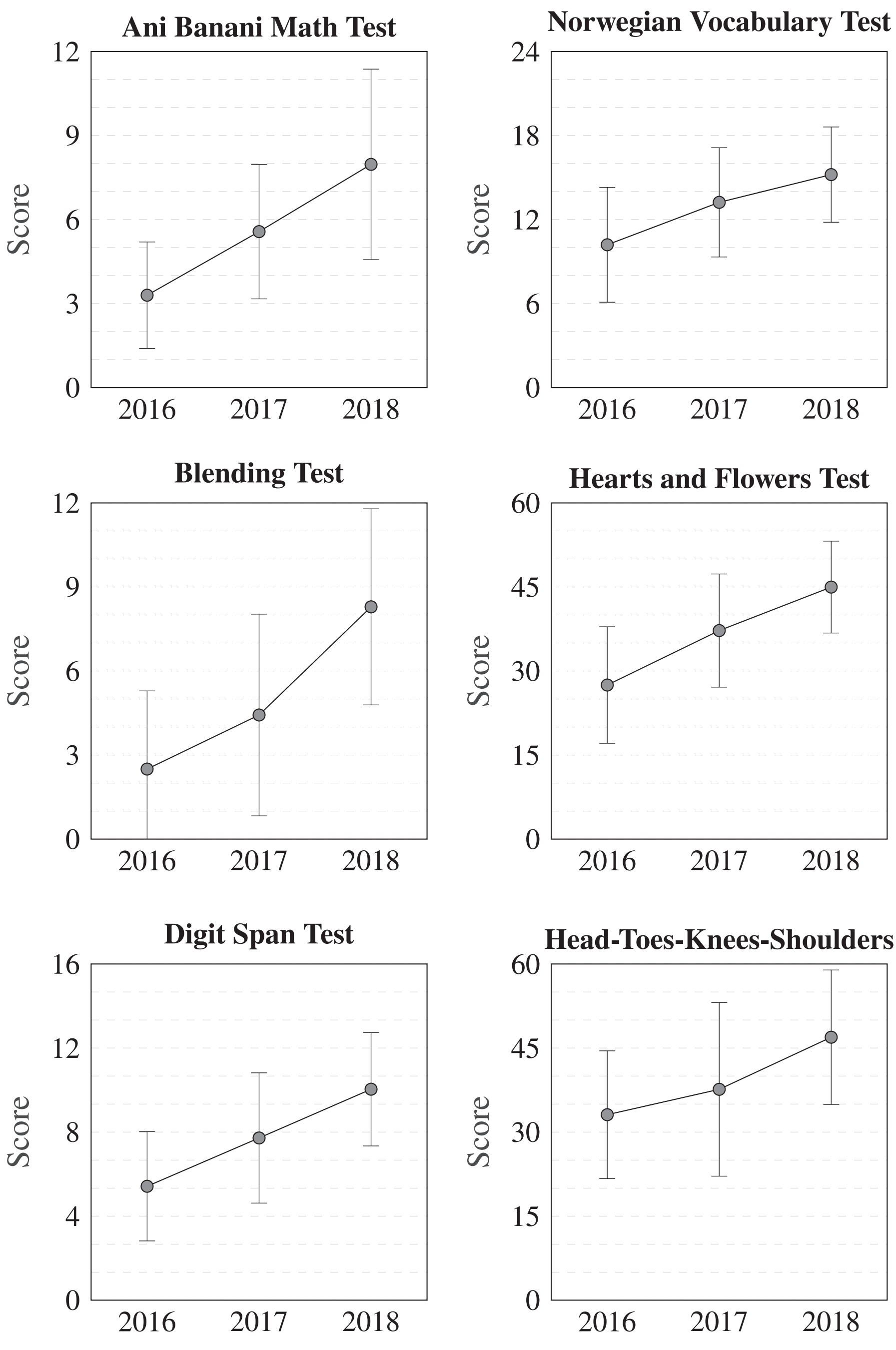

Figure 2. Mean Score and Standard Deviation by Assessment Wave

During the Ani Banani Math Test, testers asked children to help the monkey Ani Banani with such tasks as counting bananas or setting a table with the correct number of plates. The test takes about ten minutes to complete. While it covers three areas of mathematics (*i.e.*, numeracy, geometry, and problem-solving), it does not reliably distinguish between these areas; it can therefore be assumed to measure one general construct of mathematics development (ten Braak & Størksen 2021). ten Braak & Størksen (2021) confirm (i) good concurrent validity, (ii) good discriminant validity when contrasted with measures of EF skills and language, and (iii) predictive validity for mathematics achievement five years later.[8]

4.2. Language Skills

We used the Norwegian Vocabulary Test (Størksen et al. 2013) and a Blending Test to assess early language skills. During the Norwegian Vocabulary Test, pictures appeared on the tablet computer, and the tester would then ask the child to identify the picture. Children received a point for each correct answer, with a total of 20 possible points.

The Blending Test measures phonological awareness. Testers presented a target word with its phonemes and asked the children to select the corresponding picture from four appearing on the tablet computer. Each correct response earned children one point, with a total of 12 possible points. While the Blending Test theoretically reflects children's language skills, it may be a weak test in practice, since it was originally designed for pedagogical rather than research purposes. The Blending Test simply assesses whether or not children can read, so the distribution of the sum of items is not normal. By contrast, measurement instruments such as the Head-Toes-Knees-Shoulders Test are designed and validated specifically for research purposes (McClelland et al. 2007, 2014).

[8]Concurrent validity measures the extent to which a measures correlates with an established, validated instrument. Discriminant validity measures the extent to which measures that should not relate to one another are in fact not highly related. Of course, one would expect measures that are meant to measure the same construct to correlate more highly with each other than with other measures that are meant to measure other constructs. While there will always be correlations between measures of child development, the idea is that these correlations should be lower for those who are not meant to measure the same construct. Predictive validity measures the extent to which the concept being measured by the instrument predicts an outcome of interest.

4.3. Executive Function Skills

The first measure of EF skills is the Hearts and Flowers Test (Davidson et al. 2006), which assesses children's inhibitory control and cognitive flexibility using tablet computers. The children were instructed to press a key on the same side as the stimulus when they saw a heart and on the opposite side when the stimulus was a flower. The child received a point for each correctly pressed key. The test consisted of 60 stimuli.

The second measure is the Head-Toes-Knees-Shoulders Test (McClelland et al. 2014), which integrates inhibitory control, cognitive flexibility, and working memory demands into a self-regulation task. The test consists of three blocks with ten items per block. For each item, children received two points when they the task correctly, one point when they carried out an incorrect movement but ended with a correct response, and zero points for incorrect responses. McClelland et al. (2014) report the psychometric properties of this test. The test relates to inhibition, working memory, and cognitive flexibility when it comes to construct validity. Furthermore, the test predicts academic achievement, particularly from kindergarten to first grade (Lenes et al. 2020).

The last measure is the Forward and Backward Digit Span Test, a component of the Wechsler Intelligence Scale for Children (Wechsler 1991). The children were asked to listen to a sequence of digits voiced by the tester, then to repeat back the sequence of digits. The forward digit span test simply assesses short-term (auditory) memory, as children are not required to manipulate the information. By contrast, the backward digit span test measures the child's ability to manipulate verbal information in temporary storage (*i.e.*, working memory). A child's total score is the sum of the combined forward and backward digit span tests and reflects the total number of correctly repeated digit sequences.

4.4. Child and Parental Characteristics

We matched the Agder Project's assessment data to Statistics Norway's registry data. As part of the Agder Project, we also collected data on sex, birth month, randomization-block indicators, and whether we received late parental consent. In Statistics Norway's registry data, we observe education levels for the mother and father, income for both parents, and the parents' country of origin. Using the information on country of origin, we construct an indicator variable that takes on the value of one if at least one of the parents is a non-Western immigrant and zero otherwise. In Appendix A.1, we use these child and parental characteristics to conduct a balance test (see Table A.1.3). Following Rege et al. (2021), we do not include categories for the birth month or the parents' education in our model. Instead, we estimate a single parameter for each of these variables.

Table 1 reports descriptive information for the child and parental characteristics (see Table A.1.1 in Appendix A.1 for an overview of observed and missing data). In Table 1, we observe that about half of the children are female. The median educational attainment for mothers is the first stage of higher education (undergraduate level), whereas for fathers it is the upper secondary (final) year. On average, about 15 percent of the sampled children have at least one parent who is a non-Western immigrant. Mean family income is about NOK 888,416, 63 percent earned by the father (NOK 555,051) and 37 percent by the mother (NOK 330,587).

Table 1. Descriptive Statistics Child and Parental Characteristics

	(1) Mean	(2) *SD*	(3) Obs.
The Child is Female	0.49	0.50	701
Birth Month	6.00[a]	3.19	701
Education Mother	6.00[a]	1.67	676
Education Father	4.00[a]	1.59	666
Non-Western Immigrant	0.15	0.36	682
Income Mother (in NOKs)	330,587	213,546	698
Income Father (in NOKs)	555,051	268,071	683
Late Parental Consent	0.19	0.39	701

Notes. This table reports descriptive statistics for the child and parental characteristics. Sex and birth month are part of the project's data collection. The parents' education, country of origin, and income are obtained from Statistics Norway. Education comprises eight categories: (1) primary education; (2) lower secondary education; (3) upper secondary (basic); (4) upper secondary (final year); (5) post-secondary, not higher education; (6) first stage of higher education, undergraduate level; (7) first stage of higher education, graduate-level; and (8) second stage of higher education (postgraduate education).
[a] We report the median rather than the mean.

5. Empirical Strategy

This section describes our empirical strategy. In particular, we explain how differences between treated and non-treated children in primary school can be decomposed into program-induced improvements by the end of preschool. This decomposition closely follows Heckman et al. (2013). We then augment the linear decomposition methodology with the exact discrete model from the continuous-time modeling literature (see Bergstrom 1988) as applied to the technology of skill formation in Thijssen (2022). The analysis in Thijssen (2022) builds on earlier work by, among others, Hamerle et al. (1991), Hamerle et al. (1993), Singer (1993, 1995), Oud & Jansen (2000), and Voelkle & Oud (2013). We provide a detailed exposition of the exact discrete model in Appendix A.2. Since the six tests are measured with error, we specify measurement models. We present further details concerning these measurement models (and their identification) in Appendix A.3. The last section describes our multi-step estimation algorithm and bootstrap procedure. In Appendix A.4, we present additional information concerning these topics.

5.1. A Linear Framework for Decomposing Treatment Effects

We consider a linear-in-parameters production function. We assume that the first stage of development extends from August 2016 (age t) to June 2017 (age $t + 1$) and that the second developmental stage extends from June 2017 (age $t + 1$) to March 2018 (age $t + 2$). It follows from Equation (1) that the parameters are invariant within these stages with respect to time.

Let d index treatment assignment so that $d = 1$ if a child attends a treated preschool center and $d = 0$ otherwise, $d \in \{0, 1\}$. Because of our linear-in-parameters assumption, we can write the first developmental stage as a multivariate equation,[9]

$$\mathbf{S}_{d,i,t+1} = \mathbf{a}_d + \mathbf{B}\mathbf{S}_{d,i} + \mathbf{C}\mathbf{X}_{d,i} + \mathbf{w}_{d,i,t+1} \qquad (3)$$

In Equation (3), $\mathbf{S}_{d,i,t+1}$ denotes a H-dimensional vector representing the

[9] Equation (3) relates to the Neyman-Fisher-Quandt-Rubin potential outcomes framework through the following equation: $\mathbf{S}_{i,t} = \mathbf{D}\mathbf{S}_{1,i,t} + (1 - \mathbf{D})\mathbf{S}_{0,i,t}$.

counterfactual skill set at age $t + 1$ (*i.e.*, post-intervention, June 2017). Second, $\mathbf{a}_d$ is a H-dimensional vector with scalar intercept parameters. Third, $\mathbf{B}$ is (H, H)-dimensional matrix of scalar parameters that characterize the extent to which skills acquired at age t (*i.e.*, baseline, August 2016) are self-productive. The parameters on the diagonal measure how skills reinforce themselves, and the off-diagonal parameters measure how skills cross-produce each other over the period from August 2016 to June 2017. Fourth, $\mathbf{C}$ is a (H, P)-dimensional matrix with scalar parameters measuring how child and parental characteristics (*i.e.*, the child and parental characteristics described in Table 1 plus randomization-block indicators) affect skill formation. The last term, $\mathbf{w}_{d,i,t+1}$ is a H-dimensional vector representing idiosyncratic, zero mean shocks.

Since treatment assignment is random (Rege et al. 2021), we know, by definition, that observed (*i.e.*, $\mathbf{S}_{d,i,t}, \mathbf{X}_{d,i}$) and unobserved (*i.e.*, $\mathbf{w}_{d,i,t+1}$) variables balance in expectation. It follows, then, that the mean difference, $\mathbb{E}(\mathbf{S}_{1,i,t+1} - \mathbf{S}_{0,i,t+1}) = \mathbf{a}_1 - \mathbf{a}_0$, where $\mathbb{E}(\cdot)$ denotes the expectation operator, identifies program-induced improvements in skills.[10]

The second developmental stage is different. First, the program may have improved variables we did not measure (*e.g.*, social competence). Second, the program may have changed the extent to which skills acquired at age t are self-productive (*i.e.*, $\mathbf{B} = \mathbf{B}_d$). Third, the program may have changed the extent to which child and parental characteristics affect skill formation (*i.e.*, $\mathbf{C} = \mathbf{C}_d$). Let $\mathbf{U}_{d,i,t+1}$ denote a R-dimensional vector with unmeasured variables affected by the structured curriculum and let $\tilde{\mathbf{B}}_d$ denote the (H, R)-dimensional matrix with scalar parameters that measure the effect of unmeasured variables on measured skills. For the second developmental stage, we can then write the following multivariate model:

$$\mathbf{S}_{d,i,t+2} = \mathbf{a}_d + \mathbf{B}_d\mathbf{S}_{d,i,t+1} + \mathbf{C}_d\mathbf{X}_{d,i} + \tilde{\mathbf{B}}_d\mathbf{U}_{d,i,t+1} + \mathbf{w}_{d,i,t+2}. \qquad (4)$$

To simplify the decomposition, we assume that $\mathbf{B}_1 = \mathbf{B}_0$ and $\mathbf{C}_1 = \mathbf{C}_0$. The treatment affected skills in primary school, but not self-productivity or the effect of child and parental characteristics. Parenthetically, we test and fail to reject this hypothesis. For details on the intuition of this test, see Heckman et al. (2013). Thus, we proceed by writing $\mathbf{B}_d = \mathbf{B}$ and $\mathbf{C}_d = \mathbf{C}$.

[10] In Appendix A.1, we report the results of a balance test (see Table A.1.3).

We rewrite Equation (4) as follows:

$$\mathbf{S}_{d,i,t+2} = \tilde{\mathbf{a}}_d + \mathbf{B}\mathbf{S}_{d,i,t+1} + \mathbf{C}\mathbf{X}_{d,i} + \tilde{\mathbf{w}}_{d,i,t+2}, \tag{5}$$

where we defined $\tilde{\mathbf{a}}_d = \mathbf{a}_d + \tilde{\mathbf{B}}_d\mathbb{E}(\mathbf{U}_{d,i,t+1})$ as the new H-dimensional vector with intercepts and $\tilde{\mathbf{w}}_{d,i,t+2} = \mathbf{w}_{d,i,t+2} + \tilde{\mathbf{B}}_d[\mathbf{U}_{d,i,t+1} - \mathbb{E}(\mathbf{U}_{d,i,t+1})]$ as the new H-dimensional vector with random zero mean shocks.

If we further assume that measured and unmeasured skills are (statistically) independent conditional on the child and parental characteristics for the no-treatment equations, then we can identify treatment effects owing to measured skills from the difference in expectations (for further details, see Heckman et al. 2013, pp. 2060–2063),

$$\overbrace{\mathbb{E}(\mathbf{S}_{1,i,t+2} - \mathbf{S}_{0,i,t+2})}^{\substack{\text{Follow-Up}\\ \text{Treatment Effect}}} = \overbrace{(\tilde{\mathbf{a}}_1 - \tilde{\mathbf{a}}_0)}^{\substack{\text{Contribution of}\\ \text{Unmeasured}\\ \text{Variables}}} + \overbrace{\mathbf{B}\mathbb{E}(\mathbf{S}_{1,i,t+1} - \mathbf{S}_{0,i,t+1})}^{\substack{\text{Contribution of}\\ \text{Measured Variables}}} \tag{6}$$

Thus, for the present decomposition, we maintain the same identifying assumption as Heckman et al. (2013), Berger et al. (2020), Conti et al. (2016), and Kosse et al. (2020) concerning the independence of measured and unmeasured skills.

Thus far, we have maintained the assumption, albeit implicitly, that we observe children at the same age and in equally spaced unit intervals. In other words, the intervals assumed by the models in Equations (3) and Equation (5) correspond to the intervals observed in the data. Most longitudinal studies target same-aged children based on the (pre)school year. In general, the development of children born in the same year and measured with consistent regularity can differ by as much as 12 months at the time of measurement. Such age differences can have a significant impact on cognitive achievement (see, *e.g.*, Cascio & Lewis 2006, Crawford et al. 2010, Solli 2017), and parameter estimates of skill formation models (Thijssen 2022). We rarely observe all children with such regularity, however. Irregularities may, for example, arise because of the logistics involved in surveying a large number of geographically dispersed children. Therefore, we apply insights from the continuous-time modeling literature to account for these (child-specific) unequal intervals.

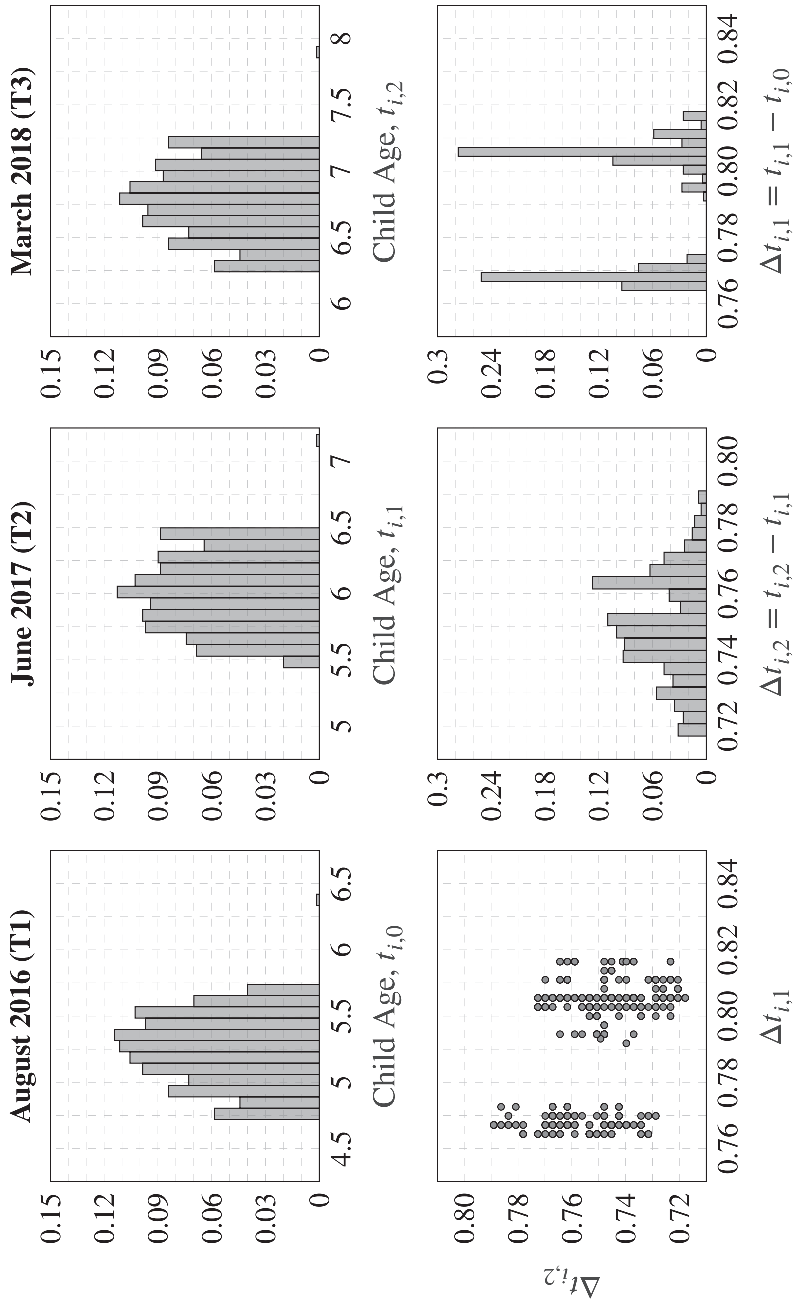

Figure 3. Descriptives on Child-Specific Unequal Observation Intervals in the Agder Project Data

5.2. The Estimation Problem of Unequally Spaced Intervals

Studies in econometric methods show that failing to account for unequally spaced intervals can lead to biased estimation (Baltagi & Wu 1999, Jones & Boadi-Boateng 1991, McKenzie 2001, Millimet & McDonough 2017, Rosner & Muñoz 1988, Sasaki & Xin 2017). Thijssen (2022) estimates the technology of cognitive skill formation and finds that estimates are greatly impacted by failing to account for unequally spaced intervals. We apply the exact discrete model approach in Thijssen (2022) and Voelkle & Oud (2013) because we assume age is the relevant timing variable for skill development. Consequently, unequally spaced intervals can be child specific. Children assessed simultaneously may not be the same age, and same-aged children may not be assessed simultaneously. For an exposition of this approach in skill formation models, see Thijssen (2022) or Appendix A.2.

We use the information on the child's birth month and assessment dates to calculate the age at the time of assessment. In Figure 3 (first row), we plot probability frequency distributions of the child's age during assessment for each assessment wave (*i.e.*, August 2016 (T1), June 2017 (T2), and March 2018 (T3)). As expected, children differ by up to 12 months in age in any given assessment wave. In Figure 3 (second row), we plot the probability frequency distributions of the temporal distance (in years) between the assessment waves. As expected, we observe less variation in the distance between assessment waves in preschool. During the first two assessment waves, the preschool centers met in a local science museum. Testers traveled to the primary schools during the third assessment wave, which is more challenging in terms of logistics. Nonetheless, the variation in the temporal distance between assessment waves appears to be limited. Finally, we show a scatter plot of the (temporal) distance between the first and second assessment waves and the second and third. If children were assessed consistently and at the same age, we should see a single dot, yet we observe some variation.

We apply insights from continuous-time modeling to account for the child-specific unequal observation intervals shown in Figure 3. In particular, we apply the exact discrete model (see Bergstrom 1988). The intuition is as follows: Equation (3) and Equation (5) belong to stochastic differential equations with continuous-time parameters that do not depend on the (temporal) distance. The solutions to these equations (see Arnold 1974, pp. 128–134) reveal constraints that link the discrete-time model parameters and continuous-time model parameters in an exact way for a given observation interval. By allowing this observation interval to be unequal and child-specific (see, *e.g.*, Thijssen 2022, Voelkle & Oud 2013) and applying those constraints during estimation, we obtain the continuous-time model parameters. We can then solve the continuous-time model for the unit interval and obtain the discrete-time parameters assumed by the model in Equation (3) and Equation (5), but taking into account the child-specific unequal intervals observed in the data (Figure 3). Appendix A.2 presents a detailed exposition of the exact discrete model and the parameter constraints.

5.3. Specifying a Linear Measurement Model

The children's skills in Equation (3) and Equation (5) are not directly observable. Instead, we observe manifestations in the form of test scores that we assume to be consistent with a particular skill level. Suppose there are multiple test scores for a given skill. In that case, we could standardize the individual test scores first, construct a simple arithmetic average, and re-standardize so that the composite variable is mean zero and standard deviation one. However, such "*measurement by fiat*" (Torgerson 1958, p. 22) is generally not recommended, as many have demonstrated (Cunha & Heckman 2008, Cunha et al. 2010) and argued (Borghans et al. 2008, Cunha et al. 2021, Skrondal & Rabe-Hesketh 2004). Therefore, we specify a measurement model that addresses (classical) measurement error and weighs each test score based on its level of informativeness regarding the skill it manifests. Appendix A.3 presents results based on a simple arithmetic average.

The "*task performance function*" in Borghans et al. (2008, p. 978) motivates our measurement model as it defines observed test scores (*i.e.*, manifest variables) as a function of unobserved skills (*i.e.*, common factors) and other latent influences (*i.e.*, unique factors). Formally, let $\mathbf{M}_{h,d,i,t}$ denote a L_h-dimensional vector with (observed) manifest variables in which skill h manifests at age t. Since we observe the same manifest variables in each period, we omit a time subscript for L_h. We assume that the manifest variables are additively separable in the common factors they represent. It follows, then, that we can write the following linear system of measurement models:[11]

$$\mathbf{M}_{h,d,i,t} = \boldsymbol{\mu}_{h,t} + \boldsymbol{\lambda}_{h,t}\mathrm{S}_{h,d,i,t} + \boldsymbol{\zeta}_{h,d,i,t}. \tag{7}$$

In Equation (7), $\boldsymbol{\mu}_{h,t}$ denotes a L_h-dimensional vector of intercepts. Furthermore, $\boldsymbol{\lambda}_{h,t}$ denotes a L_h-dimensional vector of factor loadings. These factor loadings weigh each manifest variable based on their informativeness concerning the common factor, $\mathrm{S}_{h,d,i,t}$. Lastly, $\boldsymbol{\zeta}_{h,d,i,t}$ is a L_h-dimensional vector with unique factors (*i.e.*, measurement-specific influences and measurement errors). We can conceive of $\mathrm{S}_{h,d,i,t}$ as a common factor because $\mathrm{S}_{h,d,i,t}$ is *common* across the L_h manifest variables assumed to measure skill h. We can conceive of $\boldsymbol{\zeta}_{h,d,i,t}$ as unique factors because the elements of $\boldsymbol{\zeta}_{h,d,i,t}$ manifest *uniquely* in each manifest variable. Implicit in Equation (7), we assume that the measurement properties are invariant across treatment status and children.

Since none of the right-hand-side variables in Equation (7) are observable, there is an inherent identification problem. First, we require some normalization to set a scale and location for the factors. We normalize one of the factor loadings (say the first) to one to set a scale. To set the location, we normalize the mean of the common factor to zero. Second, we assume (1) independence between the common and unique factors and (2) independence between the unique factors, conditional on the common factor. We also assume that the unique factors are independent across children. If the measures are continuous, then these assumptions and normalizations are sufficient for identifying the measurement model in Equation (7) (Anderson & Rubin 1956).

We can identify the factor loadings from the ratio of covariances. With the factor loadings identified, we can (nonparametrically) identify the distribution of the factors by applying Kotlarski's lemma (see Lemma 1, Remark 4, and Remark 5 Kotlarski 1967, pp. 70–73). However, these identification results no longer hold when the manifest variables are categorical (*e.g.*, ordinal, dichotomous). In those cases, we assume a known distribu-

[11] We assume that each common factor manifests in at most one manifest variable. We do not require this assumption for identification, but it facilitates the interpretation. Also, the measures are designed to measure specific skills (see the referenced literature in Section 3).

tion for the unique factors. We also require further normalizations since the variances in polychoric (or tetrachoric) correlation matrices are redundant. We normalize the unique factor variances to unity to achieve (local) identification. Appendix A.3 provides further notes on identification.

A final consideration is the scale of the common factor. We anchor the scale of the common factor in the scale of one of the tests. We anchor mathematical skills in item 15 of the Ani Banani Math Test. Item 15 asks each child to copy a pattern appearing on the tablet computer. For EF skills, we use the Forward and Backward Digit Span Test. The score on this test represents the total number of correctly repeated number sequences. Lastly, we anchor language in the Blending Test. The score on this test is the total number of correctly chosen alternatives from four pictures.

5.4. Estimation Procedure

Before implementing our multi-step estimation procedure, we estimate measurement models for the Norwegian Vocabulary Test, the Blending Test, the Hearts and Flowers Test, and the Head-Toes-Knees-Shoulders Test using the individual items. As discussed, when manifest variables have a categorical scale, we need to make distributional assumptions. Since we have no prior, we estimate the models under varying distributional assumptions and decide based on Akaike's Information Criterion (AIC: Akaike 1987) and the Bayesian Information Criterion (BIC: Schwarz 1978). See Appendix A.3 for details. We then use the predicted factor scores for these tests in measurement models for language and EF skills. The prediction error associated with these predictions becomes part of the measurement error.

We use a multi-step estimation algorithm, following Heckman et al. (2013, p. 2066). In the first step, we estimate the measurement models for EF skills, mathematical skills, and language skills (we present these results in Appendix A.3). In the second step, we predict (Bartlett) factor scores (Bartlett 1937, Thomson 1938) using

$$\hat{\mathrm{S}}_{h,d,i,t} = (\hat{\boldsymbol{\lambda}}'_{h,t}\widehat{\boldsymbol{\Sigma}}^{-1}_{\zeta,h,t}\hat{\boldsymbol{\lambda}}_{h,t})^{-1}\hat{\boldsymbol{\lambda}}'_{h,t}\widehat{\boldsymbol{\Sigma}}^{-1}_{\zeta,h,t}(\mathbf{M}_{h,d,i,t} - \widehat{\boldsymbol{\mu}}_{h,t}) \quad (8)$$

where $\hat{\boldsymbol{\lambda}}'_{h,t}$ is the L_h-dimensional vector with estimated factor loadings,

$\widehat{\boldsymbol{\Sigma}}^{-1}_{\zeta,h,t}$ is a (L_h, L_h)-dimensional matrix with estimated unique factor variances, and $\widehat{\boldsymbol{\mu}}_{h,t}$ is a L_h-dimensional vector with estimated intercepts. The third step estimates the models outlined in Section 5.1 using the predicted factor scores. We apply Croon's correction method in the last step (Croon 2002). The intuition behind Croon's correction method is to use our knowledge of the common factor variance and unique factor variance (from the first step) to adjust the estimates for prediction error. See Appendix A.4 for details. We assume data are missing at random, to address the issue of missing data, and estimate the models using full information parametric maximum likelihood (Anderson 1957). We assume a normal distribution for the error terms. Note that we do not require this assumption for identification. See Appendix A.1 for details regarding missingness.

After applying this correction, the standard errors are incorrect. We apply a clustered wild residual bootstrap procedure. We draw 1,000 bootstrap samples from the original data and apply the multi-step estimation algorithm to each pseudo-sample. We cluster and re-sample at the (randomization) block level to ensure that each bootstrap sample includes both treated and non-treated children. Lastly, we use the "true" variance, not the variance of the predicted factor score, which is biased, to standardize the interpretation of the parameters.

6. Results

In Section 6.1, we start by presenting the parameter estimates of self-reinforcement and cross-production for both stages of development. Then, we discuss the post-intervention and follow-up treatment effects. Lastly, we discuss how the follow-up treatment effect results from program-induced improvements in EF, mathematical, language, and unmeasured skills post-intervention.

Section 6.2 presents results that account for unequally spaced intervals. In particular, we characterize the time-course of the post-intervention and the follow-up treatment effect for two scenarios, with and without accounting for unequally spaced intervals. Finally, we use the characterization of the follow-up treatment effect to construct an area plot, the areas of which represent the relative importance of EF, mathematical, language, and unmeasured skills in producing the follow-up treatment effect.

6.1. A Linear Decomposition under Equal Intervals

Table 2 reports the self-reinforcement and cross-production parameter estimates and standard errors. Panel A reports results based on Equation (3), whereas the results in Panel B are derived from the model in Equation (5). Each column reports estimates from one model. There are two panels and six columns, so the estimates in Table 2 represent twelve models. The control variables are the child's sex, birth month, whether or not at least one of the parents is a non-Western immigrant, parental education, family income, an indicator for late parental consent, and randomization-block indicators.

The results in Table 2 suggest the following: First, EF and mathematical skills show strong persistence in both stages of development (the diagonal cells). In our preferred model specification (*i.e.*, the model with control variables), we estimate an auto-regressive parameter of 0.774 for EF skills and 0.663 for mathematical skills (both statistically significant at the one percent level) in developmental stage 1. In developmental stage 2, we find a similar level of persistence for mathematical skills, namely 0.635, but the persistence of EF skills increases to 0.943 (both statistically significant at the one percent level). High persistence implies that (effective) investments depreciate more slowly. Such parameter estimates also imply a stable rank order of children from one measurement occasion to the next. Second, we observe that skills are cross-productive in the first development stage, particularly EF and mathematical skills. In the second developmental stage, we observe that EF skills remain cross-productive. By contrast, mathematical skills promote only EF skills, and language skills do not seem to boost either EF or mathematical skills. This finding aligns with the studies that have determined that EF skills predict success in school (Blair 2002, 2006, Blair & Raver 2015, Blair & Razza 2007).

Table 3, Panel A, reports the parameter estimates and standard errors associated with the post-intervention and follow-up treatment effects. The post-intervention treatment effect answers the question of how much more the EF, mathematical, and language skills of non-treated children would have improved if the children had experienced the structured curriculum during the last year of preschool. The follow-up treatment effect measures the extent to which EF, mathematical, and language skills are different in primary school for treated and non-treated children.

Table 2. Self-Reinforcement and Cross-Production of Skills

	Mathematics		Language		EFs	
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Developmental Stage 1 (2016 - 2017)						
EFs	0.393**	0.363**	0.148*	0.167*	0.785**	0.774**
	(0.061)	(0.052)	(0.061)	(0.065)	(0.056)	(0.045)
Mathematics	0.691**	0.663**	0.292*	0.293*	0.348**	0.339**
	(0.034)	(0.039)	(0.059)	(0.065)	(0.056)	(0.060)
Language	0.074	0.090*	0.279**	0.264**	0.115**	0.113*
	(0.040)	(0.042)	(0.059)	(0.061)	(0.044)	(0.046)
Panel B: Developmental Stage 2 (2017 - 2018)						
EFs	0.390**	0.424**	0.355*	0.304*	0.966**	0.943**
	(0.054)	(0.052)	(0.059)	(0.057)	(0.066)	(0.065)
Mathematics	0.668**	0.635**	0.049	0.081	0.247**	0.261**
	(0.052)	(0.057)	(0.070)	(0.065)	(0.052)	(0.051)
Language	0.021	0.043	0.243**	0.215**	–0.018	–0.029
	(0.040)	(0.043)	(0.055)	(0.050)	(0.045)	(0.041)
Control Variables		✓		✓		✓

Notes. This table reports the self-productivity parameter estimates. The columns denote the dependent variables, and the rows denote the independent variables. The table reports estimates from twelve models. The control variables include: child sex, birth month, whether or not at least one of the parents is a non-Western immigrant, parental education, family income, an indicator for late parental consent, and randomization-block indicators. Bootstrapped (clustered) Standard errors in parentheses (1,000 repetitions). All models are estimated with full-information maximum likelihood (701 obs.). Appendix A.1 provides an overview of missingness.
* $p < 0.05$ and ** $p < 0.01$ (two-tailed).

Table 3, Panel A, reports the parameter estimates and standard errors associated with the post-intervention and follow-up treatment effects. The post-intervention treatment effect answers the question of how much more the EF, mathematical, and language skills of non-treated children would have improved if the children had experienced the structured curriculum during the last year of preschool. The follow-up treatment effect measures the extent to which EF, mathematical, and language skills are different in primary school for treated and non-treated children.

The results in Panel A suggest the following: Consistent with Rege et al. (2021), children develop more EF skills because they have experienced the structured curriculum. Treated children have about a 0.176 standard-deviation-higher level of EF skills than non-treated children (statistically significant at the one percent level). Furthermore, there is suggestive evidence that the structured curriculum boosted the mathematical skills of treated children by the end of preschool. However, we cannot rule out the possibility that this estimate occurred by chance at the conventional cut-off of five percent.

Panel A further shows that language skills and EF skills differ in favor of treated children in primary school, but these differences are imprecisely estimated, so we cannot assume a difference at the conventional cut-off of five percent. Put differently, while the follow-up treatment effect on EF and language skills are not statistically significant, treated children do have higher levels of EF and language skills in our sample. We do observe positive and statistically significant differences in favor of treated children for mathematical skills in primary school, however. The follow-up treatment effect on mathematical skills implies that treated children have about a 0.198 standard deviation higher level of mathematical skill because they experienced the structured curriculum in preschool (statistically significant at the five percent level). Appendix A.3 presents results that do not account for measurement error (except through simple averaging). We find point estimates (and levels of precision) of the post-intervention and follow-up treatment effects even closer to those reported in Rege et al. (2021).

Table 3, Panel B, decomposes the follow-up treatment effect into measured and unmeasured variables. In square brackets, we report the relative contributions of these measured and unmeasured variables. We calculate

Table 3. Treatment Effects and Decomposition

	Mathematics		Language		EFs	
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Post-Intervention and Follow-Up Treatment Effects						
Post-Intervention	0.183	0.173	0.026	0.042	0.168**	0.176**
	(0.099)	(0.098)	(0.098)	(0.104)	(0.063)	(0.063)
Follow-Up	0.205*	0.198*	0.006	0.042	0.050	0.069
	(0.100)	(0.093)	(0.111)	(0.112)	(0.088)	(0.087)
Panel B: Decomposition of Follow-Up Treatment Effect						
EFs	0.043*	0.050*	0.041*	0.037*	0.099*	0.105*
	(0.020)	(0.021)	(0.018)	(0.015)	(0.049)	(0.047)
	[21.1%]	[25.3%]	[48.2%]	[78.7%]	[57.9%]	[64.4%]
Mathematics	0.031	0.028	0.002	0.004	0.012	0.011
	(0.018)	(0.017)	(0.002)	(0.002)	(0.001)	(0.007)
	[15.2%]	[14.1%]	[2.4%]	[8.5%]	[7.0%]	[6.7%]
Control Variables		✓		✓		✓

Continued on next page

Table 3. *Continued from previous page*

	Mathematics		Language		EFs	
	(1)	(2)	(3)	(4)	(5)	(6)
Panel B: Decomposition of Follow-Up Treatment Effect						
	0.000	0.001	0.002	0.004	−0.000	−0.000
Language	(0.001)	(0.002)	(0.006)	(0.009)	(0.001)	(0.002)
	[0.0%]	[0.5%]	[2.4%]	[8.5%]	[0.0%]	[0.0%]
	0.130	0.119	−0.040	−0.002	−0.060	−0.047
Unmeasured Variables	(0.096)	(0.089)	(0.109)	(0.110)	(0.077)	(0.076)
	[63.7%]	[60.1%]	[47.1%]	[4.3%]	[35.1%]	[28.8%]
Control Variables		✓		✓		✓

Notes. This table reports the parameter estimates for the treatment effect decomposition. Panel A reports the post-intervention treatment effect (i.e., differences between treated and non-treated children post-intervention) and the follow-up treatment effect (i.e., differences between treated and non-treated children at the follow-up). Panel B decomposes the follow-up treatment effect in measured and unmeasured variables. The columns denote the dependent variables, and the rows denote the independent variables. We include the following control variables: child sex, birth month, whether or not at least one of the parents is a non-Western immigrant, parental education, family income, an indicator for late parental consent, and randomization-block indicators. We standardize the parameter estimates using the "true" variance. Standard errors (in parentheses) are computed using a wild residual (clustered) bootstrap procedure (1,000 bootstrap samples). In brackets, we report the relative contributions. We compute these relative contributions by taking the absolute value of the estimate divided by the sum of absolute values of all estimates multiplied by 100. All models are estimated with full-information maximum likelihood (total observations 701). Web Appendix A provides an overview of missingness.

* $p < 0.05$, and ** $p < 0.01$ (two-tailed).

these relative contributions using absolute values; negative values would otherwise cancel out positive values. For each skill, the control variables included in each second column (in Table 3) are similar to those included in Table 2.

It may appear counter-intuitive to decompose the follow-up treatment effect in EF skills and language skills. While there are positive differences in favor of treated children, the effects are statistically not significant (at any conventional cut-off). Note, however, that the follow-up treatment effect captures total differences between treated and nontreated children. There could be unmeasured variables that favor non-treated children, thereby reducing the estimate of the total difference. Alternatively, the total difference might be noisier (i.e., less precisely estimated) because its standard error is based on a linear combination of parameter estimates, which is generally more imprecise than of any single parameter estimate. Consequently, while the follow-up treatment effect on EF and language skills are not statistically significant, it is still worthwhile to estimate the extent to which preschool program-induced changes in EF skills contribute to the follow-up differences in language and EF skills.

The findings in Columns (1) through (4) are consistent with the second implication of self-productivity (Equation 2); if a skill is cross-productive, then (effective) investments will produce synergistic effects (all else being equal). Treated children received an effective ECEC-type of investment in preschool. Because EF skills were cross-productive, children exposed to the investment developed more mathematical and language skills in primary school. Specifically, because treated children started primary school with higher EF skills, they treated children developed about 0.050 of a standard deviation more mathematical skills and about 0.037 of a standard deviation more language skills (both statistically significant at the five percent level). While the total follow-up treatment effect in language skills was not statistically significant, the extent to which preschool program-induced changes in EF skills contributed to these follow-up differences is. Taken together, these findings illustrate the fundamental role of EF skills in learning, which is consistent with the claim that these skills are beneficial for school success (Diamond & Lee 2011). In other words, children seem to develop more mathematical and language skills because they started primary school with improved EF skills.

Based on the follow-up treatment effect of EF skills (Columns 5 and 6), it appears that the impact of the structured curriculum on these skills fades out, as differences between treated and non-treated children are (statistically) indistinguishable. However, Panel B reveals that such a conclusion would be inaccurate. Children who started primary school with superior EF skills – because they experienced the structured curriculum – showed improvement in their EF skills. Treated children developed 0.105 of a standard deviation more EF skills in primary school because they started school with better EF skills (statistically significant at the five percent level). This finding is consistent with the high auto-regressive parameter estimates reported in Table 2 and the stable rank order of children this implies. Also, this finding is consistent with an implication of self-productivity (Equation 2); if skills are self-productive, then investments will not fully depreciate over a given length of time (all else being equal).

The observation that the follow-up treatment effect in Table 3 is smaller than the effect of program-induced changes in EF skills on those skills indicates that non-treated children catch up. The unmeasured variables capture this catch-up mechanism. We do not know precisely what these unmeasured variables are, however. These unmeasured variables may include unmeasured skills, but they might also capture interactions between measured skills (which would vary by treatment assignment) and investments made in primary school (which would not vary by treatment assignment). These interactions may provide a possible explanation for the catch-up. In particular, it could be that investments in primary school (*e.g.*, the structure provided by the teacher through rules and expectations) are compensatory to low levels of EF skills, as is commonly hypothesized (see, *e.g.*, Bierman et al. 2008, Raver et al. 2011, Riggs et al. 2006). These investments may have boosted EF skills for non-treated children, who had lower levels of these skills at the start of primary school, but not the EF skills of treated children, who had already mastered them by the start of primary school. While plausible, such catch-up does not imply that the intervention was not effective. The primary school environment may substitute for lower levels of EF skills, but Table 3 showed that treated children who started school with higher levels of EF skills developed more mathematical and language skills. The non-treated children did not receive this boost because they started school with lower levels of EF skills.

6.2. Accounting for Child-Specific Unequally Spaced Intervals

This section estimates the same model as before, but accounts for child-specific unequal intervals. We use the continuous-time model parameter estimates to characterize the time course of the direct and total treatment effect based on a model with equal and child-specific unequal intervals.[12] Figure 4 shows the direct treatment effect (left-hand-side plots), the total treatment effect (center plots), and the treatment effect decomposition (right-hand-side plots) based on estimates that account for child-specific unequal intervals. The first row relates to EF skills, the second to mathematical skills, and the third to language skills. The solid line for the left-hand-side plot and center plot represents the treatment effect under equal intervals. The dash-dotted line corresponds to the situation in which we account for the child-specific unequal intervals. Note that the effect sizes under equal intervals in March 2018 differ from the point estimates in Table 3. This small discrepancy is due to Croon's correction, which also affects the intercept parameter.

The first row of plots in Figure 4 relates to EF skills. First, in the left-hand-side plot, we observe that treated children develop more advanced EF skills than non-treated children throughout the last year of preschool. These effects become more pronounced when we account for child-specific unequal intervals as indicated by the dash-dotted line. Second, in the center plot, the difference between treated and non-treated children becomes smaller as children progress through primary school. In other words, the treatment effect on EF skills is fading out. However, the observation that non-treated children catch up with treated children does not imply that the intervention is ineffective.

[12] We use the continuous-time model to compute discrete-time parameters for varying time intervals using,

$$\mathbb{E}(\mathbf{S}_{d,i}(t)) = \mathbf{B}^{-1}(\mathbf{e}^{\mathbf{B}\Delta t_{i,a}} - \mathbf{I})\tilde{\mathbf{a}}_d + \mathbf{e}^{\mathbf{B}\Delta t_{i,a}}\mathbb{E}(\mathbf{S}_{d,i}(t_0)). \tag{9}$$

The treatment effect is $\mathbb{E}(\mathbf{S}_{1,i}(t) - \mathbf{S}_{0,i}(t))$. While $\mathbb{E}(\mathbf{S}_{d,i}(t_0)) = 0$ in developmental stage 1, it is the outcome of developmental stage 1 in developmental stage 2. Note that the parameter matrix $\mathbf{B}$ also affects the vector with intercept parameters $\tilde{\mathbf{a}}_d$. Consequently, there is a small discrepancy between the last month of developmental stage 1 (June) and the first month of developmental stage 2. The decomposition is based on relative contributions and calculated based on absolute values in the case of negative contributions.

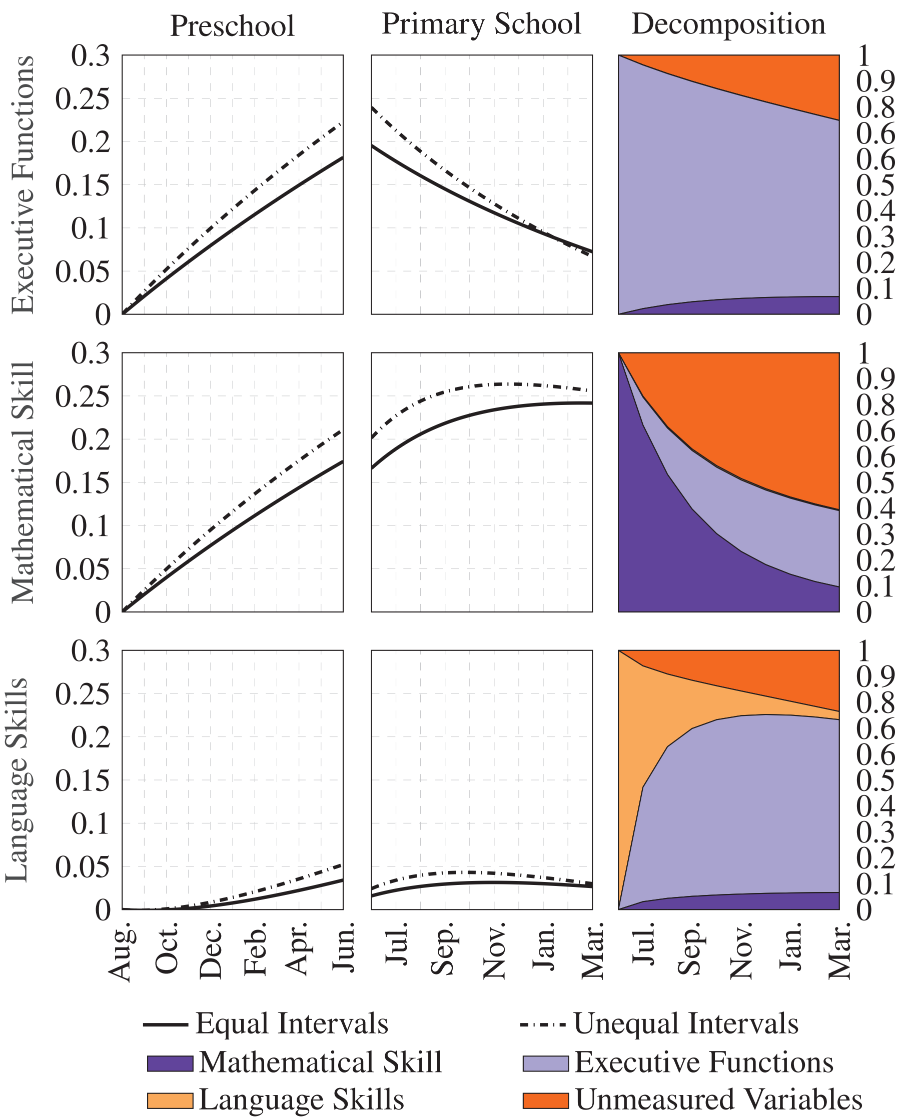

Figure 4. Dynamic Treatment Effect Decomposition

Notes. This figure shows a dynamic treatment effect decomposition. The first row relates to children's EF skills, the second to mathematical skills, and the third to language skills. The first column depicts the direct treatment effect dynamically in the last year of preschool. The second column depicts the follow-up treatment effect dynamically in the first grade of primary school. The third column depicts the decomposition of the follow-up treatment effect based on the estimates for child-specific unequal intervals.

Indeed, as Table 3 illustrated, treated children gained more mathematical and language skills because they started primary school with better EF skills. Third, in the right-hand-side plot, we decompose the total treatment effect, accounting for child-specific unequal intervals, into the relative contributions of measured and unmeasured variables. In this plot, we observe that (1) mathematical and language skills contribute relatively little to differences in EF skills between treated and non-treated children, and (2) the relative contribution of unmeasured variables becomes larger in the development of EF skills. As Table 3 illustrates, the relative contribution is positive for non-treated children, suggesting that these children experience something that we did not measure that turns out to be beneficial for developing EF skills. However, these unmeasured variables are imprecisely estimated (see Table 3).

The second row in Figure 4 relates to children's mathematical skills. First, in the left-hand-side plot, we observe that treated children develop more mathematical skills than non-treated children throughout the last year of preschool. However, we could not rule out that these differences occurred by chance at the conventional cut-off of five percent. As in the case of EF skills, we find that accounting for child-specific unequal intervals, indicated by the dash-dotted line, results in higher estimated treatment effects. Second, in the center plot, the differences between treated and non-treated children continue to increase but at a decreasing rate, which suggests that the treatment effect on mathematical skill is leveling off. Third, in the right-hand-side plot, we observe that (1) the relative contribution of children's mathematical skills decreases throughout the school year, (2) the relative contribution of EF skills increases, (3) the relative contribution of language skills is trivial, and (4) the relative contribution of unmeasured variables increases. As Table 3 illustrates, the relative contribution of unmeasured variables is positive for treated children. Such a finding might imply that the Agder Project improved skills we did not measure that turned out to promote the acquisition of mathematical skills in primary school. However, these unmeasured variables are imprecisely estimated (see Table 3), so this implication is not warranted.

The last row of plots in Figure 4 relates to language skills. The left-hand-side plot and the center plot suggest that treated children developed more language skills because of the Agder Project's structured curriculum. However, as Table 3 shows, these effects are (statistically) indistinguishable. The decomposition in the right-hand-side plot suggests that the relative contribution of EF skills in acquiring language skills increases as children progress through school. The relative contribution of unmeasured variables increases as well. Since the relative contribution is positive for non-treated children, this finding suggests that the Agder Project improved skills we did not measure that helped children acquire more language skills. However, the effect of unmeasured variables is imprecisely estimated.

7. Conclusion

We investigated the cross-productivity of EF skills. We combined the experimental variation with an econometric model of skill formation to estimate the extent to which program-induced improvements in EF skills caused children to acquire more mathematical and language skills. We found that children did in fact acquire more mathematical and language skills in primary school because they started primary school with higher levels of EF skills. Curiously, while the total follow-up difference (in favor of treated children) in language skills was not statistically significant, the extent to which preschool program-induced changes in EF skills contributed to the follow-up differences in language skills was significant. In addition, while total differences in EF skills between treated and non-treated children were (statistically) insignificant, treated children continued to benefit from the curriculum as they acquired higher levels of EF skills in primary school. These findings hint at a dynamic complementarity between the Agder Project's preschool curriculum and the investments made in primary school. Indeed, primary school investments seem to be more successful at promoting mathematical and language skills for children who experienced the structured curriculum.

Finally, accounting for child-specific unequal intervals had a relatively small impact on post-intervention and follow-up treatment effect estimates. The small impact might be due to the data collection in the Agder Project and the low number of time points for each stage of development. Nevertheless, by augmenting the framework in Heckman et al. (2013), we formulated a framework to unify the interpretation of treatment effects across different studies.

Heckman et al. (2013, p. 2079) note that their econometric framework can "*[…] unify the interpretation of the treatment effects across different studies with different interventions applied to different populations.*" Our research indicates that this claim relies on observing the assumed unit interval in the data. When this is not done, and researchers do not account for child-specific unequal spacing of observations, the proposed framework cannot unify the interpretation of treatment effects across different studies unless the length of child-specific observation intervals is the same. By augmenting their framework, we contributed to developing a framework that can unify the interpretation of treatment effects even when observation intervals are unequal and child specific.

References

Akaike, H. (1987), 'Factor analysis and AIC', *Psychometrika* **52**(3), 317–332.

Anderson, T. W. (1957), 'Maximum likelihood estimates for a multivariate normal distribution when some observations are missing', *Journal of the American Statistical Association* **52**(278), 200–203.

Anderson, T. W. & Rubin, H. (1956), *Statistical Inference in Factor Analysis*, Vol. 5 of *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 5: Contributions to Econometrics, Industrial Research, and Psychometry*, University of California Press, Berkeley, California.

Arnold, L. (1974), *Stochastic Differential Equations: Theory and Applications*, John Wiley & Sons, Inc., New York City, New York.

Attanasio, O., Cattan, S., Fitzsimons, E., Meghir, C. & Rubio-Codina, M. (2020), 'Estimating the production function for human capital: Results from a randomized controlled trial in Colombia', *American Economic Review* **110**(1), 48–85.

Baltagi, B. H. & Wu, P. X. (1999), 'Unequally spaced panel data regressions with AR(1) disturbances', *Econometric Theory* **15**(6), 814–823.

Bartlett, M. S. (1937), 'The statistical conception of mental factors', *British Journal of Psychology* **28**(1), 97–104.

Berger, E. M., Fehr, E., Hermes, H., Schunk, D. & Winkel, K. (2020), The impact of working memory training on children's cognitive and noncognitive skills, Working Paper 13338, Institute of Labor Economics, Bonn, Germany.

Bergstrom, A. R. (1988), 'The history of continuous-time econometric models', *Econometric Theory* **4**(3), 365–383.

Best, J. R. & Miller, P. H. (2010), 'A developmental perspective on executive function', *Child Development* **81**(6), 1641–1660.

Bierman, K. L., Nix, R. L., Greenberg, M. T., Blair, C. & Domitrovich, C. E. (2008), 'Executive functions and school readiness intervention: Impact, moderation, and mediation in the head start redi program', *Development and Psychopathology* **20**(3), 821–843.

Blair, C. (2002), 'School readiness: Integrating cognition and emotion in a neurobiological conceptualization of children's functioning at school entry', *American Psychologist* **57**(2), 111–127.

Blair, C. (2006), 'How similar are fluid cognition and general intelligence? a developmental neuroscience perspective on fluid cognition as an aspect of human cognitive ability', *Behavioral and Brain Sciences* **29**(2), 109–160.

Blair, C. & Raver, C. C. (2015), 'School readiness and self-regulation: A developmental psychobiological approach', *Annual Review of Psychology* **66**, 711–731.

Blair, C. & Razza, R. P. (2007), 'Relating effortful control, executive function, and false belief understanding to emerging math and literacy ability in kindergarten', *Child Development* **78**(2), 647–663.

Blau, D. & Currie, J. (2006), Pre-school, day care, and after-school care: Who's minding the kids?, *in* E. A. Hanushek & F. Welch, eds, 'Handbook of the Economics of Education', Vol. 2, Elsevier B.V., Amsterdam, Netherlands, Chapter 20, pp. 1163–1278.

Borghans, L., Duckworth, A. L., Heckman, J. J. & ter Weel, B. (2008), 'The economics and psychology of personality traits', *Journal of Human Resources* **43**(4), 972–1059.

Broidy, L. M., Nagin, D. S., Tremblay, R. E., Bates, J. E., Brame, B., Dodge, K. A., Fergusson, D., Horwood, J. L., Loeber, R., Laird, R., Lynam, D. R., Moffitt, T. E., Pettit, G. S. & Vitaro, F. (2003), 'Developmental trajectories of childhood disruptive behaviors and adolescent delinquency: A six-site, cross-national study', *Developmental Psychology* **39**(2), 222–245.

Campbell, F., Conti, G., Heckman, J. J., Moon, S. H., Pinto, R., Pungello, E. & Pan, Y. (2014), ‘Early childhood investments substantially boost adult health’, *Science* **343**(6178), 1478–1485.

Cascio, E. U. & Lewis, E. G. (2006), ‘Schooling and the armed forces qualifying test: Evidence from school-entry laws’, *Journal of Human Resources* **41**(2), 294–318.

Chor, E., Andresen, M. E. & Kalil, A. (2016), ‘The impact of universal prekindergarten on family behavior and child outcomes’, *Economics of Education Review* **55**, 168–181.

Clarke, D., Romano, J. P. & Wolf, M. (2020), ‘The romano-wolf multiple-hypothesis correction in stata’, *Stata Journal* **20**(4), 812–843.

Clements, D. H. & Sarama, J. (2011), ‘Early childhood mathematics intervention’, *Science* **333**(6045), 968–970.

Conti, G., Heckman, J. J. & Pinto, R. (2016), ‘The effects of two influential early childhood interventions on health and healthy behaviour’, *Economic Journal* **126**(596), F28–F65.

Crawford, C., Dearden, L. & Meghir, C. (2010), When you are born matters: The impact of date of birth on educational outcomes in England, Working Paper 10,06, Institute for Fiscal Studies, London, England.

Croon, M. (2002), *Using Predicted Latent Scores in General Latent Structure Models*, Lawrence Erlbaum Associates, Inc., Mahwah, New Jersey, book section 10.

Cunha, F. & Heckman, J. (2007), ‘The technology of skill formation’, *American Economic Review* **97**(2), 31–47.

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

Cunha, F., Nielsen, E. & Williams, B. (2021), 'The econometrics of early childhood human capital and investments', *Annual Review of Economics* **13**, 487–513.

Currie, J. (2001), 'Early childhood education programs', *Journal of Economic Perspectives* **15**(2), 213–238.

Davidson, M. C., Amso, D., Anderson, L. C. & Diamond, A. (2006), 'Development of cognitive control and executive functions from 4 to 13 years: Evidence from manipulations of memory, inhibition, and task switching', *Neuropsychologia* **44**(11), 2037–2078.

Del Boca, D., Flinn, C. & Wiswall, M. (2014), 'Household choices and child development', *Review of Economic Studies* **81**(1), 137–185.

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

Diamond, A. (2016), *Why Improving and Assessing Executive Functions Early in Life is Critical*, American Psychological Association, Washington, D.C., pp. 11–43.

Diamond, A. & Lee, K. (2011), 'Interventions shown to aid executive function development in children 4 to 12 years old', *Science* **333**(6045), 959–964.

Diamond, A. & Ling, D. S. (2020), *Review of the Evidence on, and Fundamental Questions About, Efforts to Improve Executive Functions, Including Working Memory*, Oxford University Press, New York City, New York, book section 8, pp. 143–431.

Dillon, M. R., Kannan, H., Dean, J. T., Spelke, E. S. & Duflo, E. (2017), 'Cognitive science in the field: A preschool intervention durably enhances intuitive but not formal mathematics', *Science* **357**(6346), 47–55.

Duncan, G. J., Dowsett, C. J., Claessens, A., Magnuson, K., Huston, A. C., Klebanov, P., Pagani, L. S., Feinstein, L., Engel, M., Brooks-Gunn, J., Sexton, H., Duckworth, K. & Japel, C. (2007), 'School readiness and later achievement', *Developmental Psychology* **43**(6), 1428–1446.

Gollob, H. F. & Reichardt, C. S. (1987), 'Taking account of time lags in causal models', *Child Development* **58**(1), 80–92.

Gormley Jr, W. T. & Gayer, T. (2005), 'Promoting school readiness in oklahoma: An evaluation of tulsa's pre-k program', *Journal of Human Resources* **40**(3), 533–558.

Hamerle, A., Nagl, W. & Singer, H. (1991), 'Problems with the estimation of stochastic differential equations using structural equations models', *The Journal of Mathematical Sociology* **16**(3), 201–220.

Hamerle, A., Singer, H. & Nagl, W. (1993), 'Identification and estimation of continuous time dynamic systems with exogenous variables using panel data', *Econometric Theory* **9**(2), 283–295.

Heckman, J. J. & Karapakula, G. (2021), 'Using a satisficing model of experimenter decision-making to guide finite-sample inference for compromised experiments', *The Econometrics Journal* **24**(2), C1–C39.

Heckman, J. J. & Mosso, S. (2014), 'The economics of human development and social mobility', *Annual Review of Economics* **6**, 689–733.

Heckman, J., Pinto, R. & Savelyev, P. (2013), 'Understanding the mechanisms through which an influential early childhood program boosted adult outcomes', *American Economic Review* **103**(6), 2052–2086.

Jenkins, J. M., Duncan, G. J., Auger, A., Bitler, M., Domina, T. & Burchinal, M. (2018), 'Boosting school readiness: Should preschool teachers target skills or the whole child?', *Economics of Education Review* **65**, 107–125.

Jones, R. H. & Boadi-Boateng, F. (1991), 'Unequally spaced longitudinal data with AR(1) serial correlation', *Biometrics* **47**(1), 161–175.

Kosse, F., Deckers, T., Pinger, P., Schildberg-Hörisch, H. & Falk, A. (2020), 'The formation of prosociality: Causal evidence on the role of social environment', *Journal of Political Economy* **128**(2), 434–467.

Kotlarski, I. (1967), 'On characterizing the gamma and the normal distribution', *Pacific Journal of Mathematics* **20**(1), 69–76.

Kraft, M. A. (2020), 'Interpreting effect sizes of education interventions', *Educational Researcher* **49**(4), 241–253.

Lenes, R., McClelland, M. M., ten Braak, D., Idsøe, T. & Størksen, I. (2020), 'Direct and indirect pathways from children's early self-regulation to academic achievement in fifth grade in norway', *Early Childhood Research Quarterly* **53**, 612–624.

Lewit, E. M. & Baker, L. S. (1995), 'School readiness', *The Future of children / Center for the Future of Children, the David and Lucile Packard Foundation* **5**(2), 128–139.

Matsumoto, K. & Tanaka, K. (2004), 'Conflict and cognitive control', *Science* **303**(5660), 969–970.

McArdle, J. J. & McDonald, R. P. (1984), 'Some algebraic properties of the reticular action model for moment structures', *British Journal of Mathematical and Statistical Psychology* **37**(2), 234–251.

McClelland, M. M., Cameron, C. E., Connor, C. M., Farris, C. L., Jewkes, A. M. & Morrison, F. J. (2007), 'Links between behavioral regulation and preschoolers' literacy, vocabulary, and math skills', *Developmental Psychology* **43**(4), 947–959.

McClelland, M. M., Cameron, C. E., Duncan, R., Bowles, R. P., Acock, A. C., Miao, A. & Pratt, M. E. (2014), 'Predictors of early growth in academic achievement: The head-toes-knees-shoulders task', *Frontiers in Psychology* **5**.

McCoy, D. C., Gonzalez, K. & Jones, S. (2019), 'Preschool self-regulation and preacademic skills as mediators of the long-term impacts of an early intervention', *Child Development* **90**(5), 1544–1558.

McKenzie, D. J. (2001), 'Estimation of AR(1) models with unequally spaced pseudo-panels', *The Econometrics Journal* **4**(1), 89–108.

Miller, H. V., Barnes, J. C. & Beaver, K. M. (2011), 'Self-control and health outcomes in a nationally representative sample', *American Journal of Health Behavior* **35**(1), 15–27.

Millimet, D. L. & McDonough, I. K. (2017), 'Dynamic panel data models with irregular spacing: With an application to early childhood development', *Journal of Applied Econometrics* **32**(4), 725–743.

Moffitt, T. E., Arseneault, L., Belsky, D., Dickson, N., Hancox, R. J., Harrington, H., Houts, R., Poulton, R., Roberts, B. W., Ross, S., Sears, M. R., Thomson, W. M. & Caspi, A. (2011), 'A gradient of childhood self-control predicts health, wealth, and public safety', *Proceedings of the National Academy of Sciences of the United States of America* **108**(7), 2693–2698.

Mol, S. E., Bus, A. G. & de Jong, M. T. (2009), 'Interactive book reading in early education: A tool to stimulate print knowledge as well as oral language', *Review of Educational Research* **79**(2), 979–1007.

Nagin, D. & Tremblay, R. E. (1999), 'Trajectories of boys' physical aggression, opposition, and hyperactivity on the path to physically violent and nonviolent juvenile delinquency', *Child Development* **70**(5), 1181–1196.

National Research Council Institute of Medicine (2000), *From neurons to neighborhoods: The science of early childhood development*, National Academy Press, Washington, D.C.

Neale, M. C., Hunter, M. D., Pritikin, J. N., Zahery, M., Brick, T. R., Kirkpatrick, R. M., Estabrook, R., Bates, T. C., Maes, H. H. & Boker, S. M. (2016), 'OpenMx 2.0: Extended structural equation and statistical modeling', *Psychometrika* **81**(2), 535–549.

OECD (2006), *Starting Strong II: Early Childhood Education and Care*, OECD Publishing, Paris, France.

Oud, J. H. L. & Jansen, R. A. R. G. (2000), 'Continuous time state space modeling of panel data by means of SEM', *Psychometrika* **65**(2), 199–215.

Raver, C. C., Jones, S. M., Li-Grining, C., Zhai, F., Bub, K. & Pressler, E. (2011), 'Csrp's impact on low-income preschoolers' preacademic skills: Self-regulation as a mediating mechanism', *Child Development* **82**(1), 362–378.

Rege, M., Størksen, I., Solli, I. F., Kalil, A., McClelland, M. M., ten Braak, D., Lenes, R., Lunde, S., Breive, S., Carlsen, M., Erfjord, I. & Hundeland, P. S. (2021), 'The effects of a structured curriculum on preschool effectiveness: A field experiment', *Journal of Human Resources* .

Riggs, N. R., Greenberg, M. T., Kusché, C. A. & Pentz, M. A. (2006), 'The mediational role of neurocognition in the behavioral outcomes of a social-emotional prevention program in elementary school students: Effects of the paths curriculum', *Prevention Science* **7**(1), 91–102.

Romano, J. P. & Wolf, M. (2005), 'Stepwise multiple testing as formalized data snooping', *Econometrica* **73**(4), 1237–1282.

Rosner, B. & Muñoz, A. (1988), 'Autoregressive modelling for the analysis of longitudinal data with unequally spaced examinations', *Statistics in Medicine* **7**(1–2), 59–71.

Rueda, M. R., Posner, M. I. & Rothbart, M. K. (2005), 'The development of executive attention: Contributions to the emergence of self-regulation', *Developmental Neuropsychology* **28**(2), 573–594.

Sasaki, Y. & Xin, Y. (2017), 'Unequal spacing in dynamic panel data: Identification and estimation', *Journal of Econometrics* **196**(2), 320–330.

Schwarz, G. (1978), 'Estimating the dimension of a model', *The Annals of Statistics* **6**(2), 461–464.

Singer, H. (1993), 'Continuous-time dynamical systems with sampled data, errors of measurement and unobserved components', *Journal of Time Series Analysis* **14**(5), 527–545.

Singer, H. (1995), 'Analytical score function for irregularly sampled continuous time stochastic processes with control variables and missing values', *Econometric Theory* **11**(4), 721–735.

Skrondal, A. & Rabe-Hesketh, S. (2004), *Generalized Latent Variable Modeling: Multilevel, Longitudinal, and Structural Equation Models*, Interdisciplinary Statistics, 1 edn, Chapman & Hall/CRC, Boca Raton, Florida.

Solli, I. F. (2017), ‘Left behind by birth month’, *Education Economics* **25**(4), 323–346.

Størksen, I., Ellingsen, I. T., Tvedt, M. S. & Idsøe, E. M. C. (2013), ‘Norsk vokabulartest (nvt) for barn i overgangen mellom barnehage og skole: Psykometrisk vurdering av en nettbrettbasert test [norwegian vocabulary test (nvt) for children in the transition between kindergarten and school: Psychometric assessment of a tablet-based test].’, *Spesialpedagogikks Forskingsdel* **4**(13), 40–54.

Størksen, I., Ertesvåg, S. K. & Rege, M. (2021), ‘Implementing implementation science in a randomized controlled trial in norwegian early childhood education and care’, *International Journal of Educational Research* **108**, 101782.

Størksen, I. & Mosvold, R. (2013), ‘Assessing early math skills with tablet computers: Development of the ani banani math test (ABMT) for young children [Utdanning2020]’.

Størksen, I., ten Braak, D., Breive, S., Lenes, R., Lunde, S., Carlsen, M. & Rege, M. (2018), *Lekbasert Læring: et Forskningsbasert Førskoleopplegg fra Agderprosjektet [Playful Learning: a Research based Preschool Curriculum from the Agder Project].*, GAN Aschehoug, Oslo, Norway.

ten Braak, D. & Størksen, I. (2021), ‘Psychometric properties of the Ani Banani Math Test’, *European Journal of Developmental Psychology* **18**(4), 610–628.

Thijssen, M. W. P. (2022), Cognitive skill production and unequally spaced intervals.

Thomson, G. H. (1938), ‘Methods of estimating mental factors’, *Nature* **141**(3562), 246.

Torgerson, W. S. (1958), *Theory and Methods of Scaling*, John Wiley & Sons, Inc., New York City, New York.

VandenBos, G. R. E. (2007), *APA Dictionary of Psychology*, APA Dictionary of Psychology, American Psychological Association, Washington, D.C.

Voelkle, M. C. & Oud, J. H. L. (2013), 'Continuous time modelling with individually varying time intervals for oscillating and non-oscillating processes', *British Journal of Mathematical and Statistical Psychology* **66**(1), 103–126.

Wald, A. (1950), *Note on the Identification of Economic Relations*, John Wiley & Sons, Inc., New York City, New York, book section 3, pp. 238–244.

Wechsler, F. (1991), *Wechsler Intelligence Scale for Children (3rd ed.) (WISC-III) Manual*, The Psychological Corporation, San Antonio, Texas.

Zahery, M., Maes, H. H. & Neale, M. C. (2017), 'CSOLNP: Numerical optimization engine for solving non-linearly constrained problems', *Twin Research and Human Genetics* **20**(4), 290–297.

Supplementary Material

Supplement to "Cross-Productivities of Executive Functions": Appendices

By

Maximiliaan W. P. Thijssen
Mari Rege
Ingeborg F. Solli
Ingunn Størksen

A.1. Further Notes on the Data

Table A.1.1 and Table A.1.2 present an overview of missingness. Table A.1.1 shows the missingness in Statistics Norway's data, and Table A.1.2 reports the missing values in the Agder Project data. To compute the missing values for the assessment data, we first estimate the measurement models for each test. Then, we predict the factor scores. We changed the predicted factor scores to missing if any of the individual items were missing. We then calculate, based on these predicted factor scores, the number of observed and missing values. This procedure implies that a single missing item will result in the deletion of the whole row. However, rarely are only a few items missing. In most cases, either all individual items are missing or none. When we construct the variable that measures a child's age at the time of assessments, we imputed some values. We missed the assessment date for a few children: 32 children in 2016, 49 in 2017, and 40 in 2018. We replace these missing values with the median of the respective preschool center. We use full information (parametric) maximum likelihood estimation (Anderson 1957). Even when data are only missing at random, full-information maximum likelihood uses all available data and provides valid point estimates.

Table A.1.1. Overview of Missing Values: Family Characteristics

		Observed		Missing	
	(1) Period	(2) Obs.	(3) Pct.	(4) Obs.	(5) Pct.
Education Father	2016	666	95.0	35	5.0
Education Mother	2016	676	96.4	25	3.6
Non-Western Immigrant	2016	682	97.3	19	2.7
Income Father	2016	683	97.4	18	2.6
Income Mother	2016	698	99.6	3	0.4
Birth Month	2016	701	100.0	0	0.0
The Child is Female	2016	701	100.0	0	0.0
Late Parental Consent	2016	701	100.0	0	0.0

Notes. This table reports the descriptive frequencies of observed and missing observations for the registry data from Statistics Norway and the child and parental characteristics collected as part of the Agder Project's data collection.

Table A.1.2. Overview of Missing Values: Agder Project Data

		Observed		Missing	
	(1) Period	(2) Obs.	(3) Pct.	(4) Obs.	(5) Pct.
Head-Toes-Knees-Shoulders Test	2016	516	73.6	185	26.4
Hearts and Flowers Test	2017	635	90.6	66	9.4
Digit Span Test	2017	641	91.4	60	8.6
Hearts and Flowers Test	2016	642	91.6	59	8.4
Head-Toes-Knees-Shoulders Test	2017	645	92.0	56	8.0
Blending Test	2017	645	92.0	55	8.0
Ani Banani Math Test	2016	646	92.3	55	7.8
Norwegian Vocabulary Test	2016	647	92.3	54	7.7
Blending Test	2016	648	92.4	53	7.6
Digit Span Test	2016	648	92.4	53	7.6
Digit Span Test	2018	653	93.2	48	6.8
Blending Test	2018	658	93.9	43	6.1
Head-Toes-Knees-Shoulders Test	2016	659	94.0	42	6.0
Norwegian Vocabulary Test	2018	659	94.0	42	6.0
Hearts and Flowers Test	2018	660	94.2	41	5.8
Ani Banani Math Test	2018	661	94.3	40	5.7
Ani Banani Math Test	2016	663	94.6	38	5.4

Notes. This table reports the descriptive frequencies related to observed and missing observations for the assessment data from the Agder Project. For the Head-Toes-Knees-Shoulders Test in 2016, some children did not complete the last ten items (10 out of a total of 30). The reason for the occurrence of these missing values is that the test stops after a child misses a certain number of items. At this point, it is unlikely that the child will complete the later items. For this reason, we replace these ten items with zeros. There is some minor variation in missingness across test items.

We regress premeasured and predetermined variables on a treatment indicator and randomization-block fixed effects to test differences in baseline characteristics. We cluster the standard errors at the randomization-block level. Table A.1.3 reports the results. All baseline characteristics are balanced, including baseline skills of children. We also allow for multiple hypothesis testing using the Romano & Wolf (2005) procedure implemented in the Stata package "`rwolf`" (Clarke et al. 2020). None of the differences are significant.

Table A.1.3. Balance Test

	Treatment Group		Control Group		Differences		
	(1) Obs.	(2) Mean	(3) Obs.	(4) Mean	(5) Mean	(6) *p*-val.	(7) RW-*p*
Panel A: Predetermined Variables							
The Child is Female	383	0.48	318	0.52	−0.03	0.44	0.85
Birth Month	383	6.15	318	6.19	−0.02	0.93	0.99
Education Mother	365	4.71	311	4.86	−0.12	0.38	0.85
Education Father	362	4.46	304	4.50	−0.05	0.76	0.99
Non-Western Immigrant	371	0.18	311	0.12	0.05	0.21	0.63
Income Mother (in NOKs)	381	321,391	317	341,640	−20, 495	0.38	0.85
Income Father (in NOKs)	372	560,080	311	549,035	11, 048	0.58	0.92
Panel B: Premeasured Variables							
Blending Test	383	2.37	318	2.63	−0.29	0.36	0.85
Head-Toes-Knees-Shoulders Test	383	21.52	318	19.98	1.47	0.40	0.85
Norwegian Vocabulary Test	383	9.93	318	9.62	0.30	0.48	0.85
Hearts and Flowers Test	383	26.34	318	25.28	1.08	0.35	0.85
Ani Banani Math Test	368	3.14	318	3.12	−0.00	0.98	0.99
Digit Span Test	359	5.39	289	5.46	−0.06	0.76	0.99

Notes. This table reports the results of a balance test. Column (5) reports the mean of the difference between the control group. Column (6) reports p-values (*p*-val.) for without correction for multiple hypothesis testing. Column (7) reports *p*-values with Romano-Wolf (RW-*p*) correction for multiple hypothesis testing using Clarke et al. (2020). See also Romano & Wolf (2005). Standard errors are clustered at the randomization-block level.
* $p < 0.05$, and ** $p < 0.01$ (two-tailed).

A.2. Further Notes on the Estimation Procedure

As an illustrative example, we show how to account for unequally spaced intervals in the second stage of development (June 2017 through March 2018). Let P denote the number of control variables. We can write Equation (5) for the unit interval, $\Delta t = 1$, as,

$$\mathbf{S}_{d,i,t} = \tilde{\mathbf{a}}_{d,\Delta t} + \mathbf{B}_{\Delta t}\mathbf{S}_{d,i,t-\Delta t} + \mathbf{C}_{\Delta t}\mathbf{X}_{d,i} + \widetilde{\mathbf{w}}_{d,i,t}. \tag{A.2.1}$$

where $\widetilde{\mathbf{w}}_{d,i,t} \sim \mathrm{N}(\mathbf{0}, \mathbf{Q}_t)$. In Equation (A.2.1), we have the following matrices of unknowns for $H = 3$. First, the intercept parameters,

$$\tilde{\mathbf{a}}_{d,\Delta t} = \begin{bmatrix} \tilde{a}_{1,d,\Delta t} \\ \tilde{a}_{2,d,\Delta t} \\ \tilde{a}_{3,d,\Delta t} \end{bmatrix}.$$

Second, the self-productivity parameters,

$$\mathbf{B}_{\Delta t} = \begin{bmatrix} b_{1,1,\Delta t} & b_{1,2,\Delta t} & b_{1,3,\Delta t} \\ b_{2,1,\Delta t} & b_{2,2,\Delta t} & b_{2,3,\Delta t} \\ b_{3,1,\Delta t} & b_{3,1,\Delta t} & b_{3,3,\Delta t} \end{bmatrix}.$$

Third, the parameters that measure the relationship between child and parental characteristics and skills,

$$\mathbf{C}_{\Delta t} = \begin{bmatrix} c_{1,1,\Delta t} & \cdots & c_{1,P,\Delta t} \\ c_{2,1,\Delta t} & \cdots & c_{2,P,\Delta t} \\ c_{3,1,\Delta t} & \cdots & c_{3,P,\Delta t} \end{bmatrix}.$$

Lastly, the variance-covariance matrix of the random shocks,

$$\mathbf{Q}_t = \begin{bmatrix} \mathrm{Var}(\widetilde{\mathrm{w}}_{1,d,t}) & \mathrm{Cov}(\widetilde{\mathrm{w}}_{1,d,t}, \widetilde{\mathrm{w}}_{2,d,t}) & \mathrm{Cov}(\widetilde{\mathrm{w}}_{1,d,t}, \widetilde{\mathrm{w}}_{3,d,t}) \\ \mathrm{Cov}(\widetilde{\mathrm{w}}_{1,d,t}, \widetilde{\mathrm{w}}_{2,d,t}) & \mathrm{Var}(\widetilde{\mathrm{w}}_{2,d,t}) & \mathrm{Cov}(\widetilde{\mathrm{w}}_{2,d,t}, \widetilde{\mathrm{w}}_{3,d,t}) \\ \mathrm{Cov}(\widetilde{\mathrm{w}}_{1,d,t}, \widetilde{\mathrm{w}}_{3,d,t}) & \mathrm{Cov}(\widetilde{\mathrm{w}}_{2,d,t}, \widetilde{\mathrm{w}}_{3,d,t}) & \mathrm{Var}(\widetilde{\mathrm{w}}_{3,d,t}) \end{bmatrix}.$$

Next, we define a filter matrix, $\mathscr{F}$. The filter matrix "filters" observed and unobserved variables. We estimate the measurement models separately, so there are no unobserved variables. As a result, the filter matrix is an identity matrix. The dimensions of this identity matrix equal the total number of variables. There are P control variables and $H = 3$ skills, so the filter matrix, $\mathscr{F}$, is $(P+3, P+3)$-dimensional. Second, define a $(P+3, P+3)$-dimensional *asymmetric* coefficient matrix, $\mathscr{A}$,

$$\underbrace{\mathscr{A}}_{(P+3,P+3)} = \begin{bmatrix} \underbrace{\mathbf{0}}_{(3,3)} & \underbrace{\mathbf{0}}_{(3,3)} & \underbrace{\mathbf{C}_0}_{(3,P)} \\ \underbrace{\mathbf{B}_1}_{(3,3)} & \underbrace{\mathbf{0}}_{(3,3)} & \underbrace{\mathbf{C}_1}_{(3,P)} \\ \underbrace{\mathbf{0}}_{(P,3)} & \underbrace{\mathbf{0}}_{(P,3)} & \underbrace{\mathbf{0}}_{(P,P)} \end{bmatrix}.$$

Lastly, we define a $(P+3, P+3)$-dimensional *symmetric* coefficient matrix, $\mathscr{S}$,

$$\underbrace{\mathscr{S}}_{(P+3,P+3)} = \begin{bmatrix} \underbrace{\mathbf{Q}_0}_{(3,3)} & \underbrace{\mathbf{0}}_{(3,3)} & \underbrace{\mathbf{0}}_{(3,P)} \\ \underbrace{\mathbf{0}}_{(3,3)} & \underbrace{\mathbf{Q}_1}_{(3,3)} & \underbrace{\mathbf{0}}_{(3,P)} \\ \underbrace{\mathbf{0}}_{(P,3)} & \underbrace{\mathbf{0}}_{(P,3)} & \underbrace{\boldsymbol{\sigma}_{\mathbf{X}}}_{(P,P)} \end{bmatrix},$$

where $\boldsymbol{\sigma}_{\mathbf{X}}$ denotes the variance-covariance matrix of control variables. The filter and coefficient matrices together represent the Reticular Action Model (McArdle & McDonald 1984). The model-implied variance-covariance matrix can be written as a function of these three matrices,

$$\boldsymbol{\Sigma}_{\mathrm{M}} = \mathscr{F}(\mathbf{I} - \mathscr{A})^{-1}\mathscr{S}(\mathbf{I} - \mathscr{A})^{-1\prime}\mathscr{F}', \tag{A.2.2}$$

and the model-implied mean structure as,

$$\boldsymbol{\mu}_{\mathrm{M}} = \mathscr{F}(\mathbf{I} - \mathscr{A})^{-1}\mathscr{M}, \tag{A.2.3}$$

where $\mathscr{M}$ is a vector with mean components (*e.g.*, $\tilde{\mathbf{a}}_{d,\Delta t}$).

With the model defined, the next step involves deriving the nonlinear constraints that link the discrete-time model parameters *exactly* to the continuous-time model parameters. We can show that Equation (A.2.1) belongs to the following stochastic differential equation,

$$\frac{\mathrm{d}\mathbf{S}_{d,i}(t)}{\mathrm{d}t} = \tilde{\mathbf{a}}_{d,i} + \mathbf{B}\mathbf{S}_{d,i}(t) + \mathbf{C}\mathbf{X}_{d,i} + \mathbf{G}\frac{\mathrm{d}\widetilde{\mathbf{w}}_{d,i}(t)}{\mathrm{d}t}, \tag{A.2.4}$$

where the parameters are no longer dependent on the observation interval Δt. We assume that the continuous-time error term, $\widetilde{\mathbf{w}}_{d,i}(t)$, follows a Wiener process. A property of this process is independent and normally distributed increments, $\Delta\widetilde{\mathbf{w}}_{d,i}(t) \equiv \widetilde{\mathbf{w}}_{d,i}(t) - \widetilde{\mathbf{w}}_{d,i}(t - \Delta t)$, with mean zero and covariance matrix $\Delta t\mathbf{I}$ (Arnold 1974, p. 46). The matrix $\mathbf{G}$ allows the continuous-time error process variance to be lower or higher than one. The variance-covariance matrix, $\mathbf{GG}'$, is the "diffusion matrix" associated with the stochastic process. Arnold (1974, pp. 45–56) explains that one cannot interpret the integral of the Wiener process as an ordinary Riemann-Stieltjes integral due to unbounded variation (see, *e.g.*, Hamerle et al. 1991).

One can interpret the integral alternatively, which gives the following solution to the stochastic differential equation (Arnold 1974, pp. 128–135),

$$\mathbf{S}_{d,i}(t) = \mathbf{e}^{\mathbf{B}(t-t_0)}\mathbf{S}_{d,i}(t_0) + \mathbf{B}^{-1}(\mathbf{e}^{\mathbf{B}(t-t_0)} - \mathbf{I})(\tilde{\mathbf{a}}_d + \mathbf{CX}_{d,i}) + \int_{t_0}^{t} \mathbf{e}^{\mathbf{B}(t-q)}\mathbf{G}\mathrm{d}\,\widetilde{\mathbf{w}}_{d,i}(q), \tag{A.2.5}$$

for initial value $\mathbf{S}_{d,i}(t_0)$ and observation interval $t - t_0$, where $\mathbf{e}^{\{\cdot\}}$ denotes the matrix exponential. Choosing for $t - t_0$ the (child-specific) unequal observation interval $\Delta t_{i,a} \equiv t_{i,a} - t_{i,a-1}$, we observe that Equation (A.2.1) equates to Equation (A.2.5) under the following constraints,

$$\mathbf{B}_{\Delta t_{i,a}} = \mathbf{e}^{\mathbf{B}\Delta t_{i,a}}, \tag{A.2.6}$$

$$\tilde{\mathbf{a}}_{d,\Delta t_{i,a}} = \mathbf{B}^{-1}(\mathbf{e}^{\mathbf{B}\Delta t_{i,a}} - \mathbf{I})\tilde{\mathbf{a}}_d, \tag{A.2.7}$$

$$\mathbf{C}_{\Delta t_{i,a}} = \mathbf{B}^{-1}(\mathbf{e}^{\mathbf{B}\Delta t_{i,a}} - \mathbf{I})\mathbf{C}_d, \tag{A.2.8}$$

and

$$\mathbf{Q}_t = \mathrm{irow}\{(\mathbf{B} \otimes \mathbf{I} + \mathbf{I} \otimes \mathbf{B})^{-1} (\mathbf{B}_{\Delta t} \otimes \mathbf{B}_{\Delta t} - \mathbf{I} \otimes \mathbf{I})\mathrm{row}\{\mathbf{GG}'\}\}, \tag{A.2.9}$$

where $\otimes$ denotes the Kronecker product, row$\{\cdot\}$ denotes the "rowvec" operation. This operation puts the elements of a matrix row-wise in a column vector. The irow$\{\cdot\}$ is the inverse of the rowvec operation.

With the matrices and nonlinear constraints defined, the final part of the estimation procedure is defining the likelihood function. Under the normality assumption, and given the matrix specifications, the parameter estimates are obtained by optimizing the following maximum likelihood function, given the restrictions in Equation (A.2.6) through Equation (A.2.9),

$$\mathscr{L} = \sum_{i=1}^{N} \{ l_i \ln(2\pi) + \ln(|\boldsymbol{\Sigma}_{\mathrm{M},i}|) + (\mathbf{Y}_i - \boldsymbol{\mu}_{\mathrm{M},i}) \boldsymbol{\Sigma}_{\mathrm{M},i}^{-1} (\mathbf{Y}_i - \boldsymbol{\mu}_{\mathrm{M},i})' \}, \quad \text{(A.2.10)}$$

with $\mathbf{Y}_i$ representing the vector of observed variables and l_i being the number of non-missing observed variables in row i.[13] We use the C + + based optimizer for solving nonlinear programs (see Zahery et al. 2017), which is available in the "`OpenMx`" package (Neale et al. 2016).

Table A.1.1 and Table A.1.2 in Appendix A.1 presented an overview of missingness. As explained in Section 5, we estimate our model using full-information maximum likelihood (Anderson 1957). Consequently, the matrices in Equation (A.2.10), $\boldsymbol{\Sigma}_{\mathrm{M},i}$, $\mathbf{Y}_i$, and $\boldsymbol{\mu}_{\mathrm{M}i}$, are "filtered." This filtering is how full-information maximum likelihood handles missing data. The filtering is performed based on the pattern of missingness.

A.3. Further Notes on Measurement Models

We first identify a measurement model in which the manifest variables are continuous. We next consider a measurement model where the manifest variables are categorical (*e.g.*, binary, ordinal). For identification, we assume, as is the case in our sample, a minimum of two *valid* measures (*i.e.*, manifest variables) in each period and a minimum of two periods.

A.3.1. Continuous Scales

Consider the following measurement model,

$$\mathrm{M}_{l,i,t} = \mu_{l,t} + \lambda_{l,t}\mathrm{S}_{i,t} + \zeta_{l,i,t}, \quad \text{(A.3.11)}$$

[13] In Equation (A.2.10), $|\cdot|$ denotes the matrix determinant.

where $M_{l,i,t}$ denotes the lth manifest variable (for $l = 1, 2$) for child i (for $i = 1, \ldots, N$) and time t (for $t = 0, 1$), $\mu_{l,t}$ denotes the intercept, $\lambda_{l,t}$ denotes the factor loading, $S_{i,t}$ denotes the unobserved skill (*i.e.*, the common factor), and $\zeta_{l,i,t}$ denotes the error term (*i.e.*, the unique factor).

We use the following normalizations: To set the location, we normalize the mean of the the common factor to 0, $\mathbb{E}(S_{i,t}) = 0 \; \forall \; t = 0, 1$, where $\mathbb{E}(\cdot)$ denotes the expectation operator. To set the scale, we normalize a factor loading (say the first) to 1, $\lambda_{1,t} = 1 \; \forall \; t = 0, 1$. Additionally, we made the following assumptions: (i) $\text{Cov}(\zeta_{l,i,t}, S_{i,t}) = 0 \; \forall \; l = 1, 2, t = 0, 1$, (ii) $\text{Cov}(\zeta_{l,i,t}, \zeta_{l',i,t'} | S_{i,t}) = 0 \; \forall \; l, l' = 1, 2, \; l \neq l', \; t = 0, 1$, and (iii) the unique factor is independent across children, where $\text{Cov}(\cdot, \cdot)$ denotes the covariance operator.

Below, we write the unknown parameters (right-hand side) as a function of known (or identified) parameters (left-hand side). We can write the following covariances,

$$\text{Cov}(M_{1,i,t}, M_{1,i,t+1}) = \text{Cov}(S_{i,t}, S_{i,t+1}), \tag{A.3.12}$$

$$\text{Cov}(M_{1,i,t}, M_{2,i,t+1}) = \lambda_{2,t+1}\text{Cov}(S_{i,t}, S_{i,t+1}), \tag{A.3.13}$$

$$\text{Cov}(M_{2,i,t}, M_{1,i,t+1}) = \lambda_{2,t}\text{Cov}(S_{i,t}, S_{i,t+1}). \tag{A.3.14}$$

We can identify the factor loading, $\lambda_{2,t+1}$, by taking the ratio of Equation (A.3.14) to Equation (A.3.12) and $\lambda_{2,t}$ by taking the ratio of Equation (A.3.13) to Equation (A.3.12).

With the factor loadings identified, we can identify the common factor variance from,

$$\frac{\text{Cov}(M_{1,i,t}, M_{2,i,t})}{\lambda_{2,t}} = \frac{\lambda_{2,t}\text{Var}(S_{i,t})}{\lambda_{2,t}} = \text{Var}(S_{i,t}), \tag{A.3.15}$$

for $t = 0, 1$. With the common factor variance identified, we can identify the unique factor variance from,

$$\text{Var}(M_{l,i,t}) - (\lambda_{l,t})^2\text{Var}(S_{i,t}) = \text{Var}(\zeta_{l,i,t}), \tag{A.3.16}$$

for $t = 0, 1$ and $l = 1, 2$. We next identify the mean structure. We can identify the intercepts from the expectations,

$$\mathbb{E}(\mathrm{M}_{l,i,t}) = \mu_{l,t} + \lambda_{l,t}\mathbb{E}(\mathrm{S}_{i,t}) = \mu_{l,t}, \tag{A.3.17}$$

for $t = 0, 1$ and $l = 1, 2$.

With the factor loadings identified, we can (nonparametrically) identify the distribution of common and unique factors using Kotlarski's lemma (see Lemma 1 and Remark 4 in Kotlarski 1967, pp. 70–73). However, these (nonparametric) identification results no longer hold when measures have an ordinal scale.

A.3.2. Categorical Scales

Consider again the measurement model in Equation (A.3.11)

$$\mathrm{M}^*_{l,i,t} = \mu_{l,t} + \lambda_{l,t}\mathrm{S}_{i,t} + \zeta_{l,i,t}, \tag{A.3.18}$$

but the observed manifest variable, $\mathrm{M}^*_{l,i,t}$ is now a latent response variable related to observed (categorical) responses through a threshold function. Consider the Head-Toes-Knees-Shoulders Test. Per item, children received two points when they performed the task correctly, one point when they carried out an incorrect move but ended with a correct response, and zero points for incorrect responses. We can write the following threshold function,

$$\mathrm{M}_{l,i,t} = \begin{cases} 0 & \text{if } \mathrm{M}^*_{l,i,t} < \tau_{1,l,t} \\ 1 & \text{if } \tau_{1,l,t} < \mathrm{M}^*_{l,i,t} < \tau_{2,l,t} \\ 2 & \text{if } \tau_{2,l,t} < \mathrm{M}^*_{l,i,t} \end{cases}$$

where $\tau_{1,l,t}$ and $\tau_{2,l,t}$ are threshold parameters that provide a mapping from the common factor, $\mathrm{S}_{i,t}$, and unique factor, $\zeta_{l,i,t}$, to the observed ranks.

Consider the case in which we assume the factors are normally distributed. Doing so, we can write the probabilities associated with a child achieving a particular score as follows:

$$\Pr(\mathrm{M}_{l,i,t} = 0) = \Phi\left(\frac{\tau_{1,l,t}}{\sqrt{(\lambda_{l,t})^2 + \mathrm{Var}(\zeta_{l,i,t})}}\right)$$

$$\Pr(\mathrm{M}_{l,i,t} = 1) = \Phi\left(\frac{\tau_{1,l,t} - \tau_{2,l,t}}{\sqrt{(\lambda_{1,l,t})^2 + \mathrm{Var}(v_{l,i,t})}}\right) - \Phi\left(\frac{\tau_{1,l,t}}{\sqrt{(\lambda_{l,t})^2 + \mathrm{Var}(\zeta_{l,i,t})}}\right)$$

$$\Pr(\mathrm{M}_{l,i,t} = 2) = 1 - \Phi\left(\frac{\tau_{1,l,t} - \tau_{2,l,t}}{\sqrt{(\lambda_{l,t})^2 + \mathrm{Var}(\zeta_{l,i,t})}}\right)$$

where $\Phi(\cdot)$ denotes the cumulative normal distribution.

Since variances are redundant in polychoric and tetrachoric correlation matrices, we require additional normalizations as there would be 13 unknowns and only 10 knowns. If we additionally normalize the unique factor variance to 1, $\mathrm{Var}(\zeta_{l,i,t}) = 1 \; \forall \; l = 1, 2, t = 0, 1$, we can establish (local) identification. See, for example, Skrondal & Rabe-Hesketh (2004, pp. 135–158). We can use our known and unknown parameters to demonstrate the local identification. Let $\boldsymbol{\vartheta}$ denote the parameter vector with unknown parameters and let $\mathbf{m}(\boldsymbol{\vartheta})$ denote the vector with reduced-form thresholds and covariances. We can then compute the (10, 13)-dimensional Jacobian, $\mathbb{J}(\boldsymbol{\vartheta}) = \boldsymbol{\partial}\mathbf{m}(\boldsymbol{\vartheta})/\boldsymbol{\partial\vartheta}$. The matrix rank of the Jacobian is 9, which is equal to the number of unknown parameters, so the model is locally identified (if $\boldsymbol{\vartheta}$ is a regular point) (Wald 1950, Skrondal & Rabe-Hesketh 2004).

A.3.3. Estimating the Measurement Models

We first estimate "lower-level" measurement models (results available on request). Specifically, we start by estimating a measurement model for the Head-Toes-Knees-Shoulders Test (30 items), the Hearts and Flowers Test (60 items), the Norwegian Vocabulary Test (20 items) and the Blending Test (12 items). We then use each estimated measurement model to assign values to the common factors. Next, we estimate the measurement models for EF skills, mathematical skills, and language skills using these predicted factor scores. For children's language skills, we only have two measures: the Norwegian Vocabulary Test and the Blending Test. Any prediction error would become part of the error term.

Before we estimated the lower-level measurement models, we had to take a position regarding the distribution and link function. Since we have no prior, we selected a model based on Akaike's Information Criterion

(AIC: Akaike 1987) and the Bayesian Information Criterion (BIC: Schwarz 1978). We considered (i) a Gaussian distribution and identity link function, (ii) a binomial (or ordinal) distribution, and (iii) logit link function, and a binomial (ordinal) distribution and probit link function.

Table Table A.3.4 presents the AIC and BIC values. In Panel A, we document the AIC and BIC for the Ani Banani Math Test. We observe that a Gaussian distribution with identity link function results in a (comparatively) better fit in the first (August 2016) and third (March 2018) assessment waves. We observe that a binomial distribution with a logit link function fits the second assessment wave better. However, to maintain consistency in estimating this measurement model, we choose a Gaussian distribution with an identity link function in each assessment wave. In Panel B, we document the AIC and BIC for the Head-Toes-Knees-Shoulders Test. We observe that an ordinal distribution with logit link function results in a better fit in the first (August 2016) and second (June 2017) assessment waves. In the third wave, a probit link function produces a better fit. We choose an ordinal distribution with a logit link function in each assessment wave to maintain consistency. In Panel C, we document the AIC and BIC for the Hearts and Flowers Test.

We observe that a binomial distribution with a logit link function fits each assessment wave better. Therefore, we use a binomial distribution with a logit link function when estimating the measurement model for further analysis. In Panel D, we document the AIC and BIC for the Norwegian Vocabulary Test. We observe that a Gaussian distribution with an identity link function fits each assessment wave better. Therefore, we use a Gaussian distribution with an identity link function when estimating the measurement model for further analysis. Lastly, in Panel E, we document the AIC and BIC for the Blending Test. We observe that a Gaussian distribution with an identity link function produces a better fit in the first (August 2016) and last (March 2018) measurement waves. A binomial distribution with a probit link function in the second wave results in a better fit. We use a Gaussian distribution with an identity link function in each wave to maintain consistency.

Table A.3.5, Table A.3.6, and Table A.3.7 present the estimates of the measurement models for EFs, mathematical skills, and language skills, respectively. The Head-Toes-Knees-Shoulders Test and Hearts and Flowers Test in Table A.3.5 and the Norwegian Vocabulary Test and Blending Test in Table A.3.7 are the predicted factor scores from the lower-level measurement models. Table A.3.8 reports the variance-covariance matrix for EF skills, mathematical skills, and language skills. Table A.3.9 reports the Pearson correlation matrix for EF skills, mathematical skills, and language skills. These measures of association are based on predicted (Bartlett) factor scores (Bartlett 1937, Thomson 1938).

Lastly, Table A.3.10 and Table A.3.11 report the main results. The difference between Table A.3.10 and Table A.3.11, compared with Table 2 and Table 3, is that we do not use a measurement model. Instead, we standardize the individual tests first, compute a simple arithmetic average, and standardize again so that the composite has mean zero and standard deviation one. The results presented in these tables show that we lose a great deal of precision by not accounting for measurement error. Nonetheless, we find that program-induced improvements in EFs in preschool lead to improvements in mathematical skills and language skills in primary school (statistically significant at the ten percent level).

Table A.3.4. Akaike and Bayesian Information Criteria

	August 2016			June 2017			March 2018		
	(1) Iden.	(2) Probit	(3) Logit	(4) Iden.	(5) Probit	(6) Logit	(7) Iden.	(8) Probit	(9) Logit
Panel A: Ani Banani Math Test									
Obs.	665	665	665	651	651	651	661	661	661
AIC	5,702	6,714	6,714	7,795	7,632	7,625	6,821	7,131	7,129
BIC	5,810	6,813	6,810	7,907	7,730	7,724	6,920	7,230	7,228
Panel B: Head-Toes-Knees-Shoulders Test									
Obs.	644	644	644	645	645	645	659	659	659
AIC	35,028	19,921	19,905	40,678	22,369	22,367	34,775	21,469	21,497
BIC	35,430	20,323	20,307	41,081	22,771	22,769	35,179	21,874	21,901
Panel C: Hearts and Flowers Test									
Obs.	642	642	642	635	635	635	660	660	660
AIC	44,965	43,008	42,992	40,141	39,490	39,462	30,636	33,553	33,542
BIC	45,474	43,517	43,501	40,638	39,998	39,970	31,148	34,065	34,054

Continued on next page

Table A.3.4. *Continued from previous page*

	August 2016			June 2017			March 2018		
	(1) Iden.	(2) Probit	(3) Logit	(4) Iden.	(5) Probit	(6) Logit	(7) Iden.	(8) Probit	(9) Logit
Panel D: Norwegian Vocabulary Test									
Obs.	647	647	647	648	648	648	659	659	659
AIC	13,043	13,183	13,178	11,372	12,317	12,319	6,904	10,538	10,535
BIC	13,223	13,362	13,357	11,551	12,496	12,497	7,084	10,718	10,715
Panel E: Blending Test									
Obs.	648	648	648	645	645	645	658	658	658
AIC	4,561	5,550	5,568	7,292	7,134	7,146	6,133	6,768	6,747
BIC	4,669	5,657	5,675	7,399	7,241	7,253	6,241	6,875	6,855

Notes. This table reports the number of observations, Akaike's Information Criterion (AIC: Schwarz 1978), and Bayesian Information Criterion (BIC: Schwarz 1978). We considered a Gaussian distribution and identity (Iden.) link functions (Columns 1, 4, and 7), a binomial (or ordinal) distribution and probit link function (Columns 2, 5, and 8), and a binomial (ordinal) distribution and logit link function (Columns 3, 6, and 9). All models are estimated with maximum likelihood.

Table A.3.5. Measurement Model Parameter Estimates: Executive Functions

	August 2016		June 2017		March 2018	
	(1) $\lambda_{l,0}$	(2) $Var(\zeta_{l,i,0})$	(3) $\lambda_{l,1}$	(4) $Var(\zeta_{l,i,1})$	(5) $\lambda_{l,2}$	(6) $Var(\zeta_{l,i,2})$
Digit Span Test	1	0.60	1	0.53	1	0.60
		(0.05)		(0.04)		(0.05)
Head-Toes-Knees-Shoulders Test	0.77	0.75	0.66	0.47	0.75	0.63
	(0.09)	(0.06)	(0.07)	(0.03)	(0.10)	(0.05)
Hearts and Flowers Test	0.92	0.59	1.03	0.48	1.14	0.59
	(0.10)	(0.05)	(0.09)	(0.04)	(0.13)	(0.05)

Notes. This table reports the measurement model parameter estimates for children's EFs. The common factor is EF skills. The manifest variables are the (Forward/Backward) Digit Span Test, the Head-Toes-Knees-Shoulders Test, and the Hearts and Flowers Test. We standardized the manifest variables to be mean zero and standard deviation one. For this reason, we omit the measurement model intercept estimates. We anchor children's EF skills in the Digit Span Test. Columns (1), (3), and (5) present factor loading estimates. Columns (2), (4), and (6) present unique factor variance estimates. We present the Huber-White standard errors in parentheses. All models are estimated with full-information maximum likelihood (total observations 701). The log-likelihood is $-7,313.07$.

Table A.3.6. Measurement Model Parameter Estimates: Mathematical Skills

	August 2016			June 2017			March 2018		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Items	$\mu_{l,0}$	$\lambda_{l,0}$	Var($\zeta_{l,i,0}$)	$\mu_{l,1}$	$\lambda_{l,1}$	Var($\zeta_{l,i,1}$)	$\mu_{l,2}$	$\lambda_{l,2}$	Var($\zeta_{l,i,2}$)
1	0.24	1	0.12	0.59	1	0.17	0.88	1	0.09
	(0.02)		(0.01)	(0.02)		(0.01)	(0.01)		(0.01)
2	0.10	0.15	0.09	0.26	0.72	0.15	0.47	2.18	0.20
	(0.01)	(0.06)	(0.01)	(0.02)	(0.09)	(0.01)	(0.02)	(0.41)	(0.01)
3	0.08	0.45	0.06	0.29	1.04	0.13	0.59	3.25	0.13
	(0.01)	(0.08)	(0.01)	(0.02)	(0.11)	(0.01)	(0.02)	(0.56)	(0.01)
4	0.37	0.58	0.21	0.60	0.42	0.23	0.82	0.27	0.15
	(0.02)	(0.10)	(0.01)	(0.02)	(0.07)	(0.01)	(0.01)	(0.17)	(0.01)
5	0.43	0.68	0.22	0.70	0.48	0.20	0.83	0.88	0.13
	(0.02)	(0.10)	(0.01)	(0.02)	(0.07)	(0.01)	(0.01)	(0.19)	(0.01)
6	0.13	0.43	0.11	0.32	0.62	0.19	0.57	1.44	0.22
	(0.01)	(0.07)	(0.01)	(0.02)	(0.07)	(0.01)	(0.02)	(0.28)	(0.01)
7	0.71	0.63	0.18	0.87	0.34	0.11	0.93	0.47	0.06
	(0.02)	(0.10)	(0.01)	(0.01)	(0.06)	(0.01)	(0.01)	(0.14)	(0.01)

Continued on next page

Table A.3.6. *Continued from previous page*

	August 2016			June 2017			March 2018		
Items	(1) $\mu_{l,0}$	(2) $\lambda_{l,0}$	(3) $Var(\zeta_{l,i,0})$	(4) $\mu_{l,1}$	(5) $\lambda_{l,1}$	(6) $Var(\zeta_{l,i,1})$	(7) $\mu_{l,2}$	(8) $\lambda_{l,2}$	(9) $Var(\zeta_{l,i,2})$
8	0.66	0.68	0.20	0.84	0.43	0.12	0.87	0.91	0.14
	(0.02)	(0.10)	(0.01)	(0.01)	(0.06)	(0.01)	(0.01)	(0.19)	(0.01)
9	0.42	0.69	0.22	0.60	0.57	0.22	0.82	1.12	0.14
	(0.02)	(0.09)	(0.01)	(0.02)	(0.08)	(0.01)	(0.02)	(0.24)	(0.01)
10	0.11	0.62	0.08	0.37	1.07	0.15	0.70	2.09	0.16
	(0.01)	(0.06)	(0.01)	(0.02)	(0.07)	(0.01)	(0.02)	(0.29)	(0.01)
11	0.03	0.19	0.03	0.15	0.71	0.09	0.49	3.08	0.15
	(0.01)	(0.05)	(0.01)	(0.01)	(0.09)	(0.01)	(0.02)	(0.55)	(0.01)

Notes. This table reports the measurement model parameter estimates for children's mathematical skills. The common factor is a mathematical skill. The manifest variables are the 11 (dichotomous) items presented in the rows. We anchor children's mathematical skills in the first item. This item asks children to copy a pattern. Note that the ordering of items is not the same as the ordering in the original Ani Banani Math Test. Columns (1), (4), and (7) present the measurement model intercepts. Columns (2), (5), and (8) present the factor loadings. Columns (3), (6), and (9) present unique factor variance estimates. We present the Huber-White standard errors in parentheses. All models are estimated with full-information maximum likelihood (total observations 701). The log-likelihood is –9,757.48.

Table A.3.7. Measurement Model Parameter Estimates: Language Skills

	August 2016		June 2017		March 2018	
	(1)	(2)	(3)	(4)	(5)	(6)
	$\lambda_{l,0}$	$\text{Var}(\zeta_{l,i,0})$	$\lambda_{l,1}$	$\text{Var}(\zeta_{l,i,1})$	$\lambda_{l,2}$	$\text{Var}(\zeta_{l,i,2})$
Blending Test	1	0.96	1	0.87	1	0.85
		(0.06)		(0.05)		(0.05)
Norwegian Vocabulary Test	3.71	0.23	2.37	0.24	2.48	0.17
	(0.69)	(0.09)	(0.26)	(0.05)	(0.30)	(0.06)

Notes. This table reports the measurement model parameter estimates for children's language skills. The common factor is language skills. The manifest variables are the Blending Test and the Norwegian Vocabulary Test. We standardized the manifest variables to be mean zero and standard deviation one. For this reason, we omit the measurement model intercept estimates. We anchor children's language skills in the Blending Test. Columns (1), (3), and (5) present factor loading estimates. Columns (2), (4), and (6) present unique factor variance estimates. We present the Huber-White standard errors in parentheses. All models are estimated with full-information maximum likelihood (total observations 701). The log-likelihood is –4, 775.79.

Table A.3.8. Variance-Covariance Matrix Children's Skills

	August 2016			June 2017			March 2018		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
1. EF Skills	0.61	0.12	0.07	0.36	0.12	0.08	0.29	0.04	0.07
2. Mathematical Skills	0.12	0.10	0.02	0.11	0.06	0.03	0.10	0.02	0.03
3. Language Skills	0.07	0.02	0.08	0.06	0.02	0.03	0.05	0.01	0.02
4. EF Skills	0.35	0.11	0.06	0.52	0.14	0.11	0.31	0.05	0.07
5. Mathematical Skills	0.12	0.06	0.02	0.14	0.10	0.04	0.11	0.02	0.02
6. Language Skills	0.08	0.03	0.03	0.11	0.04	0.17	0.07	0.01	0.04
7. EF Skills	0.29	0.10	0.05	0.31	0.11	0.07	0.44	0.04	0.05
8. Mathematical Skills	0.04	0.02	0.01	0.05	0.02	0.01	0.04	0.01	0.01
9. Language Skills	0.07	0.03	0.02	0.07	0.02	0.04	0.05	0.01	0.14

Notes. This table reports the variance-covariance matrix for children's EF skills, mathematical skills, and language skills in each of the three assessment waves. We calculate these estimates based on predicted (Bartlett) factor scores (Bartlett 1937, Thomson 1938). The numbers in the columns refer to the numbers in the rows (and the corresponding skills). For instance, in Panel A, the first row (i.e., (1) EF skills) and the second column (2) present the covariance between children's EFs and mathematical skills.

Table A.3.9. Pearson Correlation Matrix Children's Skills

	August 2016			June 2017			March 2018		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
1. EF Skills	1.00	0.49	0.32	0.65	0.50	0.24	0.56	0.44	0.22
2. Mathematical Skills	0.49	1.00	0.21	0.50	0.60	0.23	0.46	0.50	0.22
3. Language Skills	0.32	0.21	1.00	0.28	0.23	0.30	0.26	0.17	0.21
4. EF Skills	0.64	0.50	0.28	1.00	0.61	0.36	0.65	0.61	0.27
5. Mathematical Skills	0.50	0.60	0.23	0.61	1.00	0.31	0.52	0.65	0.20
6. Language Skills	0.24	0.23	0.30	0.36	0.31	1.00	0.26	0.25	0.27
7. EF Skills	0.56	0.46	0.26	0.65	0.52	0.26	1.00	0.54	0.19
8. Mathematical Skills	0.44	0.50	0.17	0.61	0.65	0.25	0.51	1.00	0.19
9. Language Skills	0.22	0.22	0.21	0.27	0.20	0.27	0.19	0.19	1.00

Notes. This table reports the Pearson correlation matrix for children's EF skills, mathematical skills, and language skills in each of the three assessment waves. We calculate these estimates based on predicted (Bartlett) factor scores (Bartlett 1937, Thomson 1938). The numbers in the columns refer to the numbers in the rows (and the corresponding skills). For instance, in Panel A, the first row (i.e., (1) EF skills) and the second column (2) present the covariance between children's EFs and mathematical skills.

Table A.3.10. Self-Reinforcement and Cross-Production of Skills using Simple Arithmetic Averages

	Mathematics		Language		EFs	
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Developmental Stage 1 (2016 - 2017)						
EF	0.219**	0.206**	0.103*	0.114*	0.534**	0.535**
	(0.038)	(0.038)	(0.036)	(0.032)	(0.043)	(0.043)
Mathematics	0.425**	0.415**	0.112*	0.119*	0.194**	0.196**
	(0.029)	(0.030)	(0.041)	(0.043)	(0.042)	(0.043)
Language	0.140**	0.145**	0.556**	0.543**	0.079**	0.070*
	(0.040)	(0.041)	(0.078)	(0.043)	(0.030)	(0.032)
Panel B: Developmental Stage 2 (2017 - 2018)						
EFs	0.287**	0.296**	0.171*	0.138*	0.568**	0.557**
	(0.034)	(0.033)	(0.039)	(0.033)	(0.038)	(0.038)
Mathematics	0.426**	0.412**	0.060	0.083*	0.187**	0.187**
	(0.041)	(0.043)	(0.047)	(0.035)	(0.044)	(0.043)
Language	0.066	0.081*	0.543**	0.499**	0.029	0.025
	(0.034)	(0.037)	(0.035)	(0.034)	(0.035)	(0.033)
Control Variables		✓		✓		✓

Notes. This table reports the self-productivity parameter estimates. The columns denote the dependent variables, and the rows denote the independent variables. The table reports estimates from twelve models. The control variables include: child sex, birth month, whether or not at least one of the parents is a non-Western immigrant, parental education, family income, an indicator for late parental consent, and randomization-block indicators. Bootstrap (clustered) standard errors in parentheses (1,000 repetitions). All models are estimated with full-information maximum likelihood (701 obs.). Appendix A.1 provides an overview of missingness.
* $p < 0.05$ and ** $p < 0.01$ (two-tailed).

Table A.3.11. Treatment Effects and Decomposition using Simple Arithmetic Averages

	Mathematics		Language		EFs	
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Post-Intervention and Follow-Up Treatment Effects						
Post-Intervention	0.167	0.152	0.015	0.022	0.108	0.113
	(0.091)	(0.089)	(0.078)	(0.077)	(0.060)	(0.058)
Follow-Up	0.227**	0.221**	0.010	0.051	0.063	0.071
	(0.060)	(0.058)	(0.089)	(0.085)	(0.053)	(0.055)
Panel B: Decomposition of Total Treatment Effect						
EFs	0.031	0.033	0.019	0.016*	0.057	0.059
	(0.018)	(0.017)	(0.010)	(0.008)	(0.036)	(0.033)
	[13.7%]	[14.9%]	[27.9%]	[30.8%]	[50.4%]	[56.2%]
Mathematics	0.070	0.062	0.010	0.013	0.031	0.028
	(0.039)	(0.036)	(0.005)	(0.007)	(0.017)	(0.017)
	[30.8%]	[28.1%]	[14.7%]	[25.0%]	[27.4%]	[26.7%]
Control Variables		✓		✓		✓

Continued on next page

Table A.3.11 *Continued from previous page*

	Mathematics		Language		EFs	
	(1)	(2)	(3)	(4)	(5)	(6)
Panel B: Decomposition of Total Treatment Effect						
Language	0.001	0.002	0.010	0.016	0.000	0.001
	(0.005)	(0.006)	(0.043)	(0.039)	(0.001)	(0.002)
	[0.4%]	[0.9%]	[14.7%]	[30.8%]	[0.0%]	[1.0%]
Unmeasured Variables	0.125	0.124	−0.029	0.007	−0.025	−0.017
	(0.080)	(0.076)	(0.083)	(0.082)	(0.061)	(0.063)
	[55.1%]	[56.1%]	[42.6%]	[13.5%]	[22.1%]	[16.2%]
Control Variables		✓		✓		✓

Notes. This table reports the parameter estimates for the treatment effect decomposition. Panel A reports the post-intervention treatment effect (*i.e.*, differences between treated and non-treated children post-intervention) and the follow-up treatment effect (*i.e.*, differences between treated and non-treated children at the follow-up). Panel B decomposes the follow-up treatment effect in measured and unmeasured variables. The columns denote the dependent variables, and the rows denote the independent variables. We include the following control variables: child sex, birth month, whether or not at least one of the parents is a non-Western immigrant, parental education, family income, an indicator for late parental consent, and randomization-block indicators. Standard errors (in parentheses) are computed using a wild residual (clustered) bootstrap procedure (1,000 bootstrap samples). In brackets, we report the relative contributions. We compute these relative contributions by taking the absolute value of the estimate divided by the sum of absolute values of all estimates multiplied by 100. All models are estimated with full-information maximum likelihood (total observations 701). Appendix A.1 provides an overview of missingness.

* $p < 0.05$, and ** $p < 0.01$ (two-tailed).

A.4. Further Notes on Croon's Correction Method

We provide further notes on Croon's correction method (Croon 2002) in this appendix. Let,

$$\varphi_{h,t} = (\boldsymbol{\lambda}'_{h,t}\boldsymbol{\Sigma}^{-1}_{\zeta,h,t}\boldsymbol{\lambda}_{h,t})^{-1}\boldsymbol{\lambda}'_{h,t}\boldsymbol{\Sigma}^{-1}_{\zeta,h,t},$$

denote the (Bartlett) factor scoring matrix, where $\boldsymbol{\lambda}_{h,t}$ is a L_h-dimensional vector with factor loadings, and $\boldsymbol{\Sigma}_{\zeta,h,t}$ is a (L_h, L_h)-dimensional matrix with unique factor variances. We can then write the factor scores as,

$$\tilde{\mathrm{S}}_{h,i,t} = \varphi_{h,t}(\mathbf{M}_{h,i,t} - \boldsymbol{\mu}_{h,t}).$$

Consider data in which we want to estimate the relationship between skill h and skill k, $h \neq r$, using the predicted factor scores,

$$\tilde{\mathrm{S}}_{h,i} = \beta_0 + \beta_1\tilde{\mathrm{S}}_{r,i} + \varepsilon_{h,i}.$$

We can write,

$$\hat{\beta}_1 = \frac{\mathrm{Cov}(\tilde{\mathrm{S}}_{h,i}, \tilde{\mathrm{S}}_{r,i})}{\mathrm{Var}(\tilde{\mathrm{S}}_{r,i})}.$$

We can write the covariance, $\mathrm{Cov}(\tilde{\mathrm{S}}_{h,i}, \tilde{\mathrm{S}}_{r,i})$, as,

$$\begin{aligned}
\mathrm{Cov}(\tilde{\mathrm{S}}_{h,i}, \tilde{\mathrm{S}}_{r,i}) &= \mathbf{Cov}(\varphi_h\mathbf{M}_h, \varphi_k\mathbf{M}_r) \\
&= \varphi_h\mathbf{Cov}(\mathbf{M}_h, \mathbf{M}_r)\varphi'_r \\
&= \varphi_h\mathbf{Cov}(\boldsymbol{\lambda}_h\mathrm{S}_{h,i} + \boldsymbol{\zeta}_{h,i}, \boldsymbol{\lambda}_r\mathrm{S}_{r,i} + \boldsymbol{\zeta}_{r,i})\varphi'_r \\
&= \varphi_h\boldsymbol{\lambda}_h\mathbf{Cov}(\mathrm{S}_{h,i} + \boldsymbol{\zeta}_{h,i}, \mathrm{S}_{r,i} + \boldsymbol{\zeta}_{r,i})\boldsymbol{\lambda}'_r\varphi'_r \\
&= \varphi_h\boldsymbol{\lambda}_h\mathrm{Cov}(\mathrm{S}_{h,i}, \mathrm{S}_{r,i})\boldsymbol{\lambda}'_r\varphi'_r.
\end{aligned}$$

We can write the variance as follows,

$$\begin{aligned}
\mathrm{Var}(\tilde{\mathrm{S}}_{r,i}) &= \mathbf{Var}(\varphi_r\mathbf{M}_r) \\
&= \varphi_r\mathbf{Var}(\mathbf{M}_r)\varphi'_r \\
&= \varphi_r\mathbf{Var}(\boldsymbol{\lambda}_r\mathrm{S}_{r,i} + \boldsymbol{\zeta}_{r,i})\varphi'_r \\
&= \varphi_r\boldsymbol{\lambda}_r(\mathrm{Var}(\mathrm{S}_{r,i}) + \mathbf{Var}(\boldsymbol{\zeta}_{r,i}))\boldsymbol{\lambda}'_r\varphi'_r.
\end{aligned}$$

It follow, then, that,

$$\hat{\beta}_1 = \frac{\text{Cov}(\tilde{\text{S}}_{h,i}, \tilde{\text{S}}_{r,i})}{\text{Var}(\tilde{\text{S}}_{r,i})} = \frac{\varphi_h \boldsymbol{\lambda}_h \text{Cov}(\text{S}_{h,i}, \text{S}_{r,i}) \boldsymbol{\lambda}'_k \varphi'_r}{\varphi_r \boldsymbol{\lambda}_r (\text{Var}(\text{S}_{r,i}) + \mathbf{Var}(\boldsymbol{\zeta}_{r,i})) \boldsymbol{\lambda}'_r \varphi'_r}$$
$$= \text{Attenuation Factor} \cdot \frac{\text{Cov}(\text{S}_{h,i}, \text{S}_{r,i})}{\text{Var}(\text{S}_{r,i})},$$

where,

$$\text{Attenuation Factor} = \frac{\varphi_h \boldsymbol{\lambda}_h \text{Var}(\text{S}_{r,i}) \boldsymbol{\lambda}'_r \varphi'_r}{\varphi_r \boldsymbol{\lambda}_k (\text{Var}(\text{S}_{r,i}) + \mathbf{Var}(\boldsymbol{\zeta}_{r,i})) \boldsymbol{\lambda}'_r \varphi'_r}$$

Since all terms in the attenuation factor are obtained from estimating the measurement model, we can divide $\hat{\beta}_1$ by the attenuation factor to obtain the corrected parameter estimates.

Chapter 4

Teacher Relationship Skills and Student Learning

By

Maximiliaan W. P. Thijssen
Mari Rege
Oddny J. Solheim

Despite extensive evidence on variation in teacher value-added, we have a limited understanding of why some teachers are more effective in promoting human capital than others. Using rich, high-quality data from Norway, we introduce and validate a new approach to measuring teachers' overall capacity to form positive relationships in the classroom, relying on student survey items previously developed and validated (at the student level) in the education literature. We denote this measure as teacher relationship skills. We find that teacher relationship skills are highly stable over time. Furthermore, there is not only substantial variation in teacher quality, as measured by students' learning outcomes conditional on past achievement, but also in teacher relationship skills, even within the same school. Finally, by relying on as-good-as random class assignment, we show that teacher relationship skills affect student learning.

Keywords. Human capital, teacher quality, teacher relationship skills, social-emotional skills, and academic achievement.

JEL codes. I21, J24, H75.

Author note. We received funding from Grant 270703 from the Norwegian Research Council. The authors acknowledge data from the Two Teachers project funded by Grant 256197 from the Norwegian Research Council. We thank Daniel Kreisman, two anonymous referees, May Linn Auestad, Eric Bettinger, Andreas Ø. Fidjeland, Ingeborg F. Solli, Erik Ø. Sørensen and Torberg Falch, participants of the European Economic Association Congress 2020, the EALE SOLE AASLE World Conference 2020, the 42nd Annual Meeting of the Norwegian Economic Association of Economists, the Quantitative Forum at the University of Stavanger, and the PhD Workshop in Economics of Education, for helpful comments.

1. Introduction

There are striking individual differences in the extent to which teachers contribute to students' development, even within the same school (Aaronson et al. 2007, Araujo et al. 2016, Jackson 2018, Kraft 2019, Rivkin et al. 2005, Rockoff 2004).[1] What is more, the effects of a good teacher seem to last into adulthood (Chetty et al. 2014*b*) and can even benefit the future peers of affected students (Opper 2019). Despite extensive evidence on teacher value-added variation, we warrant more research to understand better why some teachers are more effective in promoting human capital than others.[2]

The child development literature suggests that the child's relationship with the teacher and classmates correlates with social, emotional, and academic development (Hamre & Pianta 2001, 2005, Howes et al. 1994, Parker & Asher 1987). Children who experience warm, supportive interactions with the teacher and classmates show greater learning engagement (Klem & Connell 2004), resulting in better academic performance (Roorda et al. 2011) and social adjustment (Pianta 1997).

Forming positive and avoiding negative relationships with and among the children is ultimately the teacher's responsibility. These relationships can create an environment in which children feel competent, independent, and akin to others, which increases their motivation (Connell & Wellborn 1991). To form such positive relationships, the teacher may engage in warm and genuine interactions; respond to social, emotional, and academic needs; encourage group activities; stimulate inclusiveness and provide a structure through sufficient and accurate information on expectations as well as consequences (Connell & Wellborn 1991, Downer et al. 2015). A positive teacher-child relationship also correlates with peer acceptance, which is crucial for a warm classroom climate (Howes et al. 1994). By contrast, negative interactions (*e.g.*, yelling, humiliation) may result in emotional distress, possibly causing distractions and behavioral challenges (Parker & Asher 1987, Pianta 1997).

[1] On average, improving teacher effectiveness by one standard deviation increases performance in reading by 13 percent (range: 7–18 percent) and math by 17 percent (range: 11–25 percent) of a standard deviation (Hanushek & Rivkin 2010).

[2] Observable characteristics such as teacher education do not persistently predict teacher quality (Hanushek 2003).

The numerous studies that suggest that teacher relationship skills, as perceived by the student, are essential for learning may be biased by students' (unobserved) preference for a particular relationship. Recently, economists have started to use classroom observations to measure teacher practices that are less affected by such idiosyncrasies (*e.g.*, Araujo et al. 2016, Kane et al. 2011). However, these classroom observations are resource-intensive and may fail to capture fundamental aspects of students' sentiment that ultimately drives behavior (Connell & Wellborn 1991). Moreover, there is a need to evaluate teachers and what goes on inside the classroom using various assessments (Kane & Staiger 2012).

This study introduces and validates a new approach to measure teachers' overall capacity to form positive relationships and investigates its effect on student learning. We use rich, high-quality data on 5,830 students in 300 classes from 150 schools in Norway from the Two Teachers Project (see Solheim et al. 2017, for the experimental protocol), a randomized controlled trial evaluated in Haaland et al. (2021). We analyze the treated and nontreated first-graders together, which Section 2.2 discuss in further detail.[3] As these early years lay the foundation for the productivity of future investments, they are especially important (Cunha & Heckman 2007). Trained and certified testers assessed the children in one-to-one assessments at the *start* of first grade and the *end* of first, second, and third grade. We matched these assessment data to class and school identifiers and registry data on the family background provided by Statistics Norway. To measure teacher relationship skills, we asked the students a broad set of questions that capture several dimensions of the teachers' ability to form positive relationships with and among the students. We use a leave-out-mean specification to account for the bias that arises from the students' preferences for a particular type of relationship.

Accordingly, the contribution of our paper is twofold. We first introduce and validate a new approach to measure teachers' overall capacity to form positive relationships (at the class level), relying on student survey items previously developed and validated (at the student level) in the education literature. We denote this measure as teacher relationship skills. We validate this measure in two ways: (i) we demonstrate stability over

[3] We present several robustness checks that suggest our findings on the effects of teacher relationships skills are not likely due to the treatment.

time; and (ii) we illustrate that there is not only a considerable variation in teacher quality (as measured by learning outcomes conditional on past achievement) but also in teacher relationship skills, even within the same school. Second, we show that children taught by teachers with better relationship skills develop more academically and socially-emotionally. These results even hold in models that carefully address selection and noise from child-specific idiosyncrasies in survey responses.

Math and literacy are our primary outcome measures. Test scores may not capture all relevant aspects of development, however. Given the importance of early literacy and the growing recognition that social-emotional skills (*e.g.*, beliefs, motivations, interests, and personality traits) are critical to school performance, labor market outcomes, and social behavior (*e.g.*, Bettinger et al. 2018, Borghans et al. 2008, Heckman et al. 2006, Jackson 2018, Kraft 2019), we also focus on skills closely related to motivation in reading: self-concept in reading and reading interest. The former is a measure of children's perceived competence in reading.[4]

We leverage the as-good-as random class assignment in Norwegian primary schools when investigating how the teachers' relationship skills affect student learning. By law, school administrators in Norway should not assign children to classes based on sex, religion, ethnicity, or ability (Kunnskapdepartementet 2017). Consistent with this law, our analysis demonstrates that predetermined variables and variables measured at the *start* of first grade are not predictive of class, teacher, or peer group characteristics.[5] Furthermore, we conduct several placebo, sensitivity, and robustness analyses supporting our (identifying) assumptions. Still, as discussed carefully in our concluding section, our estimates should be interpreted with caution, as we have no source of "clean" exogenous variation in teacher relationship skills.

We find that teacher relationship skills affect math, literacy, and the students' reading motivation. A one standard deviation increase in teacher relationship skills raises math test scores by about five percent and literacy test scores by about three percent of a standard deviation. Five and three

[4]Jensen et al. (2019) find a strong correlation between the *individually* perceived emotional support and self-concept in reading.

[5]Additionally, Chi-square tests of homogeneity reveal a pattern of mean differences between classes (within schools) that is consistent with as-good-as random assignment.

percent of a standard deviation is, respectively, about 15 and 10 percent of the difference in math and literacy test scores between students of mothers with and without a college degree. Concerning a child's motivation to read, we find that a one standard deviation increase in teacher relationship skills raises reading interest by about five percent of a standard deviation and self-concept in reading by about three percent of a standard deviation.[6] We find larger point estimates for boys and children from low socioeconomic households, but the differences are not statistically significant. Lastly, the evidence indicates that the effects of the teacher relationship skills in first grade persist in second grade and for literacy even until third grade, which suggests that the first-grade findings are not an anomaly.

Our paper makes several contributions. The value-added literature referred to in the first paragraph provides ample evidence on variation in teacher value-added. The use of learning gains (conditional on prior achievement and other influences) as a measure of teacher effectiveness is ubiquitous in the literature. Still, such value-added measures only allow identifying and not replicating effective teachers, as Kane et al. (2011) rightly note. In this paper, we show that the teacher's overall capacity to form positive relationships – measured from the students' perspective – can affect student learning. Consequently, teacher relationship skills relate to effective teaching practices and provide a new perspective on teacher evaluations, which is desirable (Kane & Staiger 2012).

A growing number of studies look at objective and subjective evaluations conducted by peers or administrators to understand how to replicate effective teachers (Araujo et al. 2016, Kane et al. 2011, Rockoff & Speroni 2010). Such studies examine the effect of evaluated teacher practices on student learning. Araujo et al. (2016) filmed kindergarten classes for a full day and coded the videotapes using the Classroom Assessment Scoring System (CLASS: Pianta et al. 2008). The CLASS categorizes teacher-child interactions into three domains: emotional support, classroom organization, and instructional support. By leveraging as-good-as random assignment to classrooms, they show that teacher quality, mea-

[6]When we use a control function approach to address the bias of students' preferences for a particular type of relationship in survey responses, we find effect sizes of seven, four, nine, and six percent of a standard deviation for math, literacy, reading interest, and reading self-concept, respectively.

sured using CLASS, in year t is a strong predictor of learning outcomes in year $t+1$. Kane et al. (2011) focus on the Cincinnati public school system and investigate the effect of teachers' ability to create an environment for learning and general teaching practices on learning outcomes. Both are domains in the Teacher Evaluation System, which aims to enhance professional teaching practices. Unable to rely on random assignment, Kane et al. (2011) instead examine the sensitivity of their findings with alternative model specifications such as the inclusion of teacher fixed effects. Rockoff & Speroni (2010) examine the use of objective and subjective evaluations and find that better-evaluated teachers before their hiring (or in their first year) produce more gains in achievement with their future students.

The evidence provided in this paper is consistent with the results of Araujo et al. (2016), Kane et al. (2011), and Rockoff & Speroni (2010), even though we relied on a different type of evaluation, a distinct empirical approach, and other outcomes. Consistent findings across several studies indicate a (causal) link between a set of teacher behaviors and student learning, which has important implications for designing teacher evaluation systems, performance pay, and other potential policies.

In addition to substantiating previous findings, we contribute evidence that the child's perspective can yield meaningful insights into drivers of teacher quality variation (see, *e.g.*, Begrich et al. 2020, Fauth et al. 2020). In many regards, our approach may be more advantageous than the classroom observation approach (*cf.* Araujo et al. 2016).[7] First, our measure is less resource-intensive and is implementable at scale. Second, the child's perspective is more likely to capture cumulative experiences. One disadvantage of using observers is that the data (often) stem from a single point in time. The representativeness of this time point is difficult to infer. Lastly, multiple measures from different perspectives are necessary when studying teacher quality. As Kane & Staiger (2012) and others (*e.g.*, Begrich et al. 2020, Fauth et al. 2020) describe, various assessments can serve as tools for targeting teachers, supporting development, measuring and evaluating progress, and helping policymakers improve quality.

[7]This is not to say that self-reports are without limitations: (i) self-reports rely on the honesty of the respondent, (ii) some respondents may lack introspective abilities, (iii) there is always a concern as to whether respondents fully understand the question, (iv) if respondents answer questions in specific ways, there is a chance of response bias, and (v) the distance between item response categories is generally unknown.

Also, most studies on teacher quality focus on math and literacy test scores. We contribute by providing evidence on skills closely related to academic motivation: children's self-concept in reading and reading interest. These results relate to a growing body of evidence that social-emotional skills can improve academic performance, labor market outcomes, and social behavior (Bettinger et al. 2018, Borghans et al. 2008, Heckman & Kautz 2012, Heckman et al. 2006).[8]

Finally, a large body of evidence shows that readily observable teacher characteristics (*e.g.*, education, salary) do not persistently predict teacher quality (Hanushek 2003). In this literature's spirit, we examine if such characteristics correlate with teacher relationship skills. We find a (meaningful) positive correlation between the teacher's education level and relationship skills. However, teacher education is not a strong correlate of math, literacy, reading interest, or self-concept. While we do not warrant substantive conclusions given a sample of 300 teachers, it is curious that readily observable teacher characteristics are not strong correlates of skill development nor teacher relationship skills, but that teacher relationship skills do predict various learning outcomes. Such findings may inform policy questions about what skills schools should focus on when hiring and evaluating teachers (see, *e.g.*, Goldhaber et al. 2017, Jacob et al. 2018, Stewart et al. 2021).

2. Background

2.1. Institutional Setting

Norway is a high-income country with roughly 5.2 million inhabitants, about nine percent of whom are children in primary education (Statistics Norway 2018). Norway has some extremely rural areas with a population density of 14 people per square kilometer compared to 25 in neighboring country Sweden, 121 in the European Union, and 36 in the United States. As a result, children may have to travel substantial distances to school, especially in rural areas.

[8] See Almlund et al. (2011), Borghans et al. (2008), and Heckman et al. (2006) for a discussion on the relationship between social-emotional skills and economic preference parameters.

In 2015, Norway's per-child expenditure on primary education as a share of gross domestic product per capita was 25.5 percent compared to 22.5 percent in Sweden and 20.7 percent in the United States (OECD 2018). Compared to other countries, Norway invests heavily in compulsory education and preschool education, which the government universally subsidizes. As a result, many children (91 percent in 2016: Statistics Norway 2018) attend preschool programs (ages one through five). Formal schooling in Norway starts in the year the child turns six and lasts until 16. In Norway, a primary school consists of seven grades (ages six through 12) and lower secondary schools of three grades (ages 13 through 16).

Children can (and most do) attend the school closest to home. Furthermore, there is a strong focus on equality in the Norwegian school system, with only about 3.6 percent of the children in private primary and lower secondary schools (Statistics Norway 2018). The state is responsible for financing primary and secondary education, while the municipalities are responsible for operations and administration. Also, by law, administrators should not assign children to classes based on their sex, ethnicity, religion, or ability (Kunnskapdepartementet 2017). Finally, it is common that the assignment of students to classes remains unchanged during the first three years of primary school (*i.e.*, educational looping). The assignment of teachers to classes may change for practical reasons, however. For example, such changes may occur as a result of maternity leave or a dysfunctional relationship. Finally, as we explain in Section 2.2, schools participating in the Two Teachers project had to commit to not changing class assignments.

2.2. Data

We use the first four waves of data from a high-quality longitudinal research project in the southern part of Norway: Two Teachers (Solheim et al. 2017). We invited 6,014 first-graders to participate, which resulted in 300 classes from 150 schools. We randomly selected two classrooms in each of the 150 schools. Schools participating in the project had to commit to not changing class assignments. We obtained consent from 97 percent of the parents resulting in an analytical sample of 5,830 children. Notably, one of the two classrooms was randomly selected at each school to receive an additional teacher during literacy instruction. See Solheim et al. (2017) for more details about the Two Teachers project. The effect of this additional teacher in literacy instruction is studied in Haaland et al. (2021). In the present study, we control for the additional teacher in all specifications. Since the treatment started before we measured teacher relationship skills, we may worry that the treatment affected both teacher relationship skills and student learning. To that point, the mechanism investigation in Haaland et al. (2021) suggests that teacher relationship skills do not mediate the relationship between treatment and student learning. We provide further discussion and evidence that the additional teacher did not significantly affect the teacher relationship skills in Section 4.2.1.

We assessed each child at the *start* of first grade (August 2016: denoted G0), at the *end* of first grade (May 2017: denoted G1), at the *end* of second grade (May 2018: denoted G2), and at the *end* of third grade (May 2019: denoted G3). Each assessment was one-to-one with a trained and certified tester and took place during the school-going hours in a quiet location within the school. The examiners used tablet computers to test the children. Such assessment procedures minimize measurement errors and missing values. Each assessment wave involved a battery of tests appropriate to the child's development. We describe the various assessments below (for further details about the data, see Appendix A.1). In addition to the assessment data, we collected data on relevant child, classroom, and teacher characteristics. We matched these assessment data to Statistics Norway's registry data on relevant family background characteristics. Both the experimental data and the data provided by Statistics Norway have missing observations. For further details about missing observations, see Appendix A.2.

2.3. Measures

In each wave, we assessed the children in math, literacy, reading self-concept, and reading interest.[9] We standardize all test scores and items by subtracting the mean and dividing by the sample standard deviation. Suppose there are more measures for a given skill. In that case, we standardize the individual scores first, construct an arithmetic average, and re-standardize so that the composite variable has a mean of zero and standard deviation one.

[9] See Appendix A.1 for descriptive statistics related to each measure.

For math, our measurement instrument is an arithmetic fact test (Klausen & Reikerås 2016). The children had two minutes to solve as many addition problems as possible in this test. We observe an arithmetic fact test score in each assessment wave. For literacy, our main instruments are a reading fluency test and a spelling test. We observe both instruments in each of the four waves. We based our reading fluency test on the Sight Word Efficiency from the Test of Word Reading Efficacy (TOWRE: Torgesen et al. 1999). In this test, the children had 45 seconds to correctly read as many words as possible. We use the spelling test developed by the Norwegian Reading Centre (2013). We augment the reading fluency test and spelling test with the following literacy tests: letter recognition at the *start* of first grade (G0), reading accuracy at the *end* of first grade, second, and third grade (G1, G2, and G3), and reading comprehension at the *end* of second and third grade (G2 and G3). The literacy measure captures several aspects important in early literacy development.

To measure self-concept in reading, we asked each child a series of questions regarding their believed competence in reading and letters (Chapman & Tunmer 1995, Chapman et al. 2000, Eccles et al. 1993). The items vary in each assessment wave to account for the child's development. For our main outcome variable (*i.e.*, at the end of first grade: G1), the questions are: (1) "*How good are you in letters?*"; (2) "*How good are you at reading?*"; (3) "*How good are you at reading long stories?*"; (4) "*How good are you at finding the meaning of difficult words when you read?*"; (5) "*How good are you in letters compared to your classmates?*"; (6) "*How good are you in reading compared to your classmates?*"; (7) "*Do you think reading is difficult?*"; (8) "*Do you think it is challenging to recognize words you have read before?*"; and (9) "*Do you think it is challenging to understand the meaning of words when you read?*" (Cronbach's alpha is 0.7).[10]

Lastly, we asked the children several questions regarding their interest in reading activities using the Youth Reading Motivation Questionnaire (Coddington & Guthrie 2009). Like our measure of the children's self-concept in reading, the items may vary each assessment wave. For our main outcome variable (G1), the questions are: (1) "*Do you enjoy reading?*";

[10]Cronbach's alpha is an internal consistency statistic, which captures the covariation of items believed to measure an underlying construct (Cronbach 1951). Cronbach's alpha ranges from zero to one. Higher values suggest higher internal consistency.

(2) "*Do you enjoy reading books?*"; (3) "*Do you enjoy reading comics?*"; (4) "*Do you enjoy reading at home?*"; (5) "*Do you usually look forward to reading?*"; (6) "*Would you be happy if you got a book as a present?*"; and (7) "*Do you think reading is boring?*" (Cronbach's alpha is 0.8).

We integrated two sets of questions previously validated at the student level in the education literature to measure the teacher relationship skills. The first set of four questions broadly measures the students' relationship with the teacher. These questions are part of an adapted version of the Classroom Assessment Scoring System – Student Report (CLASS-SR: Downer et al. 2015).[11] The second set of seven questions measures the relationships among the students. These questions are part of an adapted version of the Social Integration Classroom Climate and Self Concept of School Readiness (SIKS: Rauer & Schuck 2003).[12] At the start of each assessment, the tester assured each child that nobody would get to know their answers. The examiner then introduced the questions on a tablet computer, which presented the response alternatives through different smileys.

In the Two Teachers project, we also collected data on whether the mother, father, or siblings have a self-reported reading disability (self-reported by the parents). Furthermore, we know the child's sex, birth month, birth year, number of siblings, the parent's education, income for 2015, and country of birth from Statistics Norway's registry data. Statistics Norway operationalizes income as: "*[…] the sum of wage and net business income […]. Social security and maternity benefits are included.*"

2.4. Mechanisms

The extent to which a child feels engaged to learn depends on fulfilling basic psychological needs: the need to feel competent, the need for auton-

[11] Specifically, we reduced the number of items from the CLASS-SR (see Downer et al. 2015) to four. We also transformed the questions into statements to reduce the potential for cognitive response bias (Bentler et al. 1971).

[12] Holen et al. (2013) report satisfactory reliability of this measure in a Norwegian sample of children aged seven and eight. We use an adapted version of their measure. The SIKS-scale originally consists of three different subscales: classroom climate, social integration and academic skills. We only used the classroom climate subscale and reduced the number of items from 11 to seven. We also changed the items from statements into questions, and from a binary scoring (*i.e.*, true or not true) into a 4-point Likert scale.

omy, and the need to feel related to the teacher and classmates (Connell & Wellborn 1991). By building relationships with and among the children, the teacher can foster learning engagement. For instance, the teacher can provide structure, support independence, and show interest in the children. When the teacher cannot promote such a classroom environment, we expect a child's learning engagement to decrease.

It may be particularly important for children entering school with low achievements scores to feel emotionally supported by their teacher. For this reason, Hamre & Pianta (2001) identify, among others, boys and children born into low socioeconomic households as children for whom positive relationships could be particularly important. Several studies suggest that these groups enter school with lower achievements scores (*e.g.*, Magnuson & Duncan 2016).

These differences are also replicated in our sample. The differences between children from low socioeconomic households in language and math at the *start* of formal schooling (G0) are about 25 and 26 percent of a standard deviation, respectively, both statistically significant at the one percent level.[13] The difference between boys and girls in language at the *start* of formal schooling (G0) is about 30 percent of a standard deviation, statistically significant at one percent level. In math, there are no observable differences between boys and girls, however. These differences motivated the investigation of differential effects across sex and socioeconomic status.

3. Empirical Approach

3.1. Measuring Teacher Relationship Skills

We create a measure of teacher relationship skills using all the items of the adapted CLASS-SR and SIKS scales. Table 1 reports an overview of each item used to measure teacher relationship skills at the *end* of first grade (G1), including internal consistency statistics.[14] Parenthetically, Appendix A.3 reports a similar overview but with items measured at the *end* of the second (G2) and at the *end* of third grade (G3). The Cronbach's alpha suggest that the set of items we employ as manifestations of teacher relationship skills relate strongly together.

[13] We measure socioeconomic status with family income and parental education.

[14] See Appendix A.3 for the item-specific response frequencies.

Table 1. Teacher Relationship Skills Measured at the End of First Grade (G1): Descriptive Statistics

	(1) Mean	(2) *SD*	(3) Obs.
Emotional support ($\alpha = 0.62$)	0.86	0.16	5,611
1. Do you feel as if the teacher is a good friend of yours?	0.83	0.25	5,618
2. Will the teacher help you when you have problems?	0.91	0.19	5,618
3. Do you feel the teachers appreciate you?	0.86	0.25	5,618
4. Do you get help from your teacher when reading is difficult?	0.85	0.26	5,611
Classroom climate ($\alpha = 0.69$)	0.79	0.16	5,632
5. Is everybody in the class good friends?	0.79	0.24	5,632
6. Do you stick together and look after each other?	0.80	0.25	5,632
7. Is there anyone in your class who laughs at children who are different?	0.23	0.31	5,632
8. Do you help each other in the class?	0.84	0.23	5,632
9. Are all the children in the class allowed to play along?	0.84	0.25	5,632
10. Are there students in your class who make fun of others?	0.28	0.30	5,632
11. Do you tease and annoy one another in the class?	0.22	0.27	5,632
A. Positive peer relations (average of items 5, 6, 7, and 8)	0.82	0.17	5,632
B. Negative peer relations (average of 9, 10, and 11)	0.25	0.22	5,632
Overall relationship skills (average of all items, $\alpha = 0.73$)	0.82	0.13	5,611

Notes. This table reports descriptive statistics for the teacher relationship skill items. We normalize all items to take values between 0 and 1 so that means are comparable across items. We present the Cronbach's alpha (Cronbach 1951), denoted α, in parentheses. Cronbach's alpha is a statistic between zero and one that represents how well a set of items measures an underlying latent construct. Values closer to one indicate higher internal consistency.

We construct a composite measure of teacher relationship skills at the student level by computing an arithmetic average of the individual items presented in Table 1 to aid interpretation. First, we standardize the individual items, construct an arithmetic average, and re-standardize. The composite measure is then mean zero and standard deviation one. We do this in each assessment wave where we measure the teacher relationship skills (*i.e.*, at the *end* of first (G1), second (G2), and third grade (G3)). This approach assumes that each item is equally informative concerning teacher relationship skills. When a set of items relate strongly as a group, researchers often employ factor analytic methods to separate the underlying latent construct, identified from the covariation among the items, from unique item-specific variation. In contrast to an arithmetic average, such factor analytic methods would weigh each item based on the level of informativeness. In Section 4.2.3, we apply such factor analytic methods as a robustness check and see that estimates are very similar.

As a first validation of our measure at the class level, we demonstrate that teacher relationship skills, as measured at the items presented in Table 1, are stable throughout the first three years of formal schooling (G1, G2, and G3).[15] To illustrate this stability, we compute classroom averages from the standardized (student-level) composites for classes that did not experience a change in the main teacher in the first (G1), second (G2) or third grade (G3), which is about 60 percent of the sample. The results in Table 2 illustrate this stability. It reports autocorrelation coefficients of about 0.6 between first (G1) and second grade (G2) and about 0.7 between second (G2) and third grade (G3). This stability suggests there is not much fluctuation year to year. In other words, teachers have some routine (*e.g.*, in the form of established pedagogical practices, training, or experience) in the way they interact and build relationships with and among the students year to year. A degree of stability is what one would expect of teacher quality, as Chetty et al. (2014*a*) note and policies concerning teacher quality often assume (Goldhaber & Hansen 2010). Such stability also raises interesting questions about the potential mechanisms that drive stability and conditions under which stability is most probable. We leave these inquiries for future research.

[15] The teacher relationship skill items measured at the *end* of second grade (G2) and at the *end* of third grade (G3) are presented in Appendix A.3.

We examine the between-class variation in teacher relationship skills as a second validation. If there is congruence in the way children taught by the same teacher respond to the items in Table 1, then we expect to observe a considerable variation between classes in our measure. That is if there is variation in teacher quality. By contrast, if children with the same teacher have varied responses, we expect not to find significant differences between classes (within school).

To examine the extent of variation between classes in our measures, we first estimate a series of cluster dummy variable models. We regress each of the five variables in the rows of Table 3 (*i.e.*, math, literacy, self-concept, reading interest, and the measure of teacher relationship skills at the student level measured at the *end* of first grade (G1)) on a set of class (Column 1) and school (Column 2) indicators. We report the adjusted R-squared and the results from an F test for joint significance.[16] All F statistics are significant at the one percent level, indicating significant statistical variation between classes and schools. However, significant variability between classes and schools does not inform about variation between classes *within* schools.

Table 2. Temporal Stability of Teacher Relationship Skills

	(1) End of First Grade (G1)	(2) End of Second Grade (G2)	(3) End of Third Grade (G3)
End of First Grade (G1)	1	0.61^{**}	0.55^{**}
End of Second Grade (G2)	0.61^{**}	1	0.68^{**}
End of Third Grade (G3)	0.55^{**}	0.68^{**}	1

Notes. This table reports autocorrelation coefficients for teacher relationship skills, where teacher relationship skills are measured as the class average of standardized (student-level) composites. We first standardize the items presented in Table 1 to be mean zero and standard deviation one. Second, we average these standardized items for each child. Third, we compute the average of these student-level standardized averages at the class level. We use listwise deletion to handle missing data. We only use classes that stay together with the same teacher in the first (G1), second (G2), and third grade (G3), resulting in 3,380 observations.
$^{+}$ $p < 0.10$, * $p < 0.05$, and ** $p < 0.01$ (two-tailed).

[16]To avoid the dummy variable trap, we omit the intercept.

We follow Chetty et al. (2011) in investigating between-class variation within schools. Let $Y_{G1,i,j,k}$ denote an outcome variable at the *end* of first grade (G1) for student i (for $i = 1, \ldots, N$) in class j (for $j = 1, \ldots, J$) in school k (for $k = 1, \ldots, K$). These outcome variables include math, literacy, self-concept, reading interest, and a composite measure of teacher relationship skills at the student level. We estimate the following regression model for each of the five outcome variables:

$$Y_{G1,i,j,k} = \tau_k + \upsilon_{j,k} + \mathbf{Z}'_{i,j,k}\boldsymbol{\gamma} + \epsilon_{i,j,k}, \tag{1}$$

where τ_k is a school-specific intercept, $\upsilon_{j,k}$ measures the classroom effect (*i.e.*, a classroom dummy variable), $\mathbf{Z}'_{i,j,k}$ is a vector of predetermined child and family background variables and skills measured at the *start* of first grade (G0), and $\epsilon_{i,j,k}$ is a mean zero error term. The vector of variables includes family income, the number of siblings, math measured at the start of first grade, literacy measured at the start of first grade, and indicators for the child's sex, birth month, birth year, family reading disability, education mother, and non-Western immigrant.

To avoid perfect collinearity between τ_k and $\upsilon_{j,k}$, we omit one class in each school. Column (3) reports the F statistics for the joint significance of the classroom dummies for each of the five estimated models in that column. The significant F statistics imply statistically significant variability between classes within schools. The results show that there is significant variation between classes within schools for all variables.

In Column (4), we estimate Equation (1) with $\upsilon_{j,k}$ specified as a random effect to quantify the magnitude of (or variation in) the five class effects. To determine this variation, we make the following distributional assumption: $\upsilon_{j,k}$ is mean-zero and follows a Gaussian distribution with variance $\text{Var}(\upsilon_{j,k}) = \sigma_\upsilon^2$; $\upsilon_{j,k} \sim \text{N}(0, \sigma_\upsilon^2)$. The magnitudes (in standard deviations) presented in Column (4) are sizable, especially for the teacher relationship skills. The estimates for math and literacy are within the range of previous findings from the United States (Hanushek & Rivkin 2010).

In sum, our measure of teacher relationship skills seems to be stable and detect a considerable variation between classes within schools. We next describe our empirical strategy to determine if teacher relationship skills affect students' academic (*i.e.*, math and literacy) and social-emotional learning (*i.e.*, self-concept in reading and reading interest).

Table 3. Variability at the End of First Grade (G1): Significance and Magnitude

	Cluster Dummy Variable Model				Classroom Variation		
	(1) Class		(2) School		(3)	(4) Magnitude	(5)
	Adj. R^2	(F stat.)	Adj. R^2	(F stat.)	F test	(in *SD*)	Obs.
Panel A: Academic and Social-Emotional Variables (G1)							
Math	0.087	(3.414^{**})	0.079	(4.514^{**})	1.965^{**}	0.172	5,610
Literacy	0.078	(3.151^{**})	0.075	(4.496^{**})	1.695^{**}	0.129	5,637
Self-Concept	0.032	(1.883^{**})	0.024	(2.017^{**})	1.409^{**}	0.136	5,636
Reading Interest	0.041	(2.881^{**})	0.023	(1.998^{**})	1.811^{**}	0.220	5,630
Panel B: Teacher Variables (G1)							
Relationship Skills	0.127	(5.090^{**})	0.070	(4.570^{**})	3.531^{**}	0.310	5,634

Notes. This table reports the cluster dummy variable models and the significance and magnitude of the class effects. Each cell reports results from a separate regression. All variables in the rows are at the child level and standardized to have mean zero and standard deviation one. In Column (1) and (2), we report the adjusted R^2 (F statistic in parentheses) from regressing the variables presented in the rows, alternately, on a set of class (Column 1) and school (Column 2) dummy variables, respectively. To avoid the dummy variable trap, we do not include an intercept. Column (3) reports the F statistic for the test of the joint significance of a model that includes school and class fixed effects. We omit one class per school to avoid perfect collinearity. Column (4) reports the estimated magnitude of the class effects in standard deviations (*SD*s). See Appendix A.2 for details on missing data. Column (4) does not include variables that predict missingness, such as tester fixed effects.

$^{+}\ p < 0.10$, $^{*}\ p < 0.05$, and $^{**}\ p < 0.01$ (two-tailed).

3.2. Investigating the Effects of Teacher Relationship Skills

We start from the following linear-in-parameters specification of the education production function for child i (for $i = 1, \dots, N$) in class j (for $j = 1, 2, \dots, J$) in school k (for $k = 1, \dots, K$):

$$\mathrm{Y}_{\mathrm{G1},i,j,k} = \mathbf{Y}'_{\mathrm{G0},i,j,k}\boldsymbol{\alpha} + \beta \mathrm{X}_{j,k} + \mathbf{F}'_i\boldsymbol{\gamma} + \mathbf{P}'_{(-i),j,k}\boldsymbol{\delta} + \mathbf{C}'_{j,k}\boldsymbol{\kappa} + \tau_k + \epsilon_{i,j,k}, \quad (2)$$

where $\mathrm{Y}_{\mathrm{G1},i,j,k}$ is an outcome variable (*i.e.*, children's math test score, literacy test score, self-concept, or reading interest) at the *end* of first grade (G1). $\mathrm{X}_{j,k}$ denotes the classroom average of standardized (student-level) composites of the teacher relationship skill items measured at the *end* of first grade (G1),

$$\mathrm{X}_{j,k} \equiv \frac{1}{N_{j,k}} \sum_{i=1}^{N_{j,k}} \mathrm{X}_{i,j,k}, \quad (3)$$

where $N_{j,k}$ denotes the total number of children in class j in school k. We discussed $\mathrm{X}_{i,j,k}$ — a composite measure of teacher relationship skills at the student level, computed by an arithmetic average of the individual items presented in Table 1 — at the beginning of Section 3.1. $\mathbf{Y}'_{\mathrm{G0},i,j,k}$ is a vector of skills (*i.e.*, math and literacy) measured at the *start* of first grade (G0).[17] $\mathbf{F}'_i$ is a vector of child and family background variables. $\mathbf{P}'_{(-i),j,k}$ is a vector of peer variables (*i.e.*, the skills measured at the *start* of first grade and child and family background variables for child i's peers). $\mathbf{C}'_{j,k}$ is a vector of classroom and teacher variables. τ_k is a school-specific intercept. $\epsilon_{i,j,k}$ is a mean-zero error term. We cluster the errors at the school level to correct for correlations in outcomes across children within classrooms and schools.[18]

[17]We do not include children's self-concept or reading interest because we worry that these measures are endogenous; school had already started during the first assessment. Therefore, we only include skills that are not likely to have changed within the first weeks of schooling such as math and literacy. Primary schools start around the middle of August. The first assessment was on the 22nd of August and the last on September 30. We assessed almost 60 percent of the children before September 1 and 99 percent before September 9.

[18]Although the treatment arises at the class level, clustering at the school level is more conservative in our estimation. The differences are small, however.

The error term, $\epsilon_{i,j,k}$, consists of an unmodeled influence at the class-level and child-level: $\epsilon_{i,j,k} = \upsilon_{j,k} + \zeta_{i,j,k}$, where $\upsilon_{j,k}$ captures "*correlated effects*" (Manski 1993, p.533) that may arise because children are sorted into classes. Such unobserved influences are problematic when they correlate with the teacher's overall capacity to form positive relationships, $X_{j,k}$, and the outcome variable, $Y_{G1,i,j,k}$. Suppose we assume that school administrators (as-good-as) randomly assign children and teachers to classes. In that case, these correlated effects arise at the school level (*i.e.*, $\upsilon_{j,k} = \upsilon_k$) and can thus safely be ignored in models with a school-specific intercept. Therefore, we address the problem that arises from correlated effects under the following assumption:

Assumption 1 (Random Assignment). *Children and teachers are as-good-as randomly assigned to classes (within schools) such that any systematic differences occur at the school level:* $\upsilon_{j,k} = \upsilon_k$.

Importantly, Assumption 1 does not preclude random assignment of unobserved teacher attributes (*e.g.*, didactic capabilities) or teacher behaviors (Araujo et al. 2016). These unobserved teacher characteristics may correlate with the teacher's capacity to form positive relationships and student learning. Fortunately, we have access to a rich set of student and classroom-related information such as the teacher's experience, age, sex, education, and class size. Therefore, we assume:

Assumption 2 (Exogeneity). *Conditional on the child, peer, teacher, and classroom observables, the teacher relationship skills,* $X_{j,k}$, *do not correlate with the error term,* $\epsilon_{i,j,k}$.

An arithmetic class average, $X_{j,k}$ (Equation 3) is a noisy measure when unobserved preferences for a particular teacher correlate with the children's level of effort and learning and shape their perceptions and evaluations. For instance, as Dee (2004) pointed out, some children may prefer a specific teacher identity bringing about a role-model effect that increases their learning engagement and causes them to evaluate their teacher more positively. In other words, there would be an "*own-observation problem*" (Chetty et al. 2011, p. 1635).

Formally, denote $\theta_{j,k}$ as the effect that is due to the teacher and denote $\rho_{i,j,k}$ as child i's unobserved preferences such that:

$$\mathrm{X}_{j,k} \equiv \frac{1}{N_{j,k}} \sum_{i=1}^{N_{j,k}} \mathrm{X}_{i,j,k} = \theta_{j,k} + \frac{1}{N_{j,k}} \sum_{i=1}^{N_{j,k}} \rho_{i,j,k}. \tag{4}$$

We assume no peer effects (*i.e.*, $\mathrm{Cov}(\theta_{j,k}, \rho_{i,j,k}) = \mathrm{Cov}(\rho_{i,j,k}, \rho_{i',j,k}) = 0$ for $i \neq i'$, where $\mathrm{Cov}(\cdot,\cdot)$ denotes the covariance operator). If child i, who favors the teacher for some unobserved reason, subsequently evaluates the teacher more positively, $\mathrm{Cov}(\mathrm{X}_{j,k}, \rho_{i,j,k}) > 0$, and as a result increases effort and learning, $\mathrm{Cov}(\mathrm{Y}_{\mathrm{G1},i,j,k}, \rho_{i,j,k}) > 0$, then we have an upward bias in β with finite class size (see Appendix A.4). Adopting the intuition in Angrist et al. (1999) and following the application in Chetty et al. (2011), we use a leave-out-mean specification. Removing child i from $\mathrm{X}_{j,k}$ breaks the (undesired) correlation because $\rho_{i,j,k}$ is independent of $\mathrm{X}_{(-i),j,k}$, but does attenuate the point estimate.[19] Therefore, the leave-out-mean addresses the bias due to unobserved preferences assuming:

Assumption 3 (No Peer Effects). *A child's unobserved preferences do not correlate with the teacher's overall capacity to form positive relationships,* $Cov(\theta_{j,k}, \rho_{i,j,k}) = 0$, *and unobserved preferences of child i do not affect the unobserved preferences of child i′,* $Cov(\rho_{i,j,k}, \rho_{i',j,k}) = 0$ *for* $i \neq i'$.

There is still a bias, however. Child i's unobserved preference, $\rho_{i,j,k}$, may correlate with $\mathrm{X}_{(-i),j,k}$ because our estimate is relative to the school mean and, consequently, $\rho_{i,j,k}$ correlates with ρ_k. Therefore, following Chetty et al. (2011), we also omit child i from the school mean and define $\widetilde{\mathrm{X}}_{-i,j,k} \equiv \mathrm{X}_{(-i),j,k} - \mathrm{X}_{(-i),k}$. We replace $\mathrm{X}_{j,k}$ in Equation (4) with $\widetilde{\mathrm{X}}_{(-i),j,k}$ and write:

$$\mathrm{Y}_{\mathrm{G1},i,j,k} = \mathbf{Y}'_{\mathrm{G0},i,j,k}\boldsymbol{\alpha} + b\widetilde{\mathrm{X}}_{(-i),j,k} + \mathbf{F}'_i\boldsymbol{\gamma} + \mathbf{P}'_{(-i),j,k}\boldsymbol{\delta} + \mathbf{C}'_{j,k}\boldsymbol{\kappa} + \tau_k + \epsilon_{i,j,k}. \tag{5}$$

Under Assumption 1, 2, and 3, we can interpret b as the effect of how your classmates perceive the teacher relative to the classmates (and hence

[19] In Appendix A.4, we follow Chetty et al. (2011) and use the within-class variance of $\mathrm{X}_{i,j,k}$ to estimate the extent of this attenuation bias at about five percent.

the teacher) you could have had if assigned to the other class in school (since we observe two classes in each school).[20] When we investigate the effect of teacher relationship skills on student learning, we thus estimate Equation (5) using a leave-out-mean instead of the mean to address the bias caused by children's unobserved preferences.

The assumption of no peer effects is strong. Peers can have positive and negative (spillover) effects, so it is difficult to characterize the magnitude of peers' (potential) effect. For example, a child may be so disruptive that he or she consumes all the teacher's attention, precluding the teacher from building positive relationships with the other children and helping them learn ("Bad Apple"). Alternatively, a child who favors the teacher may behave in such a way that causes the teacher to direct him or herself positively to the child. Peers may respond in kind, hoping for a similar response ("Shining Light").

We argue that the magnitude in which peers affect the preferences and perceptions of other peers is (likely) small in the early years of schooling. In our setup, any peer effect depends on how a child can influence other children's perceptions. Since we use the first grade of primary school, influential children are less of a concern. For most children, the transition to primary school marks a substantial change causing children to be preoccupied with the new environment.[21] Moreover, children in the first grade of primary school have a natural desire to please adult figures. They are more engaged to learn because the practical application of what they learn is apparent to them (Allen et al. 2011). These reasons suggest that children in the first grade mainly focus on themselves and what the teacher says. Lastly, children at this age can distinguish their preferences from the preferences of others (Fawcett & Markson 2010), which suggests that children may not conflate their preferences with the preferences of their peers. In sum, it seems reasonable to suspect that the more substantial part of the effect is due to the teacher and not from peers affecting the perceptions

[20] There is a finite sample bias in small groups due to the negative (mechanical) correlation between $X_{i,j,k}$ and $\widetilde{X}_{-i,j,k}$ (Guryan et al. 2009). Intuitively, a child cannot be his or her peer (*i.e.*, we sample without replacement). Therefore, the peers of a child who prefer the teacher comes from a group with slightly lower enthusiasm for the teacher, and vice versa. In Appendix A.5, we show results of a series of Monte Carlo simulations to examine the magnitude of this bias. The bias is negligible.

[21] We are grateful for an anonymous teacher for pointing this out to us.

and hence survey response of one another. Nevertheless, we cannot be sure and hence require caution for interpreting our results as solely caused by the teacher's overall capacity to form positive relationships. Although, one could argue that positive and negative spillover effects are both part of effective relationship management strategies maintained by the teachers.

One final point that requires elaboration is contemporaneity. In Equation (5), it is not clear whether children learn more because they benefit from the teacher relationship skills or evaluate the teacher positively in their capacity to develop relationships because they learn more. We exploit Assumption 3 to examine such simultaneity between $\mathrm{Y}_{\mathrm{G1},i,j,k}$ and $\mathrm{X}_{i,j,k}$. We regress the individual measure $\mathrm{X}_{i,j,k}$ on $\mathrm{Y}_{\mathrm{G1},i,j,k}$, a vector of child and family background variables, $\mathbf{F}_i'$, and a school-specific intercept, τ_k. The residual of this regression, $\hat{\epsilon}_{i,j,k}$, represents within-school variation not caused by $\mathrm{Y}_{\mathrm{G1},i,j,k}$, or the family background. We can then construct a leave-out-mean from the residual and use it as an instrument, which is valid under Assumption 3.

3.3. Descriptive Statistics

Table 4 presents a Pearson correlation matrix for each skill measured at the *start* of first grade (G0) and the *end* of first grade (G1). The correlation coefficients reveal two patterns. First, there is a strong correlation between the same skill over time. Second, we find strong correlations among academic skills and social-emotional skills and meaningful correlations between academic and social-emotional skills.

Panel A in Table 5 summarizes the family background variables. For 16.6 percent of the children, at least one family member experiences (self-reported) reading difficulties.[22] Also, 42.4 percent of the mothers and 55 percent of the fathers have less or equivalent to a high school degree. About 21.2 percent of the families have at least one parent born in a non-Western country. Finally, even though mothers attain more education on average, fathers earn almost double what mothers do. The average family income is NOK 990,817.

[22]Experienced reading difficulties do not necessarily imply diagnosed difficulties.

Table 4. Pearson Correlation Coefficient Matrix for Skills Measured at the Start and the End of First Grade

	At the Start of First Grade (G0)				At the End of First Grade (G1)			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A: At the Start of First Grade (G0)								
1. Math	1	0.52**	0.19**	−0.04**	0.51**	0.43**	0.13**	0.04**
2. Literacy	0.52**	1	0.33**	−0.03*	0.38**	0.56**	0.17**	0.08**
3. Self-Concept	0.19**	0.33**	1	0.26**	0.16**	0.17**	0.23**	0.19**
4. Reading Interest	−0.04**	−0.03*	0.26**	1	−0.00	−0.02*	0.07**	0.25**
Panel B: At the Start of End Grade (G0)								
5. Math	0.51**	0.38**	0.16**	−0.00	1	0.50**	0.19**	0.10**
6. Literacy	0.43**	0.56**	0.17**	−0.02*	0.50**	1	0.30**	0.14**
7. Self-Concept	0.13**	0.17**	0.23**	0.07**	0.19**	0.30**	1	0.41**
8. Reading Interest	0.04**	0.08**	0.19**	0.25**	0.10**	0.14**	0.41**	1

Notes. This table reports Pearson correlation coefficients for each skill at the *start* of first grade (G0) and the *end* of first grade (G1). The row numbers correspond to the column numbers. For example, Column (1), row two, shows the correlation between skill two and skill one (i.e., literacy and math, respectively). We use listwise deletion to handle missing values. Observations: 5,490.
$^{+}$ $p < 0.10$, * $p < 0.05$, and ** $p < 0.01$ (two-tailed).

We also measure several teacher and classroom characteristics. Panel B in Table 5 summarizes these variables. As is common in lower grades, a large share of the first-grade teachers is female: 97 percent. On average, the teachers in our sample have 13 to 14 years of experience, and most are between 30 and 59 years of age. Roughly 42 percent of the teachers are in the distribution's tails (below 25 or over 60). Furthermore, 5.8 percent of the teachers have an advanced degree (*i.e.*, a master's degree), while six percent do not have an undergraduate degree. On average, a class has 20 children, including about two who require special reading education.

Table 5. Sample Summary Statistics

	(1) Mean	(2) *SD*	(3) Obs.
Panel A: Child and Family Background Variables			
The Child is Female	47.8%		5,810
Family Reading Disability	16.6%		4,536
Siblings (no.)	1.6	1.1	5,782
Non-Western Immigrant	21.2%		5,704
Education Mother			5,474
· Not Completed High School	16.3%		892
· Only Completed High School	26.1%		1,428
· Any Higher Education	55.8%		3,055
Family Income in 2015 (NOK)	990,817	502,271	5,611
Panel B: Teacher and Classroom variables			
The Teacher is Female	96.6%		5,792
Experience (in Years)	13.5	8.5	5,764
The Teacher has an Advanced Degree	5.6%		5,810
The Teacher has an Undergraduate Degree	6.0%		5,810
The Teacher is over 50 Years of Age	28.1%		5,792
The Teacher is under 30 Years of Age	14.1%		5,792
Class Size (no.)	20.1	3.8	5,810
Special Education in Reading (no.)	1.9	1.7	5,775

Notes. This table reports summary statistics. For "education mother," we excluded the category "*any post-secondary education but not higher education*" from the table. This category represents 1.8 percent (99 observations).

As described previously, a key identifying assumption is that school administrators as-good-as randomly assign children and teachers to classes (within schools). Therefore, the availability of child and family background variables and skills measured at the *start* of first grade (G0) benefits our statistical approach in three ways: (i) we can examine if school administrators sort children into classes based these variables; (ii) we can condition in our model specification (Equation 5) on child and family background, $\mathbf{F}_i'$, and skills measured at the *start* of first grade (G0), $\mathbf{Y}_{\mathrm{G0},i,j,k}'$, to get more precise estimates of the effect; and (iii) the availability of these variables allows us to examine the sensitivity of our results to a consecutive inclusion of relevant control variables which, under as-good-as random assignment, should be minimal.[23]

3.4. Assignment to Classes

The validity of our empirical strategy relies on the assumption of as-good-as random assignment. By Norwegian law, school administrators should not group children based on sex, religion, ethnicity, or academic performance (Kunnskapdepartementet 2017). Despite this law, prior empirical evidence from the United States suggests that administrators may assign better teachers to better-performing students (Clotfelter et al. 2006).

Suppose the school administrators in our sample systematically assign better-performing children to higher-quality teachers. In that case, our identifying assumption is untenable, and we can no longer rule out explanations due to the within-school sorting of students. Therefore, we assess the plausibility of our identifying assumption by investigating if predetermined variables and variables measured at the *start* of first grade (G0) are predictive of class or teacher characteristics. The former cannot predict the latter under random assignment (Rothstein 2010). By contrast, if school administrators assign better-educated teachers to academically better students, to take a case in point, we should find a relationship between teacher education and skills measured at the *start* of first grade (G0). The results in Table 6 shed light on the class assignment based on observable characteristics.

[23]It is well known that sex, race, relative age differences, family income, and parental education are strong predictors of performance in school (Black et al. 2005, 2011, Dahl & Lochner 2012, Dee 2004, 2007, Solli 2017).

Table 6. Predictability of Predetermined Variables and Teacher and Classroom Characteristics

	(1) Teacher is Female	(2) Teacher Age	(3) Teacher Experience	(4) Teacher Education	(5) Special Education	(6) Class Size
Panel A: Child and Family Background Variables						
	0.000	0.020	0.060	−0.007	−0.005	−0.011
Child is Female	(0.003)	(0.015)	(0.104)	(0.004)	(0.017)	(0.020)
	[5,792]	[5,792]	[5,764]	[5,810]	[5,775]	[5,810]
	0.000	0.001	0.004	0.001	0.002	0.004
Birth month	(0.000)	(0.003)	(0.025)	(0.001)	(0.004)	(0.005)
	[5,784]	[5,784]	[5,756]	[5,802]	[5,767]	[5,802]
	−0.000	0.001	−0.028	−0.003	0.020^{+}	0.023
Siblings	(0.001)	(0.011)	(0.072)	(0.003)	(0.012)	(0.015)
	[5,764]	[5,764]	[5,736]	[5,782]	[5,747]	[5,782]
	−0.003	0.025	0.101	0.005	0.005	−0.064
Family Reading Disability	(0.006)	(0.037)	(0.289)	(0.011)	(0.054)	(0.064)
	[4,522]	[4,522]	[4,498]	[4,536]	[4,505]	[4,536]
	0.000	−0.010	−0.039	−0.002	0.009	0.051^{**}
Education Mother	(0.001)	(0.007)	(0.052)	(0.002)	(0.011)	(0.013)
	[5,459]	[5,459]	[5,431]	[5,474]	[5,440]	[5,474]
	−0.003	0.024	0.088	0.012	−0.020	−0.073
Non-Western Immigrant	(0.006)	(0.031)	(0.216)	(0.007)	(0.034)	(0.046)
	[5,687]	[5,687]	[5,659]	[5,704]	[5,670]	[5,704]

Continued on next page

Table 6. *Continued from previous page*

	(1) Teacher is Female	(2) Teacher Age	(3) Teacher Experience	(4) Teacher Education	(5) Special Education	(6) Class Size
	−0.000	0.003	0.108	−0.009	−0.008	0.092^{+}
Family Income	(0.004)	(0.025)	(0.174)	(0.006)	(0.031)	(0.048)
	[5,594]	[5,594]	[5,568]	[5,611]	[5,578]	[5,611]
Panel B: Academic Skills at the Start of First Grade (G0)						
	0.001	−0.016	−0.137	0.005	0.007	−0.006
Math	(0.003)	(0.012)	(0.085)	(0.004)	(0.016)	(0.020)
	[5,640]	[5,640]	[5,612]	[5,656]	[5,621]	[5,656]
	0.001	−0.001	−0.013	0.004	−0.013	0.011
Literacy	(0.002)	(0.012)	(0.087)	(0.004)	(0.015)	(0.018)
	[5,690]	[5,690]	[5,662]	[5,708]	[5,673]	[5,708]
Joint Significance	0.945	1.425	0.848	0.682	1.165	2.471^{*}

Notes. This table reports point estimates related to the predictability of predetermined and premeasured variables. Each cell reports an estimate of a separate regression, including a school-specific intercept and an indicator for treatment status. We regress the variables presented in the columns alternately on the variables in the rows. Family income is in NOK 1,000,000. "Teacher age" is a categorical variable that includes five groups: under 25; between 25 and 39; between 30 and 39; between 40 and 49; between 50 and 59; and 60 or higher. "Teacher education" is also a categorical variable that includes four groups: upper secondary school, university, or college education (less than three years); bachelor's degree (undergraduate); and master's degree (advanced degree). Special education in reading measures the number of children the teacher thinks require special education in reading. We cluster the standard errors (in parentheses) at the school level. We report the number of observations in brackets. We use listwise deletion to deal with missing data. $^{+}$ $p < 0.10$, * $p < 0.05$, and ** $p < 0.01$ (two-tailed).

In particular, in Table 6, we regress teacher's sex, age, experience, education, the number of children for whom the teacher thinks they require specialized training in reading, and class size alternately on each of the predetermined (Panel A) and premeasured variables (Panel B). Each cell reports an estimate of a separate regression, including a school-specific intercept and an indicator for treatment status.[24] We also run a series of *F* tests to test the joint significance of the variables that should not have an effect under as-good-as random assignment. Overall, the results are consistent with our identifying assumption.[25] The imbalances we find are not material (*e.g.*, increasing family income by one million Norwegian Kroner increase class size by 0.09 children). Importantly, while the findings in Table 6 are consistent with as-good-as random assignment based on observables, school administrators may still sort children into classes based on unobservables. Since we have no source of clean exogenous variation, we cannot rule out such selection on unobservables (*e.g.*, teachers' ability to handle children with behavioral problems).

4. Empirical Results

As mentioned in Section 3.1, some teachers changed during first grade (about 8.7 percent). It is not clear whether the children who experienced a teacher change had the old or new teacher in mind when asked about the teacher's relationship skills. In our main analysis, we include all children. In Appendix A.6, we present results similar to those reported here but with the children who experienced a teacher change during first grade excluded. Excluding these children does not change our conclusion.

Table 7 reports the teacher relationship skill estimates based on Equation (5). We start by running a regression model in which we only control for mean differences between schools (Column 1). In Columns (2) through

[24]Note that we do not correct the standard errors for multiple hypothesis testing. Table 6 includes 54 *t*-tests. With random sampling, one would expect to find at least two or three significant at the five percent significance level.

[25]In Appendix A.6, we provide further evidence that is consistent with our identifying assumption. We examine randomization into peer groups. Finally, we borrow from Ammermueller & Pischke (2009) and Clotfelter et al. (2006), and run a series of Pearson Chi-square tests of homogeneity. All results are consistent with our identifying assumption.

(5), we consecutively condition on: child and family background, skills measured at the *start* of first grade (G0), peer composition, and teacher and classroom characteristics to assess the sensitivity of our estimates. As explained above, in Column (6), we present estimates that account for simultaneity. Our preferred model specification is Column (5), as it best resembles the education production function. To provide some intuition for the effect sizes, consider that the difference in math and literacy test scores between students for whom the mother has a college degree and students from mothers without a college degree is about 30 percent of a standard deviation.

The effect of teacher relationship skills on math test scores is positive and statistically significant (Panel A in Table 7). In our preferred model specification (Column 5), a one standard deviation increase in teacher relationship skills increases math test scores by about 4.6 percent of a standard deviation. On the other hand, the effect is only about 2.7 percent of a standard deviation for literacy.[26] The impact of teacher relationship skills on students' interest in reading is positive and statistically significant (Panel B in Table 7). A one standard deviation increase in teacher relationship skills increases children's reading interest by about 4.9 percent of a standard deviation in our preferred specification. Finally, the coefficients for self-concept likewise suggest a positive impact of teacher relationship skills. A one standard deviation increase in teacher relationship skills improves students' self-concept in reading by about 3.4 percent of a standard deviation. Note that the leave-out-mean attenuates the point estimates by about five percent (see Appendix A.4 for details).

These estimates on literacy and self-concept are consistent with Jensen et al. (2019). They also use the Two Teachers data and find that the children's individually perceived emotional support correlates with reading test scores and self-concept in reading. However, this correlation is also consistent with other mechanisms. For example, making progress in reading may increase a child's feeling of emotional support. Alternatively, some unobserved child-level factors may result in higher individual perceived emotional support and better reading performance (*e.g.*, a sense of belonging at school could make it easier to feel connected and learn).

[26]We also estimated our models using the national literacy assessment as outcome. The results using the national assessments (not reported) present a similar story.

Table 7. The Relationship Between Teacher Relationship Skills and Student Learning

	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Academic Skills at the End of First Grade (G1)						
Math	0.055**	0.044**	0.045**	0.045**	0.046**	0.042**
	(0.016)	(0.017)	(0.016)	(0.016)	(0.016)	(0.016)
Adjusted R^2	0.088	0.160	0.373	0.373	0.373	0.373
Observations	5,610	5,610	5,610	5,610	5,610	5,605
Literacy	0.039*	0.029+	0.025+	0.023	0.027+	0.027*
	(0.016)	(0.016)	(0.015)	(0.015)	(0.014)	(0.014)
Adjusted R^2	0.103	0.190	0.439	0.440	0.440	0.439
Observations	5,637	5,637	5,637	5,637	5,637	5,633
School-Specific Intercept	✓	✓	✓	✓	✓	✓
Child and Family Background		✓	✓	✓	✓	✓
Initial Skills (G0)			✓	✓	✓	✓
Peer Composition				✓	✓	✓
Teacher and Classroom					✓	✓

Continued on next page

Table 7. *Continued from previous page*

	(1)	(2)	(3)	(4)	(5)	(6)
Panel B: Social-Emotional Skills the End of First Grade (G1)						
Reading Interest	0.056**	0.055**	0.052**	0.048**	0.049**	0.040*
	(0.017)	(0.018)	(0.018)	(0.017)	(0.017)	(0.016)
Adjusted R^2	0.027	0.042	0.047	0.047	0.049	0.049
Observations	5,630	5,630	5,630	5,630	5,630	5,629
Self-Concept	0.034*	0.035*	0.034*	0.030*	0.034*	0.039**
	(0.014)	(0.015)	(0.014)	(0.015)	(0.016)	(0.015)
Adjusted R^2	0.035	0.045	0.072	0.072	0.072	0.072
Observations	5,636	5,636	5,636	5,636	5,636	5,633
School-Specific Intercept	✓	✓	✓	✓	✓	✓
Child and Family Background		✓	✓	✓	✓	✓
Initial Skills (G0)			✓	✓	✓	✓
Peer Composition				✓	✓	✓
Teacher and Classroom					✓	✓

Notes. This table reports the effect of teacher relationship skills on student learning outcomes (*i.e.*, math, literacy, reading interest, and self-concept). Child and family background includes sex, birth month, birth year, the number of siblings, dummies for mother's education, family reading disability, non-Western immigrant, and family income (quartic family income polynomial). Initial skill level (G0) includes the scores for math and literacy measured at the *start* of first grade (G0). Peer composition consists of all child and family background variables and initial skill level variables measured as class-level leave-out-means. Teacher and classroom variables include the teacher's sex, age, education, experience, and class size. We cluster the standard errors (in parentheses) at the school level. All models include an indicator for treatment status and variables that predict missingness (see Appendix A.2 for details).

$^{+}$ $p < 0.10$, * $p < 0.05$, and ** $p < 0.01$ (two-tailed).

Curiously, the teacher relationship skills have a more substantial effect on math than literacy. While curious, such differential effects across math and literacy are not uncommon, and several explanations have been put forth in the literature. For example, Bettinger (2012) points out that math test scores may be more elastic. Many parents read with their children even before formal schooling, while most children learn math exclusively in school. In other words, the home environment contributes to the child's literacy skills more than to the child's math skills. As a result, there might be more to gain in math when school starts. Although this theory seems plausible, the distributional plots in our Appendix A.1 are not necessarily consistent with this story. The plots show that the distribution for literacy test scores (see, *e.g.*, reading fluency) centers around zero, more so than math test scores.

Another explanation could be that extrinsic motivation is more effective for math (Bettinger 2012). We described above that the relationship between and among the teacher and children might provide a feeling of competence, independence, and relatedness, increasing motivation and effort (Connell & Wellborn 1991). Since the teacher's relationship skills are external causes for children's motivation to learn, we would expect more substantial effects on math.

4.1. Assessing the Plausibility of Confounders

To substantiate that unobserved causes are not likely driving our results, we investigate the existence of (potential) confounders. This section provides evidence supporting the exogeneity of the teacher relationship skill assumption (Assumption 2).

4.1.1. Placebo Analysis

We do a placebo analysis to examine if there are confounding effects that are not consistent with our identification strategy. In Table 8, we regress math and literacy measured at the *start* of first grade (G0) alternately on the teacher relationship skills measured at the *end* of first grade (G1). As noted above, the school already started when we assessed the children. Therefore, we would not expect to find precise zeros. The purpose of the exercise is to investigate if any result is substantially larger than one can reasonably expect after a few weeks of schooling. For example, we suppose that social-emotional skills can quickly increase. A child's self-concept (in reading) may improve near instantaneously as soon as a child can compare him or herself to other children. As a result, doing a placebo analysis on social-emotional skills may not be informative for examining our identification strategy. By contrast, a child's ability to solve arithmetic problems may not necessarily improve in the span of a few weeks. The results in Table 8 do not falsify our model, suggesting that the effect of teacher relationship skills is not a placebo effect due to the children's level of literacy and math skills at the *start* of first grade (G0).

Table 8. Falsification Tests

	(1) Math	(2) Literacy	(3) Teacher Relationship Skills
Teacher Relationship Skills	−0.001 (0.012)	0.009 (0.014)	
Treatment Indicator			0.037 (0.161)
Observations	5,656	5,708	5,810
School-Specific Intercept	✓	✓	✓
Child and Family Background	✓	✓	✓
Initial Skills (G0)			✓
Peer Composition			✓
Teacher and Classroom			✓

Notes. This table reports falsification tests. In Column (1), we report estimates obtained from regressing math measured at the *start* of first grade (G0) on teacher relationship skills measured at the *end* of first grade (G1). In Column (2), we report estimates obtained from regressing literacy measured at the *start* of first grade (G0) on teacher relationship skills measured at the *end* of first grade (G1). In Column (3), we report estimates obtained from regressing teacher relationship skills (leave-out-mean specification) on the treatment. In addition to the controls indicated by the check marks, the regression reported in Column (3) also includes variables that predict missingness (see Appendix A.2). We cluster the standard errors (in parentheses) at the school level.
$^{+}$ $p < 0.10$, * $p < 0.05$, and ** $p < 0.01$ (two-tailed).

As explained in Section 2.2, we use data from a randomized controlled trial (see Solheim et al. 2017). Since the treatment had started before we measured the teacher relationship skills, we may worry that the treatment affected both the teacher relationship skills and student learning outcomes. Therefore, in Table 8 (Column 3), we regress our teacher relationship skills measure (leave-out-mean) on the treatment indicator and the control variables of our preferred specification. Consistent with the mechanism investigation in Haaland et al. (2021), our analysis suggests that the treatment does not affect the teacher's capacity to form positive relationships with and among the students. We also estimate our preferred model specification without controlling for treatment status (see Appendix A.6). The results are nearly identical.

4.1.2. Sensitivity Analysis

The gradual inclusion of control variables in Table 7 is, in a sense, a type of sensitivity analysis. Given the rich set of control variables, the stability of our point estimates and standard errors suggest that the results are not likely sensitive. Nevertheless, to formally substantiate this conclusion, we conduct a sensitivity analysis following Imbens (2003) and apply the procedures described in Harada (2013). We run a sensitivity analysis to understand how strong a confounding effect needs to be to change teacher relationship skills by half.

The intuition behind the sensitivity analysis is as follows. Confounding effects depend on two parameters. First, the parameter that captures the relationship between the confounding variable and the outcome of interest. Second, the parameter that captures the relationship between the confounding variable and the predictor of interest (*i.e.*, the teacher relationship skills). Using the procedure described in Harada (2013), we generate various "pseudo-unobservables" from the residual deviance, compute the two partial effect parameters, and plot the estimates on a curve. Like Imbens (2003), we then add the partial effect of observed covariates to the plot. We determine the magnitude of confounders comparatively to the skills measured at the *start* of first grade (G0) since those have plausibly the most impact on skills measured at the *end* of first grade (G1). We borrow this intuition from Altonji et al. (2005), who evaluate unobservables relative to the most impactful observables. The analysis suggests that our point estimates are not likely to be sensitive to unobserved confounders as they would have to have an effect stronger than skills measured at the *start* of first grade (G0) (see Appendix A.6 for the plots).

4.1.3. Robustness Analysis

We assess the robustness of our findings by estimating the effect of the teacher relationship skills on math, literacy, self-concept, and reading interest measured at the *end* of the second (G2) and the *end* of third grade (G3) (see Appendix A.6). The purpose for this investigation is to understand whether the effects of first grade represent an anomaly or whether we can find effects in later grades. Naturally, there is a concern that the magnitude drops because of the increasing time interval. Moreover, while we exploit the variation in first grade to estimate the effect of teacher relationship skills, the assigned teacher may have changed for practical reasons (*e.g.*, maternity leave) as children progress through school. Consequently, the analyses for second and third grade are more noisy.

The findings reveal patterns consistent with the estimated effects in Table 7, especially in second grade (G2). However, the magnitudes do drop, which is the case even more so in third grade (G3). This drop is especially prevalent for social-emotional skills. On the other hand, academic skills remain more stable, especially literacy. In sum, these findings suggest that the main results in Table 7 are not necessarily an anomaly of the first grade.

As a second robustness check, we construct the leave-out-mean measure of teacher relationship skills, but instead of an arithmetic average, we use factor analytic methods. When a set of items relate strongly as a group, researchers often employ factor analytic methods to separate the underlying latent construct, identified from the covariation among the items, from unique item-specific variation (and to reduce dimensionality). In Appendix A.3, we first conduct an exploratory and confirmatory factor analysis, following the analyses in the web appendices of Heckman et al. (2013) and Attanasio et al. (2020). We explore the number of factors we can reliably extract from the items. We can reliably extract three factors from the 11 items in each assessment wave. Next, we conduct an exploratory factor analysis. Based on the (rotated) loadings, we can conceive the factors as (i) the relationship between teacher and child, (ii) the positive relationships among the children, and (iii) a gloomy classroom atmosphere. Then, we conduct a confirmatory factor analysis, which shows a good model fit, and compute the variation attributable to the latent factor relative to the items' total variation to get an idea of the items' unique variation. Finally, we assign values to the factors, construct a leave-out-mean using the procedures described in Section 3, and estimate the models in Table 7. The results are comparable to using the an arithmetic average.

As a final robustness check, we estimate the effect of teacher relationship skills using a control function approach instead of the leave-out-mean. A control function method (also known as two-stage residual inclusion) adds a variable to the model that controls for the endogeneity caused by students' idiosyncratic preferences (see Wooldridge 2015, on control function approaches). In our case, the endogeneity addressed by the leave-out-mean is children's unobserved preferences for a particular teacher that affects their evaluation and learning outcomes.

By stacking the 11 items, we can contribute variation to different components. The school component captures variation between schools. The class component captures variation within schools, but between classes, the student component captures variation within schools and classes, but between students. The error term captures variation within schools, classes, and students but between items. We can assign values to each component by calculating the best linear unbiased predictions. In the second stage, we can include these predictions in Equation (5) instead of $\widetilde{\mathrm{X}}_{(-i),j,k}$, where we hold the student component fixed when we interpret the effect of the class component (see Appendix A.7 for details). The results of the control function method further support our initial findings. The estimates do increase in size. A one standard deviation increase in the teacher relationship skills increases math, literacy, reading interest, and self-concept, respectively, by seven, four, nine, and six percent of a standard deviation.

4.2. Differential Effects

We examine differential effects by sex and socioeconomic status (SES) using our preferred specification in Table 9. We operationalize socioeconomic status by constructing an 8-point ranking using parental education and family income. We then compute a geometric average and divide the sample into low (below the average) and high SES (above the average).

Columns (1) and (2) in Table 9 present the effect of teacher relationship skills by sex. Except for self-concept, we observe that boys benefit more from the teacher relationship skills, especially for reading interest. None of the differences is statistically significant, however. In Columns (4) and (5), we observe that children from low SES households benefit more than children from high SES households in all skills, especially for math and reading interest. However, we cannot rule out that these differences result from noise as none of the differences are statistically significant.

4.3. Informing Teacher Hiring and Training

A large literature suggests no consistent relationship between readily observable characteristics (*e.g.*, education, experience, salary) and student performance (see, *e.g.*, Hanushek 2003). The absence of a consistent relationship has motivated economists to borrow from psychology and education science and focus on what goes on inside the classroom (*e.g.*, emotional support, instructional practices, classroom management), using classroom observations (*e.g.*, Araujo et al. 2016, Kane et al. 2011). The findings from this literature are promising and have sparked debates on what teacher skills schools should focus on when hiring and evaluating teachers (*e.g.*, Jacob et al. 2018, Stewart et al. 2021).

Hiring and evaluating teachers is an important aspect of public education (Goldhaber et al. 2017, Jacob et al. 2018). First, there is substantial variation in teacher effectiveness (even within the same school), suggesting room for improvement. Second, effective teachers have long-term impacts on students' education and labor market outcomes. Finally, teachers commonly represent the largest budgetary expense in schools (Hanushek & Rivkin 2006). Understanding the type of information that improves hiring and job performance evaluations is thus imperative. Indeed, research suggests that policies that provide more information about job candidates, such as emphasizing important characteristics in application materials, could improve teacher quality (Goldhaber et al. 2017, Jacob et al. 2018). This section seeks to expand on how our approach to measuring teachers' overall capacity to form positive relationships using student reports might inform teacher hiring, evaluation, and training.[27]

[27]We thank the co-editor, Dan Kreisman, for this suggestion.

Table 9. Differential Effects by Sex and Socioeconomic Status

	Sex			Socioeconomic Status		
	(1) Boys	(2) Girls	(3) Wald Test	(4) Low	(5) High	(6) Wald Test
Panel A: Math at the End of First Grade (G1)						
Teacher Relationship Skills	0.044^{+}	0.039^{+}	0.1	0.066^{**}	0.029	1.7
	(0.022)	(0.020)		(0.024)	(0.022)	
Observations	2,920	2,690		2,300	3,310	
Panel B: Literacy at the End of First Grade (G1)						
Teacher Relationship Skills	0.031^{+}	0.019	0.3	0.029	0.020	0.1
	(0.018)	(0.019)		(0.025)	(0.015)	
Observations	2,935	2,702		2,313	3,324	
Panel C: Reading Interest at the End of First Grade (G1)						
Teacher Relationship Skills	0.060^{*}	0.028	1.2	0.079^{**}	0.037^{+}	1.7
	(0.024)	(0.022)		(0.026)	(0.021)	
Observations	2,930	2,700		2,307	3,323	

Continued on next page

Table 9. *Continued from previous page*

	Sex			Socioeconomic status		
	(1) Boys	(2) Girls	(3) Wald Test	(4) Low	(5) High	(6) Wald Test
Panel D: Self-Concept at the End of First Grade (G1)						
Teacher Relationship Skills	0.028	0.032^{+}	0.0	0.035	0.025	0.1
	(0.022)	(0.019)		(0.027)	(0.020)	
Observations	2,933	2,703		2,313	3,323	

Notes. This table reports the point estimates of our preferred specification (*i.e.*, Column 5 in Table 7). Columns (3), (6), and (9) present the Chi-square statistic of a Wald test for testing the difference between sub-populations. We include a school-specific intercept, an indicator for treatment status, a vector of child and family background variables, a vector of variables that measure initial skills (G0), a vector with variables that capture peer composition, and a vector of variables that measure teacher and classroom variables. In addition to these controls, the regressions also include variables that predict missingness (see Appendix A.2 for details). We cluster the standard errors (in parentheses) at the school level.
$^{+}$ $p < 0.10$, * $p < 0.05$, and ** $p < 0.01$ (two-tailed).

In Table 10, we present several Pearson correlation coefficients between readily observable teacher characteristics (*e.g.*, education, experience), teacher practices that are likely unobserved by school administrators (*i.e.*, teacher relationship skills), and student outcomes. We observe that none of the readily observable teacher characteristics strongly correlate with teacher relationship skills aside from teacher education. Second, the strength of the correlation coefficients between students' learning outcomes and teacher variables is highest for teacher relationship skills. Curiously, while teacher education correlates strongly with teacher relationship skills, it does not correlate with student learning outcomes.

In Table 11, we use our preferred specification (*i.e.*, Column (5) in Table 7) to compare these teacher variables formally within our model. Consistent with a smaller yet growing literature (*e.g.*, Araujo et al. 2016, Kane et al. 2011), teacher relationship skills impact student learning. Furthermore, consistent with a large body of literature on teacher characteristics (Hanushek 2003), we find that readily observable teacher characteristics are not predictive of student learning. Parenthetically, the significant effect reported for teachers with upper secondary education is due to two teachers. It follows, then, that the teacher's capacity to form positive relationships, measured using student reports, may provide more information about candidates and could improve teacher quality.[28]

We can consider teacher relationship skills measured using student reports as (potentially) relevant information for hiring decisions, teacher job performance evaluations, and teacher feedback. A key benefit of our approach to measuring teachers' overall capacity to form positive relationships is the relative ease of data collection. It is feasible to evaluate teacher relationship skills several times per academic year. It follows that teachers can evaluate their performance, receive regular feedback on their capacity to form positive relationships compared to themselves and others, and use such reports for future job applications. Although such a measure is beneficial, we ultimately want to evaluate teachers using different perspectives (Kane & Staiger 2012).

[28] See also the results on temporal stability of teacher relationship skills in Table 2, the variability of teacher relationship skills in Table 3, and the effects of teacher relationship skills on math, literacy, reading interest, and self-concept in Table 7.

Table 10. Pearson Correlation Coefficient Matrix Teacher and Classroom Variables

	(1)	(2)	(3)	(4)	(5)	(6)
1. Teacher Relationship Skills	1	0.02	0.04**	0.03**	0.12**	–0.07**
2. Teacher is Female	0.02	1	0.05**	0.05**	0.03*	–0.00
3. Teacher Age	0.04**	0.05**	1	0.76**	–0.17**	–0.09**
4. Teacher Experience	0.03**	0.05**	0.76**	1	–0.17**	–0.06**
5. Teacher Education	0.12**	0.03*	–0.17**	–0.17**	1	0.04**
6. Class Size	–0.07**	–0.00	–0.09**	–0.06**	0.04**	1
Math	0.07**	0.01	–0.00	0.01	0.01	0.03*
Literacy	0.07**	0.03*	0.01	0.03*	0.00	0.02+
Reading Interest	0.12**	0.01	0.04**	0.03**	0.00	–0.01
Self-Concept	0.08**	–0.00	0.02	0.03*	–0.01	–0.02

Notes. This table reports Pearson correlation coefficients between the teacher and classroom characteristics and student learning. The row numbers correspond to the column numbers. For example, Column (1), row three, shows the correlation between teacher relationship skills and teacher age. Teacher relationship skills are measured measured as the classroom average from the standardized (student-level) composites. We first standardize the items presented in Table 1 to be mean zero and standard deviation one. Second, we average these standardized items for each child. Third, we compute the average of these student-level standardized averages at the class level. We use listwise deletion to handle missing values. Observations: 5,561.

$^{+}$ $p < 0.10$, * $p < 0.05$, and ** $p < 0.01$ (two-tailed).

Table 11. Comparing Readily Observable Teacher Characteristics and Teacher Relationship Skills

	(1) Math	(2) Literacy	(3) Reading Interest	(4) Self-Concept
Panel A: Teacher Relationship Skills				
Teacher Relationship Skills	0.046**	0.027+	0.049**	0.034*
	(0.016)	(0.014)	(0.017)	(0.016)
Panel B: Readily Observable Teacher Characteristics				
Teacher is Female	0.010	0.123	0.026	−0.106
	(0.130)	(0.079)	(0.127)	(0.157)
Teacher Age (Base: Under 25)				
25 – 29	0.067	0.152	0.202	0.106
	(0.153)	(0.163)	(0.154)	(0.157)
30 – 39	0.147	0.087	0.191	0.125
	(0.155)	(0.154)	(0.146)	(0.153)
40 – 49	0.167	0.126	0.362*	0.103
	(0.141)	(0.154)	(0.155)	(0.152)
50 – 59	0.153	0.163	0.351*	0.049
	(0.160)	(0.156)	(0.153)	(0.156)
60 or Older	0.165	0.201	0.487*	0.169
	(0.159)	(0.162)	(0.190)	(0.173)

Continued on next page

Table 11. *Continued from previous page*

	(1) Math	(2) Literacy	(3) Reading Interest	(4) Self-Concept
Teacher Experience	−0.002	0.003	−0.005	0.003
	(0.003)	(0.003)	(0.004)	(0.004)
Teacher Education (Base: Bachelor Degree)				
Upper Secondary	0.873**	0.150	0.628**	0.695**
	(0.137)	(0.126)	(0.216)	(0.154)
No Bachelor Degree	0.016	−0.109	−0.205+	−0.091
	(0.092)	(0.096)	(0.107)	(0.097)
Master Degree	−0.015	0.056	−0.154	−0.143
	(0.088)	(0.076)	(0.112)	(0.091)
Observations	5,610	5,637	5,630	5,636
School-Specific Intercept	✓	✓	✓	✓
Child and Family Background	✓	✓	✓	✓
Initial Skills (G0)	✓	✓	✓	✓
Peer Composition	✓	✓	✓	✓

Notes. This table reports the point estimates from OLS regressions. Child and family background includes sex, birth month, birth year, the number of siblings, dummies for mother's education, family reading disability, non-Western immigrant, and family income (quartic family income polynomial). Initial skills (G0) includes the scores for math and literacy measured at the *start* of first grade (G0). Peer composition consists of all family background variables and initial skill level variables specified as leave-out-means. We cluster the standard errors (in parentheses) at the school level. All models include variables that predict missingness (see Appendix A.2 details). For teacher sex and age, we miss data on one teacher. For teacher experience, we miss data on three teachers. For these missing teachers, we replace the missing values with the median of the school.

$^{+}$ $p < 0.10$, * $p < 0.05$, and ** $p < 0.01$ (two-tailed).

Most teacher training involves pedagogic theory. Consequently, items such as those used in this study do not necessarily measure aspects of which the importance is unknown to the teacher. While teachers are likely aware, we did estimate substantial between-class variation within schools (see Table 3). Given the large degree of unity in teacher training, such variation is curious. It may suggest an interaction between other unobserved teacher skills (*e.g.*, social-emotional skills) and the teachers' capacity to form positive relationships. Teachers' social-emotional skills may be a fruitful avenue for future research as a determinant of teacher quality. Alternatively, such variation may indicate that the degree of seriousness by which teachers approach relationships with and among the students varies.

5. Conclusion

Numerous studies document a salient variation in teacher quality, yet our understanding of the drivers is minimal. We introduce and validate a new approach for measuring teachers' overall capacity to form positive relationships in the classroom, relying on student survey items previously developed and validated (at the student level) in the education literature. Building on earlier work by Araujo et al. (2016), Kane et al. (2011), and Rockoff & Speroni (2010), we present evidence that teacher relationship skills affect math, literacy, reading interest, and self-concept in reading. These effects may arise because the relationship the teacher builds with and among the students provides them with a feeling of security, competence, independence, and relatedness that increases their effort and learning achievements (Connell & Wellborn 1991). Furthermore, boys and children born into low socioeconomic households seem to benefit more from positive classroom relationships. However, the differences across sex and socioeconomic background are not statistically significant.

While we document consistent effects, we still require caution when interpreting our findings. That is, we can only conceive the teacher as the sole cause of our findings under assumptions that may not hold. For example, random assignment to classrooms does not preclude random assignment of teacher attributes to teachers (*e.g.*, didactic). As a result, the assumption that the teacher's overall capacity to form positive relationships is exogenous conditional on the observables may not hold. Furthermore, we only observe the same teacher with the same group of students. As a result, we cannot conclude that the estimated effects are solely due to the teacher, as there might be peer effects from sources other than the teacher. For instance, an overly disruptive child, who despises the teacher, may affect how other children perceive the teacher independently of teacher behavior changes. Future research that uses student reports may thus benefit from observing the same teacher with multiple classes to disentangle peer effects from teacher effects.

While our findings provide some interesting insights concerning classroom observations from the student's perspective, there are also limitations. First, the items used in this study are likely context-specific. That is to say, the questions asked to students may be less appropriate at middle or high school. Second, some assumptions, such as Assumption 3 of no peer effects, may not be tenable once children age. Teachers may develop a particular reputation that persists and is hard to change. In other words, student reports should be evaluated considering multiple performance indicators, not in isolation. Finally, we warrant more research as our sample had only 300 teachers. Moreover, we relied on within-school variation with only two classes per school. This small number of classes within schools likely affected the precision associated with the comparisons in Table 11.

In conclusion, our findings illustrate the benefit of student reports, with which we further substantiate the importance of using multiple measures from different perspectives to research and evaluate teacher quality. While prior research used trained (external) observers to examine teacher quality, we focus on what the children say. Kane & Staiger (2012) acknowledge the importance of multiple measures to evaluate teachers. Test scores provide a quantification of performance; asking children how they perceive the relationship with the teacher and other children is equally important. We show that such information is predictive of learning outcomes. As a result, while test scores enable identifying excellent and poor teachers, teacher relationship skills measures allow us to understand why some teachers are potentially more effective than others and how we may replicate excellent teachers.

References

Aaronson, D., Barrow, L. & Sander, W. (2007), 'Teachers and student achievement in the Chicago public high schools', *Journal of Labor Economics* **25**(1), 95–135.

Allen, J. P., Pianta, R. C., Gregory, A., Mikami, A. Y. & Lun, J. (2011), 'An interaction-based approach to enhancing secondary school instruction and student achievement', *Science* **333**(6045), 1034–1037.

Almlund, M., Duckworth, A. L., Heckman, J. & Kautz, T. (2011), Personality psychology and economics, *in* 'Handbook of the Economics of Education', Vol. 4, Elsevier B.V., Chapter 1, pp. 1–181.

Altonji, J. G., Elder, T. E. & Taber, C. R. (2005), 'Selection on observed and unobserved variables: Assessing the effectiveness of catholic schools', *Journal of Political Economy* **113**(1), 151–184.

Ammermueller, A. & Pischke, J. S. (2009), 'Peer effects in european primary schools: Evidence from the progress in international reading literacy study', *Journal of Labor Economics* **27**(3), 315–348.

Angrist, J. D., Imbens, G. W. & Krueger, A. B. (1999), 'Jackknife instrumental variables estimation', *Journal of Applied Econometrics* **14**(1), 57–67.

Araujo, M. C., Carneiro, P., Cruz-Aguayo, Y. & Schady, N. (2016), 'Teacher quality and learning outcomes in kindergarten', *Quarterly Journal of Economics* **131**(3), 1415–1453.

Attanasio, O., Cattan, S., Fitzsimons, E., Meghir, C. & Rubio-Codina, M. (2020), 'Estimating the production function for human capital: Results from a randomized controlled trial in Colombia', *American Economic Review* **110**(1), 48–85.

Begrich, L., Fauth, B. & Kunter, M. (2020), 'Who sees the most? differences in students' and educational research experts' first impressions of classroom instruction', *Social Psychology of Education* **23**(3), 673–699.

Bentler, P. M., Jackson, D. N. & Messick, S. (1971), 'Identification of content and style: A two-dimensional interpretation of acquiescence', *Psychological Bulletin* **76**(3), 186–204.

Bettinger, E., Ludvigsen, S., Rege, M., Solli, I. F. & Yeager, D. (2018), 'Increasing perseverance in math: Evidence from a field experiment in norway', *Journal of Economic Behavior and Organization* **146**, 1–15.

Bettinger, E. P. (2012), 'Paying to learn: The effect of financial incentives on elementary school test scores', *Review of Economics and Statistics* **94**(3), 686–698.

Black, S. E., Devereux, P. J. & Salvanes, K. G. (2005), 'Why the apple doesn't fall far: Understanding intergenerational transmission of human capital', *American Economic Review* **95**(1), 437–449.

Black, S. E., Devereux, P. J. & Salvanes, K. G. (2011), 'Too young to leave the nest? the effects of school starting age', *Review of Economics and Statistics* **93**(2), 455–457.

Borghans, L., Duckworth, A. L., Heckman, J. J. & ter Weel, B. (2008), 'The economics and psychology of personality traits', *Journal of Human Resources* **43**(4), 972–1059.

Cattell, R. B. (1966), 'The scree test for the number of factors', *Multivariate Behavioral Research* **1**(2), 245–276.

Chapman, J. W. & Tunmer, W. E. (1995), 'Development of young children's reading self-concepts: An examination of emerging subcomponents and their relationship with reading achievement', *Journal of Educational Psychology* **87**(1), 154–167.

Chapman, J. W., Tunmer, W. E. & Prochnow, J. E. (2000), 'Early reading-related skills and performance, reading self-concept, and the development of academic self-concept: A longitudinal study', *Journal of Educational Psychology* **92**(4), 703–708.

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

Chetty, R., Friedman, J. N. & Rockoff, J. E. (2014*a*), 'Measuring the impacts of teachers I: Evaluating bias in teacher value-added estimates', *American Economic Review* **104**(9), 2593–2632.

Chetty, R., Friedman, J. N. & Rockoff, J. E. (2014*b*), 'Measuring the impacts of teachers II: Teacher value-added and student outcomes in adulthood', *American Economic Review* **104**(9), 2633–2679.

Clotfelter, C. T., Ladd, H. F. & Vigdor, J. L. (2006), 'Teacher-student matching and the assessment of teacher effectiveness', *Journal of Human Resources* **41**(4), 778–820.

Coddington, C. S. & Guthrie, J. T. (2009), 'Teacher and student perceptions of boys' and girls' reading motivation', *Reading Psychology* **30**(3), 225–249.

Connell, J. P. & Wellborn, J. G. (1991), *Competence, Autonomy, and Relatedness: A Motivational Analysis of Self-System Processes*, The Minnesota Symposia on Child Psychology, Volume 23., Lawrence Erlbaum Associates, Inc, Hillsdale, New Jersey, pp. 43–77.

Cronbach, L. J. (1951), 'Coefficient alpha and the internal structure of tests', *Psychometrika* **16**(3), 297–334.

Cunha, F. & Heckman, J. (2007), 'The technology of skill formation', *American Economic Review* **97**(2), 31–47.

Dahl, G. B. & Lochner, L. (2012), 'The impact of family income on child achievement: Evidence from the earned income tax credit', *American Economic Review* **102**(5), 1927–1956.

Dee, T. S. (2004), 'Teachers, race, and student achievement in a randomized experiment', *Review of Economics and Statistics* **86**(1), 195–210.

Dee, T. S. (2007), 'Teachers and the gender gaps in student achievement', *Journal of Human Resources* **42**(3), 528–554.

Downer, J. T., Stuhlman, M., Schweig, J., Martínez, J. F. & Ruzek, E. (2015), 'Measuring effective teacher-student interactions from a student perspective: A multi-level analysis', *Journal of Early Adolescence* **35**(5-6), 722–758.

Eccles, J., Wigfield, A., Harold, R. D. & Blumenfeld, P. (1993), 'Age and gender differences in children's self- and task perceptions during elementary school', *Child Development* **64**(3), 830–847.

Fauth, B., Göllner, R., Gerlinde, L., Praetorius, A. & Wagner, W. (2020), 'Who sees what? conceptual considerations on the measurement of teaching quality from different perspectives', *Zeitschrift für Pädagogik* **66**(1), 138–155.

Fawcett, C. A. & Markson, L. (2010), 'Children reason about shared preferences', *Developmental Psychology* **46**(2), 299–309.

Gelman, A. & Hill, J. (2006), *Data Analysis Using Regression and Multilevel/Hierarchical Models*, Cambridge University Press, Cambridge, England.

Goldhaber, D., Grout, C. & Huntington-Klein, N. (2017), 'Screen twice, cut once: Assessing the predictive validity of applicant selection tools', *Education Finance & Policy* **12**(2), 197–223.

Goldhaber, D. & Hansen, M. (2010), 'Using performance on the job to inform teacher tenure decisions', *American Economic Review* **100**(2), 250–255.

Guryan, J., Kroft, K. & Notowidigdo, M. J. (2009), 'Peer effects in the workplace: Evidence from random groupings in professional golf tournaments', *American Economic Journal: Applied Economics* **1**(4), 34–68.

Haaland, V. F., Rege, M. & J., S. O. (2021), Complementary effects of increased teacher-student ratio and teacher professional development.

Hamre, B. K. & Pianta, R. C. (2001), 'Early teacher-child relationships and the trajectory of children's school outcomes through eighth grade', *Child Development* **72**(2), 625–638.

Hamre, B. K. & Pianta, R. C. (2005), 'Can instructional and emotional support in the first-grade classroom make a difference for children at risk of school failure?', *Child Development* **76**(5), 949–967.

Hanushek, E. A. (2003), 'The failure of input-based schooling policies', *Economic Journal* **113**(485), F64–F98.

Hanushek, E. A. & Rivkin, S. G. (2006), Teacher quality, *in* E. A. Hanushek & F. Welch, eds, 'Handbook of the Economics of Education', Vol. 2, Elsevier B.V., Amsterdam, Netherlands, Chapter 18, pp. 1051–1078.

Hanushek, E. A. & Rivkin, S. G. (2010), 'Generalizations about using value-added measures of teacher quality', *American Economic Review* **100**(2), 267–271.

Harada, M. (2013), Generalized sensitivity analysis. working paper, Report.

Heckman, J. J. & Kautz, T. (2012), 'Hard evidence on soft skills', *Labour Economics* **19**(4), 451–464.

Heckman, J. J., Stixrud, J. & Urzua, S. (2006), 'The effects of cognitive and noncognitive abilities on labor market outcomes and social behavior', *Journal of Labor Economics* **24**(3), 411–482.

Heckman, J., Pinto, R. & Savelyev, P. (2013), 'Understanding the mechanisms through which an influential early childhood program boosted adult outcomes', *American Economic Review* **103**(6), 2052–2086.

Holen, S., Waaktaar, T., Lervåg, A. & Ystgaard, M. (2013), 'Implementing a universal stress management program for young school children: Are there classroom climate or academic effects?', *Scandinavian Journal of Educational Research* **57**(4), 420–444.

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

Howes, C., Hamilton, C. E. & Matheson, C. C. (1994), 'Children's relationships with peers: Differential associations with aspects of the teacher-child relationship', *Child Development* **65**(1), 253–263.

Imbens, G. W. (2003), ‘Sensitivity to exogeneity assumptions in program evaluation’, *American Economic Review* **93**(2), 126–132.

Jackson, C. K. (2018), ‘What do test scores miss? the importance of teacher effects on non–test score outcomes’, *Journal of Political Economy* **126**(5), 2072–2107.

Jacob, B. A., Rockoff, J. E., Taylor, E. S., Lindy, B. & Rosen, R. (2018), ‘Teacher application hiring and teacher performance: Evidence from DC public schools’, *Journal of Public Economics* **166**, 81–97.

Jensen, M. T., Solheim, O. J. & Idsøe, E. M. C. (2019), ‘Do you read me? associations between perceived teacher emotional support, reader self-concept, and reading achievement’, *Social Psychology of Education* **22**(2), 247–266.

Kaiser, H. F. (1960), ‘The application of electronic computers to factor analysis’, *Educational and Psychological Measurement* **20**(1), 141–151.

Kane, T. J. & Staiger, D. O. (2012), Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains, Report.

Kane, T. J., Taylor, E. S., Tyler, J. H. & Wooten, A. L. (2011), ‘Identifying effective classroom practices using student achievement data’, *Journal of Human Resources* **46**(3), 587–613.

Klausen, T. & Reikerås, E. (2016), *Regnefaktaprøven [Assessment of Arithmetic Facts]*, Norwegian Reading Centre, Stavanger, Norway.

Klem, A. M. & Connell, J. P. (2004), ‘Relationships matter: Linking teacher support to student engagement and achievement’, *Journal of School Health* **74**(7), 262–273.

Kraft, M. A. (2019), ‘Teacher effects on complex cognitive skills and social-emotional competencies’, *Journal of Human Resources* **54**(1), 1–36.

Kunnskapdepartementet (2017), ‘Veiledning om organisering av elevene: Opplæringsloven 8–2 m.m.’.

Magnuson, K. & Duncan, G. J. (2016), 'Can early childhood interventions decrease inequality of economic opportunity?', *RSF: The Russell Sage Foundation Journal of the Social Sciences* **2**(2), 123–141.

Manski, C. F. (1993), 'Identification of endogenous social effects the reflection problem', *Review of Economic Studies* **60**(3), 531–542.

Norwegian Reading Centre (2013), *Lesesenterets Staveprøve [The Reading Centres Spelling Test]*, Norwegian Reading Centre, Stavanger, Norway.

OECD (2018), *Education at a Glance 2018: OECD Indicators*, OECD Publishing, Paris, France.

Opper, I. M. (2019), 'Does helping John help Sue? evidence of spillovers in education', *American Economic Review* **109**(3), 1080–1115.

Parker, J. G. & Asher, S. R. (1987), 'Peer relations and later personal adjustment: Are low-accepted children at risk?', *Psychological Bulletin* **102**(3), 357–389.

Pianta, R. C. (1997), 'Adult-child relationship processes and early schooling', *Early Education and Development* **8**(1), 11–26.

Pianta, R. C., LaParo, K. & Hamre, B. K. (2008), *Classroom Assessment Scoring System: Manual K-3*, Paul H. Brookes Publishing, Baltimore, Maryland.

Rauer, W. & Schuck, K. D. (2003), *Frageborgen zur Erfassung emotionaler und sozialer schulerfahrungen von Grundschulkindern dritter und vierter Klassen: FEESS 3-4: [Questionnaire to Assess Emotional and Social School Experiences of Elementary School Children, First and Second Classes (FEESS 3-4)].*, Beltz Test GMBH, Göttingen, Germany.

Rivkin, S. G., Hanushek, E. A. & Kain, J. F. (2005), 'Teachers, schools, and academic achievement', *Econometrica* **73**(2), 417–458.

Rockoff, J. E. (2004), 'The impact of individual teachers on student achievement: Evidence from panel data', *American Economic Review* **94**(2), 247–252.

Rockoff, J. E. & Speroni, C. (2010), ‘Subjective and objective evaluations of teacher effectiveness’, *American Economic Review* **100**(2), 261–266.

Roorda, D. L., Koomen, H. M. Y., Spilt, J. L. & Oort, F. J. (2011), ‘The influence of affective teacher-student relationships on students’ school engagement and achievement: A meta-analytic approach’, *Review of Educational Research* **81**(4), 493–529.

Rothstein, J. (2010), ‘Teacher quality in educational production: Tracking, decay, and student achievement’, *Quarterly Journal of Economics* **125**(1), 175–214.

Solheim, O. J., Rege, M. & McTigue, E. (2017), ‘Study protocol: “two teachers”: A randomized controlled trial investigating individual and complementary effects of teacher-student ratio in literacy instruction and professional development for reachers’, *International Journal of Educational Research* **86**, 122–130.

Solli, I. F. (2017), ‘Left behind by birth month’, *Education Economics* **25**(4), 323–346.

Statistics Norway (2018), *Facts about Education in Norway 2018 - Key Figures 2016*, Statistics Norway, Oslo, Norway.

Stewart, K., Sass, T. R. & Goldring, T. (2021), *Principals’ selection of teacher candidates*, Metro Atlanta Policy Lab for Education: Georgia Policy Labs, Atlanta, Georgia.

Torgesen, J. K., Rashotte, C. A. & Wager, R. K. (1999), *TOWRE: Test of Word Reading Efficiency*, Pro-ed, Inc., Austin, Texas.

Velicer, W. F. (1976), ‘Determining the number of components from the matrix of partial correlations’, *Psychometrika* **41**(3), 321–327.

Wooldridge, J. M. (2015), ‘Control function methods in applied econometrics’, *Journal of Human Resources* **50**(2), 420–445.

Supplementary Material

Supplement to "Teacher Relationship Skills and Student Learning": Appendices

By

Maximiliaan W. P. Thijssen
Mari Rege
Oddny J. Solheim

A.1. Further Notes on Data

The tables and figure in this section provide further details on the social-emotional and academic skills. For reading self-concept, we report the following tables. First, Table A.1.1 reports the answer frequencies related to self-concept in reading at the *start* of first grade (G0). Second, Table A.1.2 reports the answer frequencies related to self-concept in reading at the *end* of first grade (G1). Third, Table A.1.3 reports the answer frequencies related to self-concept in reading at the *end* of second grade (G2). Lastly, Table A.1.4 reports the answer frequencies related to self-concept in reading at the *end* of third grade (G3).

For reading interest, we report the following tables. First, Table A.1.5 reports the answer frequencies related to reading interest at the *start* of first grade (G0). Second, Table A.1.6 reports the answer frequencies related to reading interest at the *end* of first grade (G1). Third, Table A.1.7 reports the answer frequencies related to reading interest at the *end* of second grade (G2). Lastly, Table A.1.8 reports the answer frequencies related to reading interest at the *end* of third grade (G3).

For the academic skills (*i.e.*, math and literacy), we report summary statistics and approximate representations of the distributions. Table A.1.9 reports summary statistics (*i.e.*, number of observations, mean, standard deviation, minimum, and maximum) for assessment data. Figure A.1.1 presents probability frequency distributions for the assessment data.

Table A.1.1. Self-Concept in Reading Answer Frequencies: At the Start of First Grade (G0)

	Likert Scale			
	[1]	[2]	[3]	[4]
1. How many letters can you write that are in your name?	66 (1.1%)	159 (2.8%)	686 (11.8%)	4,775 (82.2%)
2. How easy is it to learn to read?	1,282 (22.1%)	1,777 (30.6%)	934 (16.1%)	1,693 (29.1%)
3. How many letters do you know that are in your name?	105 (1.8%)	207 (4.7%)	814 (14.0%)	4,497 (77.4%)
4. How easy is it to learn the letters?	477 (8.2%)	968 (16.7%)	1,256 (21.6%)	2,985 (51.4%)
5. How easy is it to read words?	989 (17.0%)	1,686 (29.0%)	1,166 (20.1%)	1,845 (31.8%)
6. How easy is it to write letters?	281 (4.84%)	496 (8.54%)	1,097 (18.88%)	3,812 (65.61%)
7. How many letters do you know compared to your classmates?	365 (6.3%)	921 (15.9%)	2,034 (35.0%)	2,365 (40.7%)

Notes. This table reports the absolute and relative (in parentheses) answer frequencies for each item of the self-concept in reading assessment at the *start* of first grade (G0). Cronbach's alpha is 0.73. The children answered on a 4-point Likert scale: [1] is low, and [4] is high. We presented the items through different smileys.

Table A.1.2. Self-Concept in Reading Answer Frequencies: At the End of First Grade (G1)

	Likert Scale				
	[1]	[2]	[3]	[4]	[5]
1. How good are you in letters?	35 (0.6%)	29 (0.5%)	322 (5.5%)	995 (17.1%)	4,254 (73.2%)
2. How good are you at reading?	66 (1.1%)	67 (1.2%)	647 (11.1%)	1,717 (29.6%)	3,138 (54.0%)
3. How good are you at reading long stories?	448 (7.7%)	531 (9.1%)	1,587 (27.3%)	1,846 (31.8%)	1,223 (21.1%)
4. How good are you at finding the meaning of difficult words on your own?	500 (8.6%)	712 (12.3%)	1,802 (31.0%)	1,505 (25.9%)	1,116 (19.2%)
5. How good are you in letters compared to your classmates?	106 (1.8%)	206 (4.5%)	2,318 (39.9%)	1,466 (25.2%)	1,486 (25.6%)
6. How good are you at reading compared to your classmates?	110 (1.9%)	236 (4.1%)	1,654 (28.5%)	1,608 (27.7%)	2,028 (34.9%)
7. Do you think reading is difficult?	159 (2.7%)	106 (1.8%)	1,416 (24.4%)	1,241 (21.4%)	2,714 (46.7%)
8. Do you think it is difficult to recognize the words you have read before?	309 (5.3%)	239 (4.1%)	946 (16.3%)	1,061 (18.3%)	3,081 (53.0%)
9. Do you think it is difficult to understand the meaning of words when you read?	323 (5.6%)	379 (6.5%)	1,875 (32.3%)	1,544 (26.6%)	1,515 (26.1%)

Notes. This table reports the absolute and relative (in parentheses) answer frequencies for each item of the self-concept in reading assessment at the *end* of first grade (G1). Cronbach's alpha is 0.72. The children answered on a 5-point Likert scale: [1] is low, and [5] is high. We presented the items through different smileys.

Table A.1.3. Self-Concept in Reading Answer Frequencies: At the End of Second Grade (G2)

	Likert Scale				
	[1]	[2]	[3]	[4]	[5]
1. How good are you at reading?	33 (0.6%)	43 (0.7%)	692 (11.9%)	1,994 (34.3%)	2,586 (44.5%)
2. How good are you at reading long stories?	162 (2.8%)	451 (7.8%)	1,684 (29.0%)	2,091 (36.0%)	960 (16.5%)
3. How good are you at reading compared to your classmates?	61 (1.1%)	367 (6.3%)	2,942 (50.6%)	1,367 (23.5%)	611 (10.5%)
4. Do you think reading is difficult?	60 (1.0%)	71 (1.2%)	1,444 (24.9%)	1,768 (30.4%)	2,005 (34.5%)
5. Do you think it is difficult to understand the meaning of words when you read?	93 (1.6%)	190 (3.3%)	2,110 (36.3%)	2,067 (35.6%)	888 (15.3%)

Notes. This table reports the absolute and relative (in parentheses) answer frequencies for each item of the self-concept in reading assessment at the *end* of second grade (G2). Cronbach's alpha is 0.70. The children answered on a 5-point Likert scale: [1] is low, and [5] is high. We presented the items through different smileys.

Table A.1.4. Self-Concept in Reading Answer Frequencies: At the End of Third Grade (G3)

	Likert Scale				
	[1]	[2]	[3]	[4]	[5]
1. How good are you at reading?	34 (0.6%)	56 (1.0%)	698 (12.0%)	2,322 (40.0%)	1,985 (34.2%)
2. How good are you at reading long stories?	103 (1.8%)	384 (6.6%)	1,616 (27.8%)	2,134 (36.7%)	859 (14.8%)
3. How good are you at reading compared to your classmates?	53 (0.9%)	460 (7.9%)	2,769 (47.7%)	1,371 (23.6%)	443 (7.6%)
4. Do you think reading is difficult?	34 (0.6%)	110 (1.9%)	1,423 (24.5%)	2,139 (36.8%)	1,390 (23.9%)
5. Do you think it is difficult to understand the meaning of words when you read?	32 (0.6%)	150 (2.6%)	1,871 (32.2%)	2,347 (40.4%)	695 (12.0%)

Notes. This table reports the absolute and relative (in parentheses) answer frequencies for each item of the self-concept in reading assessment at the *end* of third grade (G3). Cronbach's alpha is 0.73. The children answered on a 5-point Likert scale: [1] is low, and [5] is high. We presented the items through different smileys.

Table A.1.5. Reading Interest Answer Frequencies: At the Start of First Grade (G0)

	Likert Scale			
	[1]	[2]	[3]	[4]
1. How fun do you think it is to look at books and flip through pages yourself?	375 (6.5%)	477 (8.2%)	1,312 (22.6%)	3,522 (60.6%)
2. How fun do you think it is to visit the library?	411 (7.1%)	472 (8.1%)	1,122 (19.3%)	3,681 (63.4%)
3. How fun do you think it is when somebody reads for you at home?	292 (5.0%)	411 (7.1%)	1,036 (17.8%)	3,947 (67.9%)
4. How fun do you think it is to get a book as a present?	327 (5.6%)	467 (8.0%)	1,091 (18.8%)	3,801 (65.4%)
5. How fun do you think it is to look in books and flip through pages with others?	353 (6.1%)	445 (7.7%)	1,198 (20.6%)	3,690 (63.5%)
6. How fun do you think it is to look at comic books?	376 (6.5%)	638 (11.0%)	989 (17.0%)	3,683 (63.4%)

Notes. This table reports the absolute and relative (in parentheses) answer frequencies for each item of the reading interest assessment at the *start* of first grade (G0). Cronbach's alpha is 0.65. The children answered on a 4-point Likert scale: [1] is low, and [4] is high. We presented the items through different smileys.

Table A.1.6. Reading Interest Answer Frequencies: At the End of First Grade (G1)

	Likert Scale				
	[1]	[2]	[3]	[4]	[5]
1. Do you enjoy reading?	502 (8.6%)	167 (2.9%)	679 (11.7%)	1,145 (19.7%)	3,137 (54.0%)
2. Do you enjoy reading books?	481 (8.3%)	240 (4.1%)	816 (14.0%)	1,236 (21.3%)	2,857 (49.2%)
3. Do you enjoy reading comics?	941 (16.2%)	308 (5.3%)	793 (13.7%)	949 (16.3%)	2,639 (45.4%)
4. Do you enjoy reading at home?	488 (8.4%)	370 (6.4%)	1,252 (21.6%)	1,160 (20.0%)	2,360 (40.6%)
5. Do you usually look forward to reading?	928 (16.0%)	510 (8.8%)	1,356 (23.3%)	1,121 (19.3%)	1,715 (29.5%)
6. Would you be happy if you got a book as a present? (7.1%)	412 (2.5%)	147 (10.4%)	603 (16.9%)	983 (60.0%)	3,485
7. Do you think reading is boring?	556 (9.6%)	263 (4.5%)	1,148 (19.8%)	1,154 (19.9%)	2,509 (43.2%)

Notes. This table reports the absolute and relative (in parentheses) answer frequencies for each item of the reading interest assessment at the *end* of first grade (G1). Cronbach's alpha is 0.81. The children answered on a 5-point Likert scale: [1] is low, and [5] is high. We presented the items through different smileys.

Table A.1.7. Reading Interest Answer Frequencies: At the End of Second Grade (G2)

	Likert Scale				
	[1]	[2]	[3]	[4]	[5]
1. Do you enjoy reading?	343 (5.9%)	187 (3.2%)	840 (14.5%)	1,381 (23.8%)	2,583 (44.5%)
2. Do you enjoy reading books?	306 (5.3%)	262 (4.5%)	815 (14.0%)	1,476 (25.4%)	2,475 (42.6%)
3. Do you enjoy reading comics?	660 (11.4%)	399 (6.9%)	802 (13.8%)	1,130 (19.5%)	2,342 (40.3%)
4. Do you enjoy reading at home?	403 (6.9%)	432 (7.4%)	1,354 (23.3%)	1,381 (23.8%)	1,764 (30.4%)
5. Do you usually look forward to reading?	695 (12.0%)	561 (9.7%)	1,645 (28.3%)	1,340 (23.1%)	1,092 (18.8%)
6. Would you be happy if you got a book as a present?	278 (4.8%)	188 (3.2%)	776 (13.4%)	1,239 (21.3%)	2,853 (49.1%)
7. Do you think reading is boring?	354 (6.1%)	242 (4.2%)	1,461 (25.2%)	1,484 (25.5%)	1,793 (30.9%)

*Notes.*This table reports the absolute and relative (in parentheses) answer frequencies for each item of the reading interest assessment at the *end* of second grade (G2). Cronbach's alpha is 0.85. The children answered on a 5-point Likert scale: [1] is low, and [5] is high. We presented the items through different smileys

Table A.1.8. Reading Interest Answer Frequencies: At the End of Third Grade (G3)

	Likert Scale				
	[1]	[2]	[3]	[4]	[5]
1. Do you enjoy reading?	283 (4.9%)	248 (4.3%)	834 (14.4%)	1,535 (26.4%)	2,179 (37.5%)
2. Do you enjoy reading books?	219 (3.8%)	279 (4.8%)	818 (14.1%)	1,539 (26.5%)	2,224 (38.3%)
3. Do you enjoy reading comics?	433 (7.5%)	418 (7.2%)	847 (14.6%)	1,088 (18.7%)	2,292 (39.5%)
4. Do you enjoy reading at home?	395 (6.8%)	447 (7.7%)	1,225 (21.1%)	1,487 (25.6%)	1,525 (26.3%)
5. Do you usually look forward to reading?	595 (10.2%)	643 (11.1%)	1,603 (27.6%)	1,359 (23.4%)	879 (15.1%)
6. Would you be happy if you got a book as a present?	240 (4.1%)	230 (4.0%)	797 (13.7%)	1,371 (23.6%)	2,441 (42.0%)
7. Do you think reading is boring?	247 (4.3%)	272 (4.7%)	1,475 (25.4%)	1,785 (30.7%)	1,300 (22.4%)

Notes.This table reports the absolute and relative (in parentheses) answer frequencies for each item of the reading interest assessment at the *end* of third grade (G3). Cronbach's alpha is 0.87. The children answered on a 5-point Likert scale: [1] is low, and [5] is high. We presented the items through different smileys.

Table A.1.9. Summary Statistics Assessments

	(1) Obs.	(2) Mean	(3) *SD*	(4) Min.	(5) Max.
Panel A: Start of First Grade (G0)					
Spelling	5,652	2.02	3.06	0.00	10.00
Reading Fluency	5,632	2.65	5.77	0.00	75.00
Letter Recognition	5,695	16.58	6.43	0.00	24.00
Math	5,656	4.68	5.10	0.00	32.00
Panel B: End of First Grade (G1)					
Spelling	5,620	6.82	2.60	0.00	15.00
Reading Fluency	5,633	23.00	10.94	0.00	80.00
Reading Accuracy	5,629	30.75	8.87	0.00	36.00
Math	5,610	20.30	9.09	0.00	79.00
Panel C: End of Second Grade (G2)					
Spelling	5,317	5.45	3.72	0.00	28.00
Reading Fluency	5,324	37.51	13.78	0.00	89.00
Reading Accuracy	5,342	36.58	11.76	0.00	43.00
Reading Comprehension	5,313	11.87	6.60	0.00	32.00
Math	5,319	31.59	10.90	0.00	90.00
Panel D: End of Third Grade (G3)					
Spelling	5,070	6.39	4.22	0.00	15.00
Reading Fluency	5,077	48.13	13.66	0.00	92.00
Reading Accuracy	5,091	35.97	13.83	0.00	43.00
Reading Comprehension	5,076	14.90	8.05	0.00	35.00
Math	5,050	39.26	11.18	0.00	85.00

Notes. This table reports statistics related to the individual assessments. We base the summary statistics for spelling, letter recognition, reading accuracy, and reading comprehension on the total of the individual items.

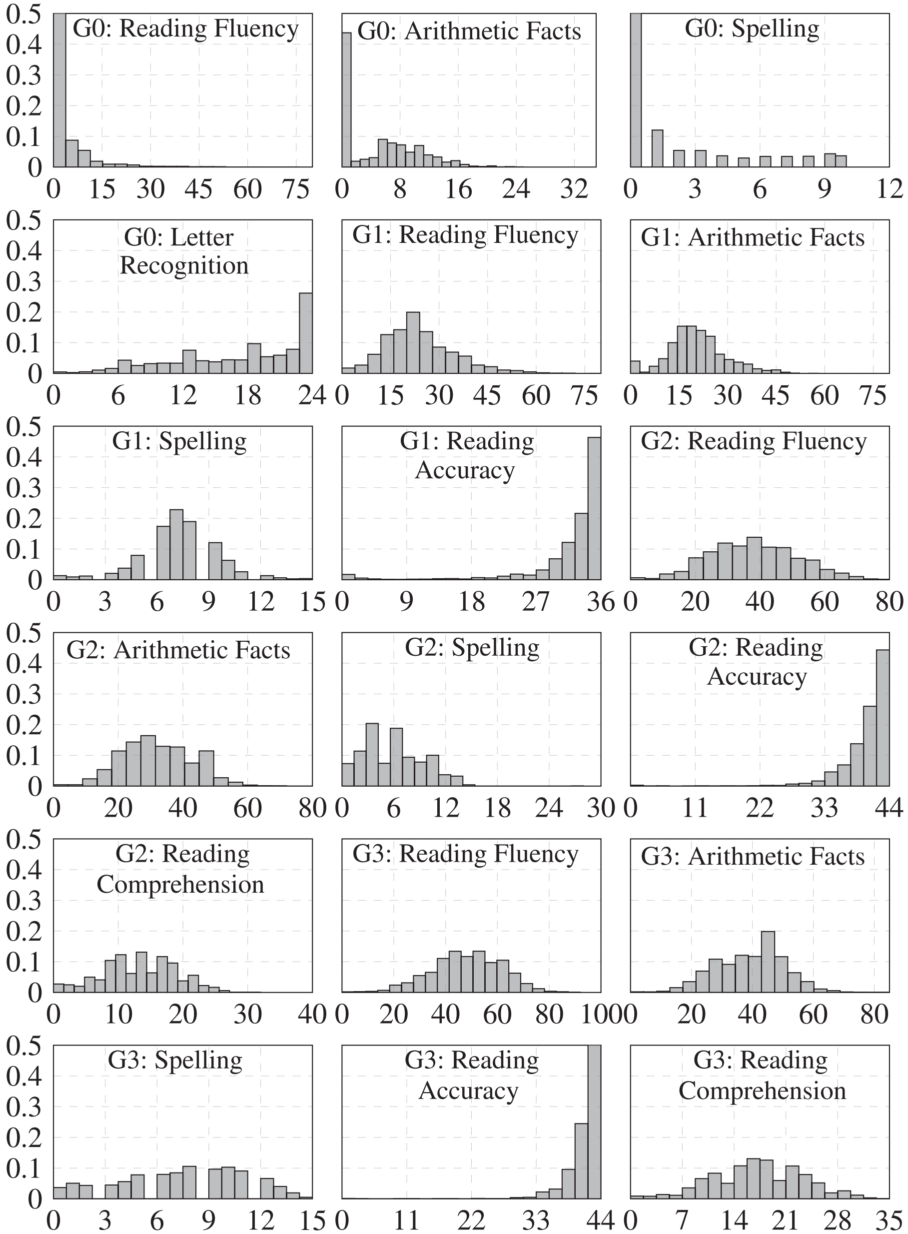

Figure A.1.1. Probability Frequency Distributions Assessment Variables

Notes. This figure reports probability frequency distributions for the assessment variables summarized in Table A.1.9. The plots show the probability (y-axis) of how often a particular score (x-axis) will happen. We use listwise deletion to handle missing values.

A.2. Further Notes on Missing Data

We assume the missing data are missing at random (*i.e.*, the probability a data point is missing does not depend on the value of the missing data point but only on available information). In other words, conditional on the observed data, we assume to have enough information to ignore the missing data mechanism. Formally, let X_i (for $i = 1, 2, \ldots, N$) denote an observation in the sample. We assume:

Assumption A.2.1 (Missing at Random). *Missing observations in X_i do not depend on their value but could depend on other observations $X_{i'}$, such that $Pr(X_i = missing \mid X_i, X_{i'}) = Pr(X_i = missing \mid X_{i'})$ for all $i \neq i'$.*

Given our rich data, we believe this to be a reasonable assumption. Under the Assumption A.2.1, it is acceptable to ignore missing cases (Gelman & Hill 2006, pp. 529–543), given that one controls for all relevant variables that may predict missingness.

We miss data in some of the variables provided by Statistics Norway, which is likely due to these families being immigrants to Norway. As a result, the government has not yet collected all records. For many of the parents, we know the country of birth. Consequently, conditional on the country of birth, and other variables we have on the family, we assume that the data provided by Statistics Norway are missing at random.

For the Two Teachers data, we also miss observations. Fortunately, we have detailed information on the assessments. First, for each tester, we observe a unique identifier. Some testers might be more able to make students complete the assessment than others when the student struggles. Consequently, testers may predict missingness. Second, as part of the assessments, testers documented whether students were ill/away, spoke a foreign language, moved to another place, and so forth. Conditional on knowing why testers did not assess students, it is likely that missing is at random. Lastly, we know the month and day on which the assessment occurred. It might be the case that specific dates disproportionally affected the likelihood of being assessed. Conditional on this information, we believe it is reasonable to assume that data are missing at random.

In sum, in addition to a dummy variable that controls for the treatment in the Two Teachers project, we add the following three variables in all our regressions.[29] We assume that the addition of these variables makes the missing at random assumption more plausible. First, dummies for testers who assessed the children at the *start* of first grade (G0). Second, dummies for the assessment background at the *start* of first grade (G0) (*e.g.*, students that were sick). Lastly, an interaction between the day and month of the assessment at the *start* of first grade (G0). Parenthetically, these variables are exogenous in the sense that they affect the probability of missingness in, for example, the skills measured at the *start* of first grade (G0), but are not affected by the teacher relationship skills measured at the *end* of first grade (G1). Figure A.2.2 gives an overview of percentage of missing data for each variable.

Since we do not have many missing observations, we employ the following strategy when we estimate the relationship between teacher relationship skills and student learning for the independent variables only. For categorical variables, we simply add another category that represents missingness. For continuous variables, like the teacher relationship skills, we impute all missing values with the median in the school. We include dummies that indicate missingness and add interactions between the variable of interest (*i.e.*, the teacher relationship skills) and the missingness indicator dummies of the other variables to allow the slope to be different for missing-data groups. See Gelman & Hill (2006, pp. 529–543) for more details and footnote 18 in Araujo et al. (2016) for a similar application. We also ran our models without the previously explained simple imputation. Aside from literacy, which drops in magnitude, the results are similar. Note, however, that we lose many observations and precision.

[29]For about two percent of the observations, we miss a tester identifier and the day of assessment. For those observations, we impute using the median of the class (rounded off). In most cases, testers assessed an entire class around the same time. Furthermore, often, a single tester assessed an entire class.

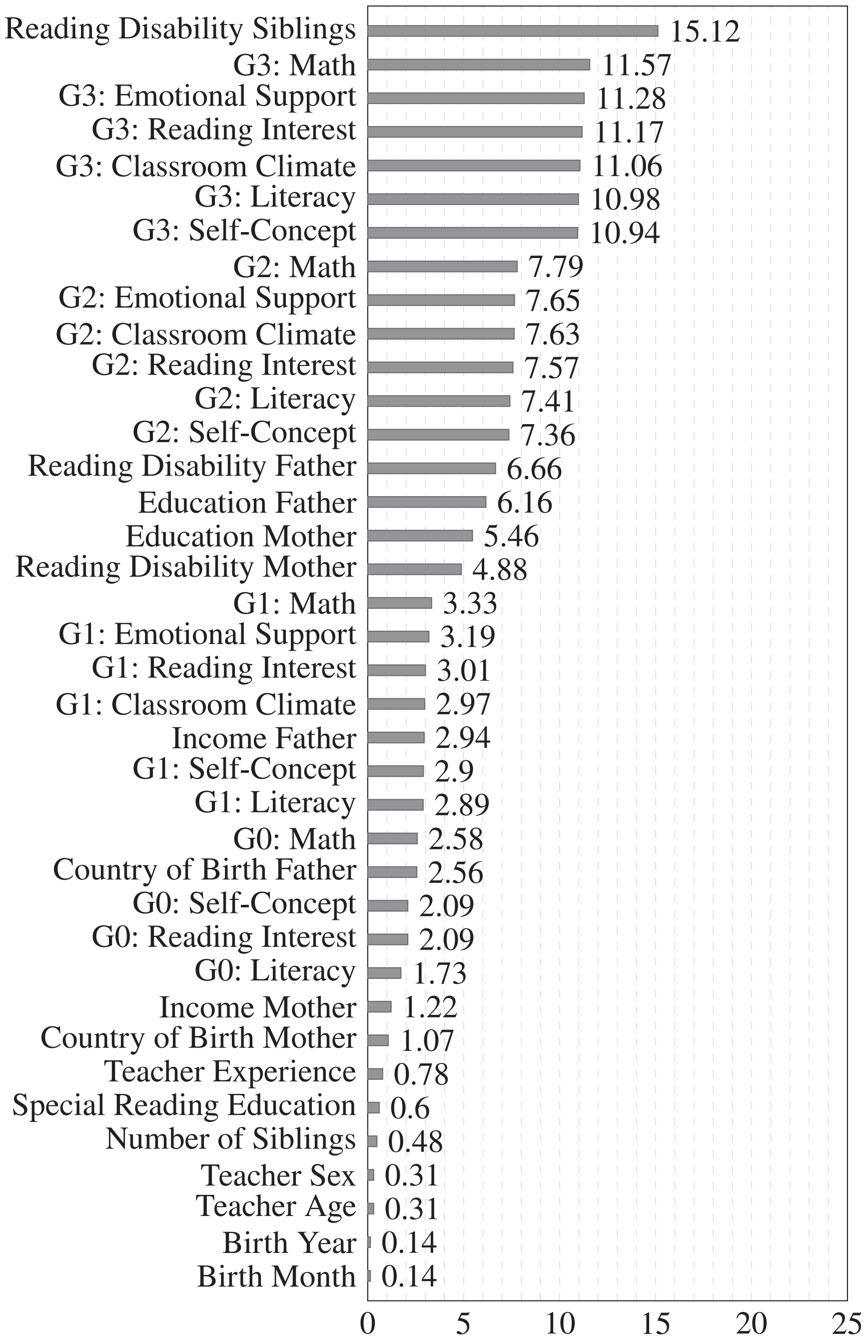

Figure A.2.2. Percentage of Missing Data for Each Variable

A.3. Further Notes on Teacher Relationship Skill Measures

In this appendix, we investigate the teacher relationship skill measure in more detail. First, we present some descriptives related to each of the items. In particular, Table A.3.10 reports the answer frequencies for questions related to emotional support at the *end* of first grade (G1). Second, Table A.3.11 reports the answer frequencies for questions related to emotional support at the *end* of second grade (G2). Third, Table A.3.12 reports the answer frequencies for questions related to emotional support at the *end* of third grade (G3). Fourth, Table A.3.13 reports the answer frequencies for questions related to classroom climate at the *end* of first grade (G1). Fifth, Table A.3.14 reports the answer frequencies for questions related to classroom climate at the *end* of second grade (G2). Sixth, Table A.3.15 reports the answer frequencies for questions related to classroom climate at the *end* of third grade (G3). Seventh, Table A.3.16 reports a polychoric correlation coefficients for the emotional support and classroom climate items at the *end* of first grade (G1). Polychoric correlations are commonly used to estimate associations between two assumed normally distributed continuous *latent* variables, from ordinal items. Eighth, Table A.3.17 reports a polychoric correlation coefficients for the emotional support and classroom climate items at the *end* of second grade (G2). Lastly, Table A.3.18 reports a polychoric correlation coefficients for the emotional support and classroom climate items at the *end* of third grade (G3).

In Table A.3.19, we evaluate the number of factors we could reasonably extract using Kaiser's eigenvalue rule (Kaiser 1960), Cattell's scree plot (Cattell 1966), Horn's parallel test (Horn 1965), and Velicer's minimum average partial correlation rule (Velicer 1976). The results in Table A.3.19 suggest that the data is rich enough for extracting three factors. In Table A.3.20, we perform an exploratory factor analysis using "quartimin" rotation. Quartimin rotation aims to rotate the loadings such that items load mostly on a single factor. The results in Table A.3.19 suggest that in each assessment wave, three factors underlie the 11 items. The first factor appears to relate mostly to the relationship between the student and the teacher. The second factor pertains to positive relationships among the students, and the final factor represents the negative relations among the students. Parenthetically, at the end of second grade, we see that "*Do you help each other in the class?*" loads on the second and third factors.

Table A.3.10. Answer Frequencies for Questions Related to Emotional Support: End of First Grade (G1)

	Likert Scale				
	[1]	[2]	[3]	[4]	[5]
1. Do you feel as if the teacher is a good friend of yours?	178 (3.2%)	115 (2.1%)	822 (14.6%)	1,215 (21.6%)	3,288 (58.5%)
2. Will the teacher help you when you have problems?	67 (1.2%)	47 (0.8%)	411 (7.3%)	801 (14.3%)	4,292 (76.4%)
3. Do you feel the teachers appreciate you?	281 (3.9%)	88 (1.6%)	568 (10.1%)	920 (16.4%)	3,824 (68.1%)
4. Do you get help from your teacher when reading is difficult?	233 (4.2%)	96 (1.7%)	591 (10.5%)	989 (17.6%)	3,702 (66.0%)

Notes.This table reports the absolute and relative (in parentheses) answer frequencies for each item related to emotional support at the *end* of first grade (G1). Cronbach's alpha is 0.62. The children answered on a 5-point Likert scale: [1] is low, and [5] is high.

Table A.3.11. Answer Frequencies for Questions Related to Emotional Support: End of Second Grade (G2)

	Likert Scale				
	[1]	[2]	[3]	[4]	[5]
1. Do you feel as if the teacher is a good friend of yours?	98	99	745	1,596	2,790
	(1.8%)	(1.9%)	(14.0%)	(30.0%)	(52.4%)
2. Will the teacher help you when you have problems?	18	36	297	942	4,036
	(0.3%)	(0.7%)	(5.6%)	(17.7%)	(75.7%)
3. Do you feel the teachers appreciate you?	60	55	365	1,088	3,760
	(1.1%)	(1.0%)	(6.9%)	(20.4%)	(70.6%)
4. Do you get help from your teacher when reading is difficult?	99	81	444	1,366	3,339
	(1.9%)	(1.5%)	(8.3%)	(25.6%)	(62.7%)

Notes. This table reports the absolute and relative (in parentheses) answer frequencies for each item related to emotional support at the *end* of second grade (G2). Cronbach's alpha is 0.69. The children answered on a 5-point Likert scale: [1] is low, and [5] is high.

Table A.3.12. Answer Frequencies for Questions Related to Emotional Support: End of Third Grade (G3)

	Likert Scale				
	[1]	[2]	[3]	[4]	[5]
1. Do you feel as if the teacher is a good friend of yours?	36 (0.7%)	84 (1.7%)	509 (10.0%)	1,691 (33.4%)	2,751 (54.3%)
2. Will the teacher help you when you have problems?	8 (0.2%)	23 (0.5%)	237 (4.7%)	1,048 (20.7%)	3,754 (74.0%)
3. Do you feel the teachers appreciate you?	26 (0.5%)	46 (0.9%)	269 (5.3%)	1,045 (20.6%)	3,685 (72.7%)
4. Do you get help from your teacher when reading is difficult?	101 (2.0%)	69 (1.4%)	406 (8.0%)	1,420 (28.0%)	3,074 (60.6%)

Notes.This table reports the absolute and relative (in parentheses) answer frequencies for each item related to emotional support at the *end* of third grade (G3). Cronbach's alpha is 0.72. The children answered on a 5-point Likert scale: [1] is low, and [5] is high.

Table A.3.13. Answer Frequencies for Questions Related to Classroom Climate: End of First Grade (G1)

	Likert Scale			
	[1]	[2]	[3]	[4]
1. Is everybody in the class good friends?	116	475	2,332	2,709
	(2.1%)	(8.4%)	(41.4%)	(48.1%)
2. Do you stick together and look after each other?	180	432	2,046	2,974
	(3.2%)	(10.5%)	(26.2%)	(55.7%)
3. Is there anyone in your class who laughs at children who are different?	429	589	1,476	3,138
	(7.6%)	(10.5%)	(26.2%)	(55.7%)
4. Do you help each other in the class?	111	256	1,857	3,408
	(2.0%)	(4.6%)	(33.0%)	(60.5%)
5. Are all the children in the class allowed to play along?	142	423	1,426	3,641
	(2.5%)	(7.5%)	(25.3%)	(64.7%)
6. Are there students in your class who make fun of the others?	375	777	2,039	2,441
	(6.7%)	(13.8%)	(36.2%)	(43.3%)
7. Do you tease and annoy one another in the class?	243	471	2,112	2,806
	(4.3%)	(8.4%)	(37.5%)	(49.8%)

Notes. This table reports the absolute and relative (in parentheses) answer frequencies for each item related to classroom climate at the *end* of first grade (G1). Cronbach's alpha is 0.69. The children answered on a 4-point Likert scale: [1] is low, and [4] is high.

Table A.3.14. Answer Frequencies for Questions Related to Classroom Climate: End of Second Grade (G2)

	Likert Scale			
	[1]	[2]	[3]	[4]
1. Is everybody in the class good friends?	49 (0.9%)	379 (7.1%)	2,852 (53.5%)	2,050 (38.5%)
2. Do you stick together and look after each other?	76 (1.4%)	369 (6.9%)	2,403 (45.1%)	2,482 (46.6%)
3. Is there anyone in your class who laughs at children who are different?	155 (2.9%)	493 (9.3%)	1,803 (33.9%)	2,876 (54.0%)
4. Do you help each other in the class?	31 (0.6%)	217 (4.1%)	2,323 (43.6%)	2,759 (51.8%)
5. Are all the children in the class allowed to play along?	62 (1.2%)	333 (6.3%)	1,674 (31.4%)	3,260 (61.2%)
6. Are there students in your class who make fun of the others?	143 (2.7%)	585 (11.0%)	2,560 (48.0%)	2,041 (38.3%)
7. Do you tease and annoy one another in the class?	88 (1.7%)	402 (7.5%)	2,713 (50.9%)	2,127 (39.9%)

Notes.This table reports the absolute and relative (in parentheses) answer frequencies for each item related to classroom climate at the *end* of second grade (G2). Cronbach's alpha is 0.74. The children answered on a 4-point Likert scale: [1] is low, and [4] is high.

Table A.3.15. Answer Frequencies for Questions Related to Classroom Climate: End of Third Grade (G3)

	Likert Scale			
	[1]	[2]	[3]	[4]
1. Is everybody in the class good friends?	20 (0.4%)	317 (6.2%)	3,087 (60.7%)	1,662 (32.7%)
2. Do you stick together and look after each other?	47 (0.9%)	290 (5.7%)	2,521 (49.6%)	2,228 (43.8%)
3. Is there anyone in your class who laughs at children who are different?	83 (1.6%)	403 (7.9%)	1,817 (35.7%)	2,783 (54.7%)
4. Do you help each other in the class?	27 (0.5%)	152 (3.0%)	2,399 (47.2%)	2,509 (49.3%)
5. Is there anyone in your class who laughs at children who are different?	33 (0.7%)	244 (4.8%)	1,628 (32.0%)	3,182 (62.6%)
6. Are there students in your class who make fun of the others?	79 (1.6%)	437 (8.6%)	2,506 (49.3%)	2,065 (40.6%)
7. Do you tease and annoy one another in the class?	66 (1.3%)	378 (7.4%)	2,897 (57.0%)	1,745 (34.3%)

Notes.This table reports the absolute and relative (in parentheses) answer frequencies for each item related to classroom climate at the *end* of third grade (G3). Cronbach's alpha is 0.76. The children answered on a 4-point Likert scale: [1] is low, and [4] is high.

Table A.3.16. Polychoric Correlations for the Teacher Relationship Skill Items (End of First Grade: G1)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
(1)	1	0.32	0.46	0.35	0.27	0.31	0.12	0.33	0.23	0.12	0.16
(2)	0.32	1	0.43	0.49	0.21	0.28	0.14	0.29	0.25	0.14	0.23
(3)	0.46	0.43	1	0.38	0.18	0.32	0.17	0.29	0.26	0.19	0.18
(4)	0.35	0.49	0.38	1	0.19	0.26	0.12	0.30	0.21	0.13	0.16
(5)	0.27	0.21	0.18	0.19	1	0.31	0.19	0.42	0.40	0.33	0.33
(6)	0.31	0.28	0.32	0.26	0.31	1	0.16	0.54	0.43	0.22	0.30
(7)	0.12	0.14	0.17	0.12	0.19	0.16	1	0.13	0.25	0.42	0.38
(8)	0.33	0.29	0.29	0.30	0.42	0.54	0.13	1	0.38	0.25	0.32
(9)	0.23	0.25	0.26	0.21	0.40	0.43	0.25	0.38	1	0.31	0.35
(10)	0.12	0.14	0.19	0.13	0.33	0.22	0.42	0.25	0.31	1	0.46
(11)	0.16	0.23	0.18	0.16	0.33	0.30	0.38	0.32	0.35	0.46	1

Notes. This table reports Polychoric correlation coefficients for the teacher relationship skill items at the *end* of first grade (G1). (1) Do you feel as if the teacher is a good friend of yours? (2) Will the teacher help you when you have problems? (3) Do you feel the teachers appreciate you? (4) Do you get help from your teacher when reading is difficult? (5) Is everybody in the class good friends? (6) Do you stick together and look after each other? (7) Is there anyone in your class who laughs at children who are different? (8) Do you help each other in the class? (9) Are all the children in the class allowed to play along? (10) Are there students in your class who make fun of the others? (11) Do you tease and annoy one another in the class?

Table A.3.17. Polychoric Correlations for the Teacher Relationship Skill Items (End of Second Grade: G2)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
(1)	1	0.41	0.53	0.41	0.31	0.36	0.16	0.32	0.24	0.20	0.26
(2)	0.41	1	0.53	0.53	0.30	0.32	0.19	0.33	0.31	0.24	0.25
(3)	0.53	0.53	1	0.46	0.27	0.34	0.21	0.31	0.27	0.24	0.25
(4)	0.41	0.53	0.46	1	0.24	0.27	0.17	0.29	0.23	0.19	0.22
(5)	0.31	0.30	0.27	0.24	1	0.44	0.27	0.45	0.43	0.41	0.40
(6)	0.36	0.32	0.34	0.27	0.44	1	0.21	0.55	0.40	0.33	0.39
(7)	0.16	0.19	0.21	0.17	0.27	0.21	1	0.14	0.31	0.48	0.38
(8)	0.32	0.33	0.32	0.29	0.45	0.55	0.14	1	0.34	0.26	0.36
(9)	0.24	0.31	0.27	0.23	0.43	0.40	0.31	0.34	1	0.39	0.41
(10)	0.20	0.24	0.24	0.19	0.41	0.33	0.48	0.25	0.39	1	0.50
(11)	0.26	0.25	0.26	0.22	0.40	0.39	0.38	0.36	0.41	0.50	1

Notes. This table reports Polychoric correlation coefficients for the teacher relationship skill items at the *end* of second grade (G2). (1) Do you feel as if the teacher is a good friend of yours? (2) Will the teacher help you when you have problems? (3) Do you feel the teachers appreciate you? (4) Do you get help from your teacher when reading is difficult? (5) Is everybody in the class good friends? (6) Do you stick together and look after each other? (7) Is there anyone in your class who laughs at children who are different? (8) Do you help each other in the class? (9) Are all the children in the class allowed to play along? (10) Are there students in your class who make fun of the others? (11) Do you tease and annoy one another in the class?

Table A.3.18. Polychoric Correlations for the Teacher Relationship Skill Items (End of Third Grade: G3)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
(1)	1	0.46	0.58	0.45	0.34	0.36	0.23	0.34	0.29	0.24	0.27
(2)	0.46	1	0.60	0.54	0.27	0.35	0.22	0.37	0.33	0.22	0.29
(3)	0.58	0.60	1	0.49	0.32	0.39	0.27	0.38	0.34	0.28	0.31
(4)	0.45	0.54	0.49	1	0.25	0.30	0.17	0.31	0.29	0.21	0.21
(5)	0.34	0.27	0.32	0.25	1	0.45	0.34	0.44	0.43	0.45	0.45
(6)	0.36	0.35	0.39	0.30	0.45	1	0.29	0.56	0.45	0.37	0.43
(7)	0.23	0.22	0.27	0.17	0.34	0.29	1	0.22	0.33	0.54	0.37
(8)	0.34	0.37	0.38	0.31	0.44	0.56	0.22	1	0.36	0.33	0.38
(9)	0.29	0.33	0.34	0.29	0.43	0.45	0.33	0.36	1	0.40	0.42
(10)	0.24	0.22	0.28	0.21	0.45	0.37	0.54	0.33	0.40	1	0.55
(11)	0.27	0.29	0.31	0.21	0.45	0.43	0.37	0.38	0.42	0.55	1

Notes. This table reports Polychoric correlation coefficients for the teacher relationship skill items at the *end* of third grade (G3). (1) Do you feel as if the teacher is a good friend of yours? (2) Will the teacher help you when you have problems? (3) Do you feel the teachers appreciate you? (4) Do you get help from your teacher when reading is difficult? (5) Is everybody in the class good friends? (6) Do you stick together and look after each other? (7) Is there anyone in your class who laughs at children who are different? (8) Do you help each other in the class? (9) Are all the children in the class allowed to play along? (10) Are there students in your class who make fun of the others? (11) Do you tease and annoy one another in the class?

Table A.3.19. Number of Factors to Extract

	(1) Kaiser's Eigenvalue Rule	(2) Cattell's Scree Plot	(3) Horn's Parallel Analysis	(4) Velicer's Minimum Average Partial Correlation
End of First Grade (G1)	3	3	3	1
End of Second Grade (G2)	3	3	2	2
End of Third Grade (G3)	2	3	2	2

Notes. Kaiser's eigenvalue criterion (Kaiser 1960) extract the number of factors with eigenvalues larger than 1. Cattell's scree plot (Cattell 1966): extract the number of factors before which the eigenvalues level off. Horn's parallel analysis (Horn 1965): extract the number of factors for which the eigenvalue is larger than the eigenvalue of randomly sampled samples. Velicer's minimum average partial correlation (Velicer 1976): extract the number of factors for which we find the smallest average squared partial correlation.

In Table A.3.21, for a given factor, we estimate the following linear-in-parameter specification:

$$\mathrm{X}_{l,i,t} = \lambda_{1,l,t} f_{i,t} + v_{l,i,t}, \tag{A.3.1}$$

where $\mathrm{X}_{l,i,t}$ is the lth observed item for child i at time t, $\lambda_{1,l,t}$ is the factor loading, $f_{i,t}$ is the (common) factor, and $v_{l,i,t}$ is the unique factor. We assume: (i) $\mathrm{Cov}(f_{i,t}, v_{l,i,t}) = 0 \ \forall \ l = 1, \ldots, L_t, t = 0, 1, \ldots, T$, (ii) $\mathrm{Cov}(v_{l,i,t}, v_{l',i,t'} | f_{i,t}) = 0 \ \forall \ l, l' = 1, \ldots, L_t, \ l \neq l', t, t' = 0, 1, \ldots, T$, and (iii) the unique factor is independent across children. Parenthetically, there is no intercept because we standardized all items to have mean zero and standard deviation one before the analysis. Without further normalization, we cannot identify the scale of the factor. Therefore, to ensure identification, we normalize the sum of factor loadings to be equal to the number of items assumed to measure $f_{i,t}$. If there are four items for a given factor, then we use the following normalization: $\lambda_{1,4,t} = 4 - \lambda_{1,3,t} - \lambda_{1,2,t} - \lambda_{1,1,t} \ \forall \ t = 0, 1, \ldots, T$. Table A.3.21, indicate a good fit. Furthermore, there is substantial variation and covariation in and among the factors.

In Table A.3.22, we compute signal-to-noise ratios. The variance of an item is: $\text{Var}(X_{l,i,t}) = (\lambda_{1,l,t})^2\text{Var}(f_{i,t}) + \text{Var}(\upsilon_{l,i,t})$. Consequently, the variation explained by the factor is:

$$\text{Signal}_{l,t} = \frac{(\lambda_{1,l,t})^2\text{Var}(f_{i,t})}{(\lambda_{1,l,t})^2\text{Var}(f_{i,t}) + \text{Var}(\upsilon_{l,i,t})}, \tag{A.3.2}$$

and the noise is $1 - \text{Signal}_{l,t}$. Finally, we assign values to the latent factors using regression scoring and construct leave-out-means (Table A.3.23). The results are similar to the main results.

Table A.3.20. Rotated Factor Loadings Based on Three Extracted Factors

	End of First Grade (G1)			End of Second Grade (G2)			End of Third Grade (G3)		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Items	$f_{1,i,1}$	$f_{2,i,1}$	$f_{3,i,1}$	$f_{1,i,2}$	$f_{2,i,2}$	$f_{3,i,2}$	$f_{1,i,3}$	$f_{2,i,3}$	$f_{3,i,3}$
(1)	0.49	–	–	0.58	–	–	0.63	–	–
(2)	0.64	–	–	0.69	–	–	0.73	–	–
(3)	0.63	–	–	0.72	–	–	0.76	–	–
(4)	0.63	–	–	0.67	–	–	0.67	–	–
(5)	–	0.50	–	–	0.48	–	–	0.44	–
(6)	–	0.64	–	–	0.64	–	–	0.66	–
(7)	–	0.68	–	–	0.68	–	–	0.67	–
(8)	–	0.48	–	–	*0.31*	*0.35*	–	0.33	–
(9)	–	–	0.60	–	–	0.64	–	–	0.65
(10)	–	–	0.62	–	–	0.67	–	–	0.72
(11)	–	–	0.54	–	–	0.50	–	–	0.47

Notes. This table reports the estimated factor loadings for a total of three factors. We hide factor loadings of less than 0.3. (1) Do you feel as if the teacher is a good friend of yours? (2) Will the teacher help you when you have problems? (3) Do you feel the teachers appreciate you? (4) Do you get help from your teacher when reading is difficult? (5) Is everybody in the class good friends? (6) Do you stick together and look after each other? (7) Is there anyone in your class who laughs at children who are different? (8) Do you help each other in the class? (9) Are all the children in the class allowed to play along? (10) Are there students in your class who make fun of the others? (11) Do you tease and annoy one another in the class?

Table A.3.21. Estimated Factor Loadings and Variance-Covariance Matrix of Factors

	End of First Grade (G1)			End of Second Grade (G2)			End of Third Grade (G3)		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	$f_{1,i,1}$	$f_{2,i,1}$	$f_{3,i,1}$	$f_{1,i,2}$	$f_{2,i,2}$	$f_{3,i,2}$	$f_{1,i,3}$	$f_{2,i,3}$	$f_{3,i,3}$
Panel A: Estimated Factor Loadings									
$\lambda_{1,1,t}$	1.00	0.90	0.86	0.97	1.02	0.84	0.99	1.00	0.85
	(0.03)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
$\lambda_{1,2,t}$	0.97	1.06	1.08	1.01	1.08	1.10	1.01	1.08	1.13
	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
$\lambda_{1,3,t}$	1.07	1.06	1.07	1.09	0.98	1.05	1.11	0.98	1.01
	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
$\lambda_{1,4,t}$	0.95	0.97		0.94	0.92		0.89	0.94	
	(0.02)	(0.02)		(0.02)	(0.02)		(0.02)	(0.02)	
Panel B: Variance-Covariance Matrix of Factors									
$f_{1,i,t}$	0.29	0.18	0.11	0.37	0.20	0.15	0.41	0.23	0.17
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
$f_{2,i,t}$	0.18	0.31	0.19	0.20	0.34	0.25	0.23	0.34	0.28
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
$f_{3,i,t}$	0.11	0.19	0.34	0.15	0.25	0.36	0.17	0.28	0.39
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)

Continued on next page

Table A.3.21. *Continued from previous page*

	End of First Grade (G1)			End of Second Grade (G2)			End of Third Grade (G3)		
	(1) $f_{1,i,1}$	(2) $f_{2,i,1}$	(3) $f_{3,i,1}$	(4) $f_{1,i,2}$	(5) $f_{2,i,2}$	(6) $f_{3,i,2}$	(7) $f_{1,i,3}$	(8) $f_{2,i,3}$	(9) $f_{3,i,3}$
Panel C: Model Fit Indices									
CFI[a]	0.947	(0.933)		0.956	(0.943)		0.967	(0.954)	
TLI[b]	0.944	(0.932)		0.954	(0.942)		0.965	(0.953)	
SRMR[c]	0.031	(0.041)		0.031	(0.043)		0.967	(0.954)	
RMSEA[d]	0.040	(0.044)		0.042	(0.047)		0.039	(0.045)	
Obs.		5,609			5,312			5,062	

Notes. This table reports the estimated factor loadings using parametric maximum likelihood. We standardize all variables before the analysis to have mean 0 and standard deviation 1. Panel A presents the loadings and standard errors (in parentheses). Panel B presents the variance-covariance matrix of the factors. Panel C presents various fit indices. The fit indices in parentheses are for a model in which we model a higher-order factor to account for the covariation between the three lower-order factors.

[a] Comparative Fit Index (CFI) ranges from zero to one; values close to one suggest a good fit.

[b] Tucker-Lewis Index (TLI) ranges from zero to one; values close to one suggest a good fit.

[c] Standardized Root Means Squared Residual (SRMR) ranges from zero to one; values close to zero suggest a good fit.

[d] Root Means Squared Error of Approximation (RMSEA) ranges from zero to one; values close to zero suggest a good fit.

Table A.3.22. Signal-To-Noise Ratios

	Factors: End of First Grade (G1)		Factors: End of Second Grade (G2)		Factors: End of Third Grade (G3)	
Items	(1) Signal	(2) Noise	(3) Signal	(4) Noise	(5) Signal	(6) Noise
(1)	0.30	0.70	0.34	0.66	0.40	0.60
(2)	0.27	0.73	0.37	0.63	0.41	0.59
(3)	0.34	0.66	0.43	0.57	0.50	0.50
(4)	0.27	0.73	0.32	0.68	0.32	0.68
(5)	0.25	0.75	0.35	0.65	0.34	0.66
(6)	0.35	0.65	0.39	0.61	0.40	0.60
(7)	0.35	0.65	0.32	0.68	0.32	0.68
(8)	0.30	0.70	0.28	0.72	0.30	0.70
(9)	0.25	0.75	0.26	0.74	0.29	0.71
(10)	0.39	0.61	0.44	0.56	0.51	0.49
(11)	0.38	0.62	0.40	0.60	0.40	0.60

Notes. This table reports the signal-to-noise ratios for each of the items. The signal is the fraction of the variance attributed to the factor over the total variance. The noise is the fraction of the variance attributed to measurement error over the total variance. (1) Do you feel as if the teacher is a good friend of yours? (2) Will the teacher help you when you have problems? (3) Do you feel the teachers appreciate you? (4) Do you get help from your teacher when reading is difficult? (5) Is everybody in the class good friends? (6) Do you stick together and look after each other? (7) Is there anyone in your class who laughs at children who are different? (8) Do you help each other in the class? (9) Are all the children in the class allowed to play along? (10) Are there students in your class who make fun of the others? (11) Do you tease and annoy one another in the class?

Table A.3.23. The Effect of Teacher Relationship Skills on Academic and Social-Emotional Skills at the End of First Grade Using Regression Scoring to Assign Values to Latent Variables

	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Academic Skills at the End of First Grade (G1)						
Math	0.054^{**}	0.043^{**}	0.044^{**}	0.044^{**}	0.045^{**}	0.042^{**}
	(0.016)	(0.016)	(0.016)	(0.016)	(0.016)	(0.015)
Adjusted R^2	0.088	0.160	0.373	0.373	0.373	0.373
Observations	5,610	5,610	5,610	5,610	5,610	5,605
Literacy	0.040^{*}	0.030^{+}	0.026^{+}	0.024^{+}	0.028^{+}	0.029^{*}
	(0.016)	(0.016)	(0.014)	(0.015)	(0.014)	(0.014)
Adjusted R^2	0.103	0.190	0.439	0.440	0.440	0.440
Observations	5,637	5,637	5,637	5,637	5,637	5,633
School-Specific Intercept	✓	✓	✓	✓	✓	✓
Child and Family Background		✓	✓	✓	✓	✓
Initial Skills (G0)			✓	✓	✓	✓
Peer Composition				✓	✓	✓
Teacher and Classroom					✓	✓

Continued on next page

Table A.3.23. *Continued from previous page*

	(1)	(2)	(3)	(4)	(5)	(6)
Panel B: Social-Emotional Skills the End of First Grade (G1)						
Reading Interest	0.059**	0.058**	0.054**	0.050**	0.051**	0.043**
	(0.017)	(0.018)	(0.018)	(0.016)	(0.016)	(0.016)
Adjusted R^2	0.027	0.042	0.047	0.047	0.050	0.050
Observations	5,630	5,630	5,630	5,630	5,630	5,629
Self-Concept	0.034*	0.034*	0.033*	0.029+	0.033*	0.039**
	(0.014)	(0.015)	(0.014)	(0.015)	(0.015)	(0.015)
Adjusted R^2	0.035	0.045	0.072	0.072	0.073	0.073
Observations	5,636	5,636	5,636	5,636	5,636	5,633
School-Specific Intercept	✓	✓	✓	✓	✓	✓
Child and Family Background		✓	✓	✓	✓	✓
Initial Skills (G0)			✓	✓	✓	✓
Peer Composition				✓	✓	✓
Teacher and Classroom					✓	✓

Notes. This table reports the point estimates from an OLS regression. We cluster the standard errors at the school level and present them in parentheses. Family background includes sex, birth month, birth year, the number of siblings, dummies for mother's education, family reading disability, non-Western immigrant, and family income (quartic family income polynomial). Initial skills (G0) includes the scores for math and reading measured at the *start* of first grade (G0). We also include a cubic polynomial. Peer composition consists of all family background variables and initial skill level variables specified as leave-out-means. Teacher and classroom variables include the teacher's sex, education, experience, and class size. All models include an indicator for treatment status and variables that predict missingness (see Appendix A.2 for details). We cluster the standard errors at the school level and present them in parentheses.
$^+$ $p < 0.10$, * $p < 0.05$, and ** $p < 0.01$ (two-tailed).

A.4. Further Notes on the Own-Observation Problem

Below, we follow Chetty et al. (2011). We keep all control variables implicit and write the true model as:

$$\mathrm{Y}_{\mathrm{G1},i,j,k} = \theta_{j,k} B + \rho_{i,j,k}\eta + \tau_k + \epsilon_{i,j,k}.$$

We do not observe $\theta_{j,k}$ or $\rho_{i,j,k}$. We observe $\mathrm{X}_{i,j,k}$ instead:

$$\mathrm{X}_{i,j,k} = \theta_{j,k} + \rho_{i,j,k}.$$

We estimate:

$$\mathrm{Y}_{\mathrm{G1},i,j,k} = \mathrm{X}_{j,k}\beta + \tau_k + \epsilon_{i,j,k}.$$

The class average of $\mathrm{X}_{i,j,k}$ is:

$$\mathrm{X}_{j,k} \equiv \frac{1}{N_{j,k}} \sum_{i=1}^{N_{j,k}} \mathrm{X}_{i,j,k} = \theta_{j,k} + \frac{1}{N_{j,k}} \sum_{i=1}^{N_{j,k}} \rho_{i,j,k}.$$

We assume there are no peer effects so that $\mathrm{Cov}(\theta_{j,k}, \rho_{i,j,k}) = 0$ and $\mathrm{Cov}(\rho_{i,j,k}, \rho_{i',j,k}) = 0$ for $i \neq i'$. Also, we assume $\mathrm{Cov}(\rho_{i,j,k}, \epsilon_{i,j,k}) = 0$ and $\mathrm{Cov}(\theta_{j,k}, \epsilon_{i,j,k}) = 0$. By demeaning the variables at the school level, we get rid of the school-specific intercept:

$$\mathrm{Y}_{\mathrm{G1},i,j,k} - \mathrm{Y}_{\mathrm{G1},k} = (\theta_{j,k} - \theta_k)B + (\rho_{i,j,k} - \rho_k)\eta + (\epsilon_{i,j,k} - \epsilon_k),$$

$$\mathrm{X}_{j,k} - \mathrm{X}_k = (\theta_{j,k} - \theta_k) + \left(\frac{1}{N_{j,k}} \sum_{i=1}^{N_{j,k}} \rho_{i,j,k} - \frac{1}{\sum_{j=1}^{J_k} N_{j,k}} \sum_{j=1}^{J_k} \sum_{i=1}^{N_{j,k}} \rho_{i,j,k} \right).$$

Below, we evaluate the consistency of the mean, and the leave-out-mean as the number of schools grows large $K \to \infty$. Let:

$$\rho_{j,k} \equiv \frac{1}{N_{j,k}} \sum_{i=1}^{N_{j,k}} \rho_{i,j,k}.$$

$$\rho_k \equiv \frac{1}{\sum_{j=1}^{J_k} N_{j,k}} \sum_{j=1}^{J_k} \sum_{i=1}^{N_{j,k}} \rho_{i,j,k}.$$

We can then write:

$$\begin{aligned}
\text{plim } \hat{\beta}_{K\to\infty} &= \frac{\text{Cov}(\text{Y}_{\text{G1},i,j,k} - \text{Y}_{\text{G1},k}, \text{X}_{j,k} - \text{X}_k)}{\text{Var}(\text{X}_{j,k} - \text{X}_k)} \\
&= \frac{\text{Cov}(B(\theta_{j,k} - \theta_k) + \eta(\rho_{i,j,k} - \rho_k), (\theta_{j,k} - \theta_k) + (\rho_{j,k} - \rho_k)}{\text{Var}((\theta_{j,k} - \theta_k) + (\rho_{j,k} - \rho_k))} \\
&= \frac{\text{Cov}(B(\theta_{j,k} - \theta_k) + \eta(\rho_{i,j,k} - \rho_k), (\theta_{j,k} - \theta_k) + (\rho_{j,k} - \rho_k)}{\text{Var}(\theta_{j,k} - \theta_k) + \left(\frac{1}{N_{j,k}} - \frac{1}{\sum_{j=1}^{J_k} N_{j,k}}\right)\text{Var}(\rho_{i,j,k})} \\
&= \frac{B\text{Var}(\theta_{j,k} - \theta_k) + \eta\text{Cov}(\rho_{i,j,k} - \rho_k, \rho_{j,k} - \rho_k)}{\text{Var}(\theta_{j,k} - \theta_k) + \text{Var}(\rho_{i,j,k})\left(\frac{(\sum_{j=1}^{J_k} N_{j,k}) - N_{j,k}}{N_{j,k}(\sum_{j=1}^{J_k} N_{j,k})}\right)} \\
&= \frac{B\text{Var}(\theta_{j,k} - \theta_k) + \eta\text{Var}(\rho_{i,j,k})\left(\frac{(\sum_{j=1}^{J_k} N_{j,k}) - N_{j,k}}{N_{j,k}(\sum_{j=1}^{J_k} N_{j,k})}\right)}{\text{Var}(\theta_{j,k} - \theta_k) + \text{Var}(\rho_{i,j,k})\left(\frac{(\sum_{j=1}^{J_k} N_{j,k}) - N_{j,k}}{N_{j,k}(\sum_{j=1}^{J_k} N_{j,k})}\right)},
\end{aligned}$$

which shows that we obtain plim $\hat{\beta}_{K\to\infty} \neq 0$ if $\eta \neq 0$ even if $B = 0$ because class size $N_{j,k}$ is finite. However, if we use the leave-out-mean, we obtain:

$$\begin{aligned}
\text{plim } \hat{b}_{K\to\infty} &= \frac{\text{Cov}(\text{Y}_{\text{G1},i,j,k} - \text{Y}_{\text{G1},k}, \text{X}_{(-i),j,k} - \text{X}_{(-i),k})}{\text{Var}(\text{X}_{(-i),j,k} - \text{X}_{(-i),k})} \\
&= B \cdot \text{Attenuation Factor},
\end{aligned}$$

now we obtain plim $\hat{b}_{K\to\infty} = 0$ when $B = 0$. The attenuation factor is:

$$\text{Attenuation Factor} = \left[\frac{\left(\frac{\sum_{j=1}^{J_k} N_{j,k}}{(\sum_{j=1}^{J_k} N_{j,k}) - 1}\right)\text{Var}(\theta_{j,k} - \theta_k)}{\left(\frac{\sum_{j=1}^{J_k} N_{j,k}}{(\sum_{j=1}^{J_k} N_{j,k}) - 1}\right)^2 \text{Var}(\theta_{j,k} - \theta_k) + \frac{\text{Var}(\rho_{i',j,k})}{-1 + N_{j,k}} + \frac{\text{Var}(\rho_{i',j,k})}{1 - (\sum_{j=1}^{J_k} N_{j,k}})\right]$$

Parenthetically, for space considerations, we do not show the full derivation of the leave-out-mean as we did for the mean.

We follow the intuition in Chetty et al. (2011) and use the within-class variance of $X_{i,j,k}$ as an estimate of $\text{Var}(\rho_{i',j,k})$. First, we rewrite the attenuation factor as a function of $\text{Var}(\theta_{j,k} - \theta_k)$. Second, we fill in the components:

$$\left(\frac{37.69}{38.74}\right)^2 \left[1 - \left(\frac{0.87}{-1 + 19.37} + \frac{0.87}{1 - 38.73}\right)\right] = 0.92.$$

Third, we calculate the attenuation bias using the attenuation factor formuala:

$$1 - \left[\frac{\left(\frac{38.73}{37.69}\right)0.92}{\left(\frac{38.73}{37.69}\right)^2 0.92 + \frac{0.87}{-1 + 19.37} + \frac{0.87}{1 - 38.73}}\right] = 0.05.$$

In sum, the leave-out-mean removes the bias due to a child's unobserved preference but does attenuate the point estimate for teacher relationship skills by about five percent.

A.5. Further Notes on the Monte Carlo Simulations

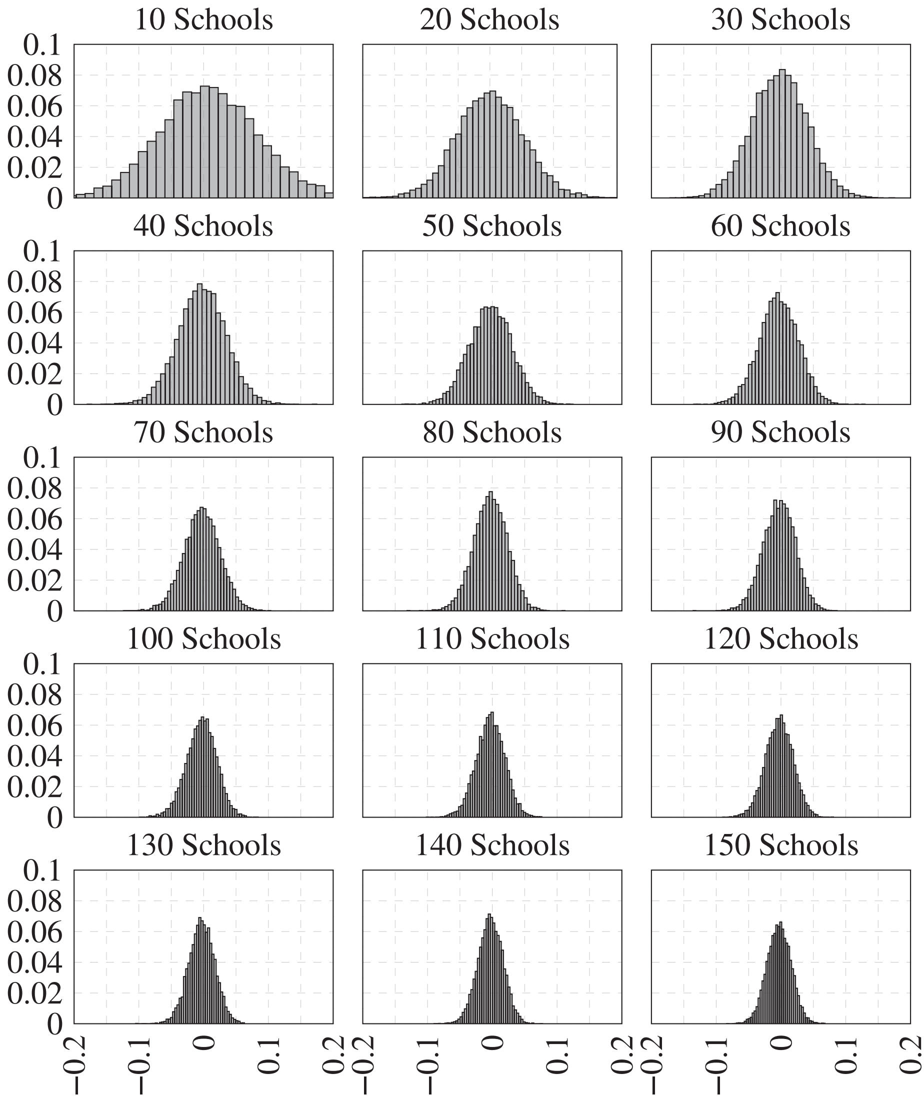

Figure A.5.3. Results Monte Carlo Simulation: Finite Sample Bias

Notes. We use 10,000 repetitions for each of the Monte Carlo simulations.

A.6. Supplementary Tables and Figures

This appendix presents some supplementary findings. In Table A.6.24, we investigate if child characteristics predict classroom peer characteristics conditional on school fixed effects and school-level peer characteristics to account for biases in these types of tests (Guryan et al. 2009). In Table A.6.24, we regress the peer characteristics (in the rows) on that same characteristic for the student while controlling for the school peer characteristics. Each row is a separate regression.

In Figure A.6.3, we examine randomization using Chi-square tests of homogeneity. If administrators randomly assign children to classes (within schools), then characteristics should balance between classes. If this process was random, the distribution of *p*-values should be uniform. To conduct the same test for family income, math (G0), and literacy (G0), we construct two equal groups for these variables (*i.e.*, high and low).

Table A.6.24. Predictability of Predetermined Variables and Classroom Characteristics

	(1) Estimate	(2) SE	(3) Obs.
The Child is Female	0.037	0.022	5,810
Birth Month	−0.014	0.018	5,802
Family Reading Disability	−0.004	0.021	4,536
Education Mother	0.001	0.024	5,474
Education Father	0.038	0.027	5,428
Non-Western Immigrant	−0.018	0.030	5,704
Family Income	−0.029	0.020	5,611
Number of Siblings	0.009	0.021	5,782
Math (G0)	0.016	0.022	5,656
Literacy (G0)	0.013	0.018	5,708

Notes. This table reports ten point estimates of a series of OLS regressions where we regress peer (class) characteristics on student characteristics. We control for the peer (school) characteristics (see Guryan et al. 2009). We also add a school-specific intercept in each model. We cluster the standard errors at the school level and present them in parentheses.
$^{+}$ $p < 0.10$, * $p < 0.05$, and ** $p < 0.01$ (two-tailed).

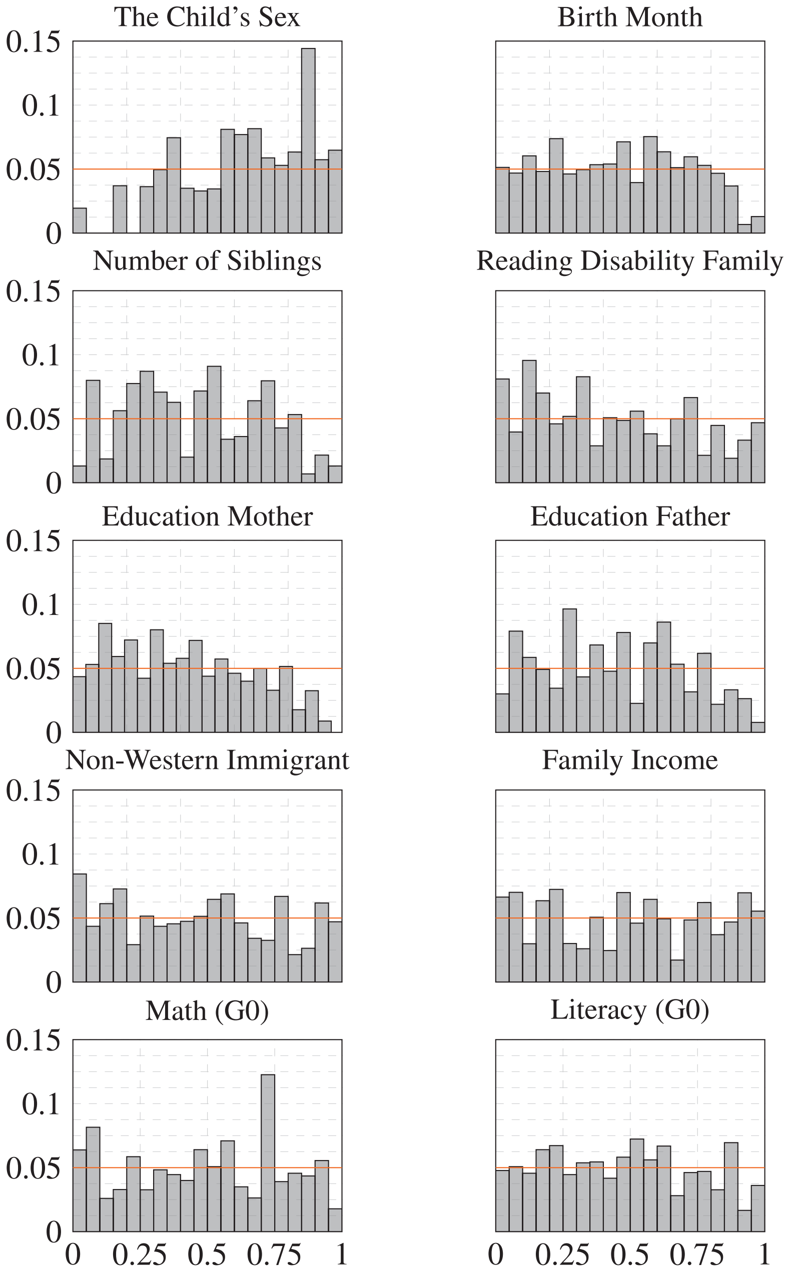

Figure A.6.4. Chi-Square Tests of Homogeneity for Predetermined and Premeasured Variables

Table A.6.25 reports regression results of our preferred model specification with and without treatment indicator. Table A.6.26 reports the regression results, similar to Table 7, but with classes that experienced teacher changes during first grade excluded. Substantively, the main conclusions in the text remain the same. Lastly, in Table A.6.27 and Table A.6.28, we present similar results as Table 7, but with the outcomes measured in second and third grade, respectively. While we exploit the variation in first grade to estimate the effect of teacher relationship skills, teachers may have changed as children progress through school. Consequently, the analysis for second and third grade is more noisy.

Finally, we visualize the results of our sensitivity analysis (*cf.* Altonji et al. 2005, Harada 2013, Imbens 2003). First, Figure A.6.3 illustrates the results of our sensitivity analysis for math. Second, Figure A.6.4 illustrates the results of our sensitivity analysis for literacy. Third, Figure A.6.4 illustrates the results of our sensitivity analysis for self-concept. Fourth, Figure A.6.5 illustrates the results of our sensitivity analysis for reading interest. The observation that math and literacy measured at the *start* of first grade (G0) lie on the left-hand side of the curve implies that any unobserved confounder must have a larger effect than those skills for the effect of teacher relationship skills to drop by half.

Table A.6.25. The Effect of Teacher Relationship Skills on Academic and Social-Emotional Skills With and Without Treatment Indicator

	(1)	(2)
Panel A: Academic Skills at the End of First Grade (G1)		
Math	0.04597**	0.04603**
	(0.016)	(0.016)
Adjusted R^2	0.373	0.373
Observations	5,610	5,610
Literacy	0.02720+	0.02770*
	(0.014)	(0.013)
Adjusted R^2	0.440	0.440
Observations	5,637	5,637
Panel B: Social-Emotional Skills at the End of First Grade (G1)		
Reading Interest	0.04855**	0.04899**
	(0.017)	(0.017)
Adjusted R^2	0.049	0.049
Observations	5,630	5,630
Self-Concept	0.03353*	0.03400*
	(0.016)	(0.016)
Adjusted R^2	0.072	0.072
Observations	5,636	5,636
Treatment Indicator	✓	
Control Variables	✓	✓

Notes. This table reports the point estimates from OLS regressions. We cluster the standard errors (in parentheses) at the school level. We include the same control variables as in our main analyses. First, child and family background includes sex, birth month, birth year, the number of siblings, dummies for mother's education, family reading disability, non-Western immigrant, and family income (quartic family income polynomial). Second, initial skill level (G0) includes the scores for math and reading measured at the *start* of first grade (G0). We also include a cubic polynomial. Third, peer composition consists of all child and family background variables and initial skill level variables specified as leave-out-means. Lastly, teacher and classroom variables include the teacher's sex, education, experience, and class size. All models include variables that predict missingness (see Appendix A.2 for details).
$^+$ $p < 0.10$, * $p < 0.05$, and ** $p < 0.01$ (two-tailed).

Table A.6.26. The Relationship Between Teacher Relationship Skills and Student Learning (Excluding Classes that Experienced a Teacher Change During First Grade)

	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Academic Skills at the End of First Grade (G1)						
Math	0.063**	0.051**	0.051**	0.054**	0.050**	0.046*
	(0.018)	(0.019)	(0.019)	(0.019)	(0.019)	(0.018)
Adjusted R^2	0.097	0.165	0.375	0.375	0.374	0.374
Observations	5,115	5,115	5,115	5,115	5,115	5,110
Literacy	0.048**	0.037*	0.034*	0.034*	0.034*	0.034*
	(0.017)	(0.017)	(0.016)	(0.017)	(0.015)	(0.015)
Adjusted R^2	0.105	0.190	0.436	0.436	0.436	0.435
Observations	5,141	5,141	5,141	5,141	5,141	5,137
School-Specific Intercept	✓	✓	✓	✓	✓	✓
Child and Family Background		✓	✓	✓	✓	✓
Initial Skills (G0)			✓	✓	✓	✓
Peer Composition				✓	✓	✓
Teacher and Classroom					✓	✓

Continued on next page

Table A.6.26. *Continued from previous page*

	(1)	(2)	(3)	(4)	(5)	(6)
Panel B: Social-Emotional Skills the End of First Grade (G1)						
Reading Interest	0.045**	0.044*	0.039*	0.029+	0.028+	0.033*
	(0.017)	(0.018)	(0.018)	(0.017)	(0.016)	(0.016)
Adjusted R^2	0.038	0.055	0.060	0.060	0.063	0.063
Observations	5,134	5,134	5,134	5,134	5,134	5,133
Self-Concept	0.036*	0.037*	0.036*	0.032+	0.035*	0.050**
	(0.016)	(0.016)	(0.016)	(0.018)	(0.017)	(0.017)
Adjusted R^2	0.038	0.047	0.077	0.077	0.077	0.076
Observations	5,140	5,140	5,140	5,140	5,140	5,137
School-Specific Intercept	✓	✓	✓	✓	✓	✓
Child and Family Background		✓	✓	✓	✓	✓
Initial Skills (G0)			✓	✓	✓	✓
Peer Composition				✓	✓	✓
Teacher and Classroom					✓	✓

Notes. This table reports the point estimates from an OLS regression. Family background includes sex, birth month, birth year, the number of siblings, dummies for mother's education, family reading disability, non-Western immigrant, and family income (quartic family income polynomial). Initial skill level (G0) includes the scores for math and literacy measured at the *start* of first grade (G0). Peer composition consists of all family background variables and initial skill level variables specified as leave-out-means. Teacher and classroom variables include the teacher's sex, age, education, experience, and class size. We cluster the standard errors at the school level and present them in parentheses. All models include an indicator for treatment status and variables that predict missingness (see Appendix A.2 for details).

$^+$ $p < 0.10$, * $p < 0.05$, and ** $p < 0.01$ (two-tailed).

Table A.6.27. The Effect of Teacher Relationship Skills on Academic and Social-Emotional Skills and the End of Second Grade

	(1)	(2)	(3)	(4)	(5)
Panel A: Academic Skills at the End of Second Grade (G2)					
Math	0.027	0.022	0.021	0.029^{+}	0.032^{+}
	(0.017)	(0.018)	(0.017)	(0.017)	(0.017)
Adjusted R^2	0.084	0.149	0.332	0.333	0.334
Observations	5,319	5,319	5,319	5,319	5,319
Literacy	0.038*	0.033*	0.024*	0.026^{+}	0.026
	(0.015)	(0.014)	(0.012)	(0.013)	(0.013)
Adjusted R^2	0.104	0.184	0.409	0.409	0.410
Observations	5,345	5,345	5,345	5,345	5,345
School-Specific Intercept	✓	✓	✓	✓	✓
Child and Family Background		✓	✓	✓	✓
Initial Skills (G0)			✓	✓	✓
Peer Composition				✓	✓
Teacher and Classroom					✓

Continued on next page

Table A.6.27. *Continued from previous page*

	(1)	(2)	(3)	(4)	(5)
Panel B: Social-Emotional Skills the End of Second Grade (G2)					
Reading Interest	0.021	0.019	0.014	0.012	0.017
	(0.018)	(0.019)	(0.018)	(0.018)	(0.017)
Adjusted R^2	0.025	0.047	0.063	0.063	0.065
Observations	5,334	5,334	5,334	5,334	5,334
Self-Concept	0.030^{+}	0.031^{+}	0.027	0.028^{+}	0.025
	(0.016)	(0.017)	(0.016)	(0.017)	(0.017)
Adjusted R^2	0.020	0.036	0.076	0.076	0.077
Observations	5,348	5,348	5,348	5,348	5,348
School-Specific Intercept	✓	✓	✓	✓	✓
Child and Family Background		✓	✓	✓	✓
Initial Skills (G0)			✓	✓	✓
Peer Composition				✓	✓
Teacher and Classroom					✓

Notes. This table reports the point estimates from an OLS regression. Family background includes sex, birth month, birth year, the number of siblings, dummies for mother's education, family reading disability, non-Western immigrant, and family income (quartic family income polynomial). Initial skills (G0) includes the scores for math and literacy measured at the *start* of first grade (G0). Peer composition consists of all family background variables and initial skill level variables specified as leave-out-means. Teacher and classroom variables include the teacher's sex, age, education, experience, and class size. We cluster the standard errors at the school level and present them in parentheses. All models include an indicator for treatment status and variables that predict missingness (see Appendix A.2 for details).
$^{+}$ $p < 0.10$, * $p < 0.05$, and ** $p < 0.01$ (two-tailed).

Table A.6.28. The Effect of Teacher Relationship Skills on Academic and Social-Emotional Skills and the End of Third Grade

	(1)	(2)	(3)	(4)	(5)
Panel A: Academic Skills at the End of Third Grade (G3)					
Math	0.019	0.018	0.018	0.014	0.017
	(0.019)	(0.019)	(0.018)	(0.019)	(0.018)
Adjusted R^2	0.090	0.139	0.272	0.274	0.274
Observations	5,050	5,050	5,050	5,050	5,050
Literacy	0.027	0.025	0.023^{+}	0.028^{+}	0.033^{*}
	(0.016)	(0.015)	(0.014)	(0.014)	(0.014)
Adjusted R^2	0.104	0.181	0.357	0.357	0.356
Observations	5,093	5,093	5,093	5,093	5,093
School-Specific Intercept	✓	✓	✓	✓	✓
Child and Family Background		✓	✓	✓	✓
Initial Skills (G0)			✓	✓	✓
Peer Composition				✓	✓
Teacher and Classroom					✓

Continued on next page

Table A.6.28 *Continued from previous page*

	(1)	(2)	(3)	(4)	(5)
Panel B: Social-Emotional Skills the End of Third Grade (G3)					
Reading Interest	−0.002	−0.004	−0.008	−0.015	−0.011
	(0.020)	(0.020)	(0.020)	(0.019)	(0.018)
Adjusted R^2	0.050	0.085	0.102	0.101	0.101
Observations	5,079	5,079	5,079	5,079	5,079
Self-Concept	−0.018	−0.017	−0.018	−0.022	−0.021
	(0.017)	(0.016)	(0.016)	(0.018)	(0.018)
Adjusted R^2	0.019	0.049	0.105	0.103	0.103
Observations	5,096	5,096	5,096	5,096	5,096
School-Specific Intercept	✓	✓	✓	✓	✓
Child and Family Background		✓	✓	✓	✓
Initial Skills (G0)			✓	✓	✓
Peer Composition				✓	✓
Teacher and Classroom					✓

Notes. This table reports the point estimates from an OLS regression using Equation (4) with outcomes measured at the *end* of third grade (G3). We cluster the standard errors at the school level and present them in parentheses. Family background includes sex, birth month, birth year, the number of siblings, dummies for mother's education, family reading disability, non-Western immigrant, and family income (quartic family income polynomial). Initial skill level (G0) includes the scores for math and reading measured at the *start* of first grade (G0). We also include a cubic polynomial. Peer composition consists of all family background variables and initial skill level variables specified as leave-out-means. Teacher and classroom variables include the teacher's sex, education, experience, and class size. All models include an indicator for treatment status and variables that predict missingness (see Appendix A.2 for details).
$^{+}$ $p < 0.10$, * $p < 0.05$, and ** $p < 0.01$ (two-tailed).

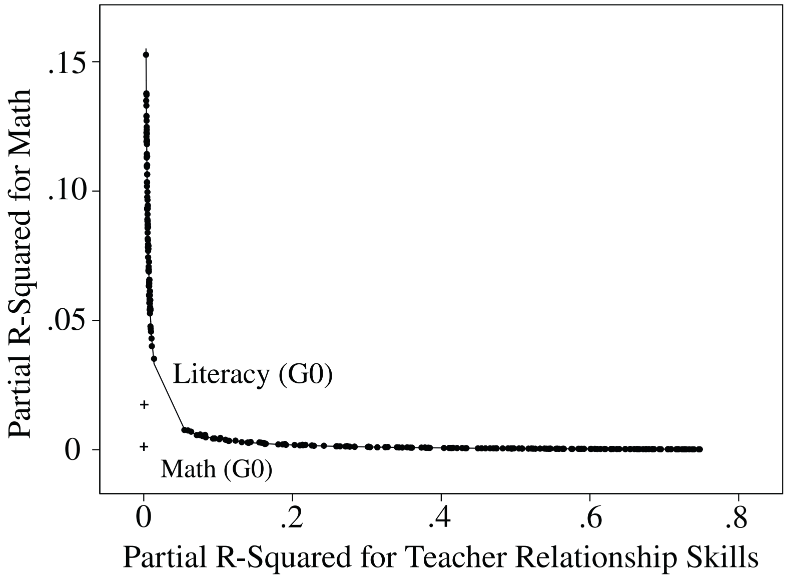

Figure A.6.5. Sensitivity Analysis: Math

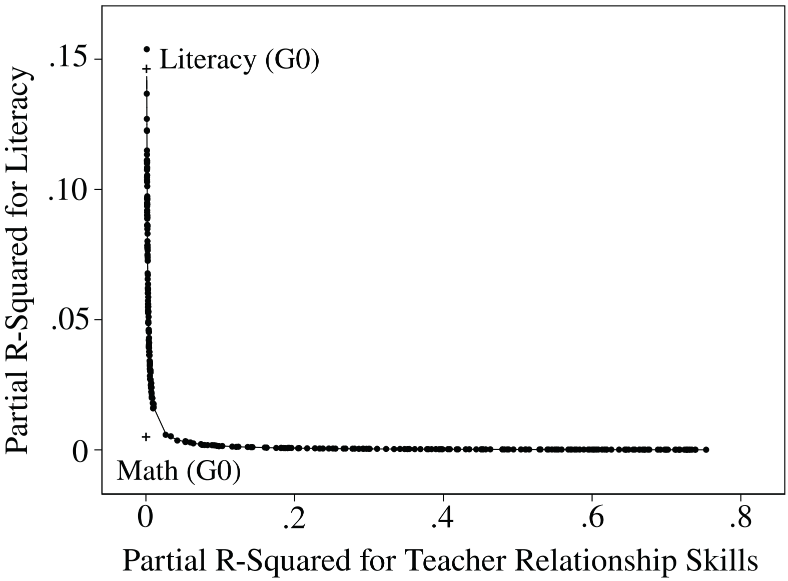

Figure A.6.6. Sensitivity Analysis: Literacy

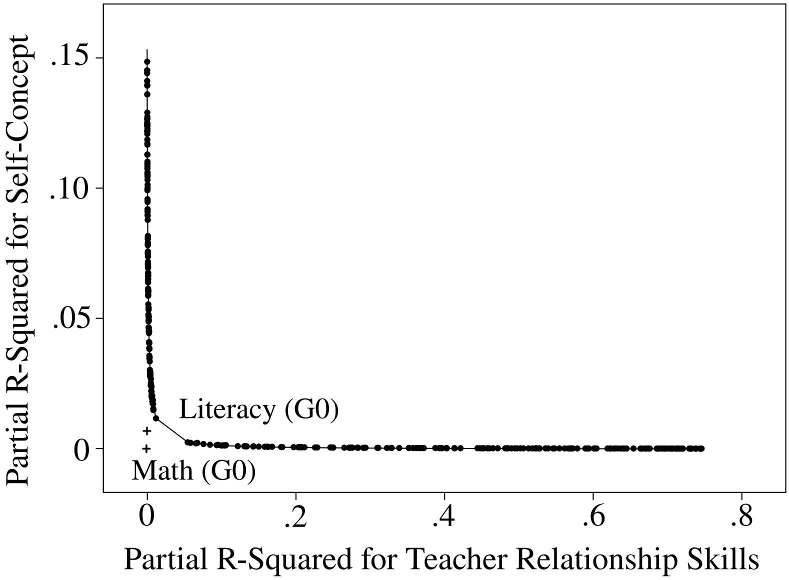

Figure A.6.7. Sensitivity Analysis: Self-Concept

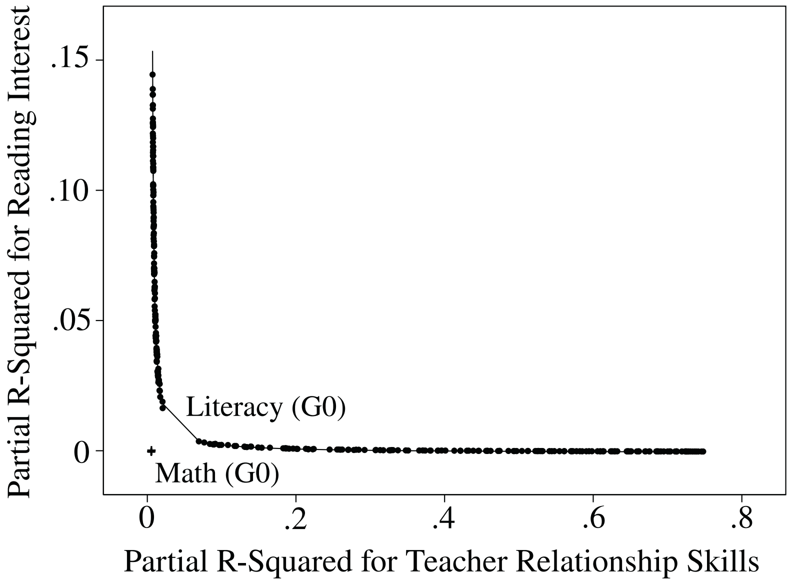

Figure A.6.8. Sensitivity Analysis: Reading Interest

A.7. Further Notes on the Control Function Approach

The intuition behind this alternative approach is to use a two-stage residual inclusion to address the endogeneity due to a child's unobserved preference. In the first stage, we assign values to the unobserved (or latent) variables $\theta_{j,k}$ and $\rho_{i,j,k}$. Then, in the second stage, we regress $Y_{G1,i,j,k}$ on the best linear unbiased predictions (or BLUPs) $\hat{\theta}_{j,k}$ and $\hat{\rho}_{i,j,k}$ from the first stage.

If we stack the 11 items, we can write the lth equation as:

$$X_{l,i,j,k} = \mu + \phi_k + \theta_{j,k} + \rho_{i,j,k} + \iota_{l,i,j,k}, \qquad \text{(A.7.3)}$$

where $X_{l,i,j,k}$ is the lth item for child i in class j in school k, μ is the sample average, ϕ_k is a school component, $\theta_{j,k}$ is the teacher/class component, $\rho_{i,j,k}$ is a student-specific component, and $\iota_{l,i,j,k}$ is variation unique to item l. Each of these components captures variation at a different level of clustering. The school component captures variation between schools, the teacher/class component captures variation within schools, but between classes, the student component captures variation within schools and within classes, but between students. The error term captures variation within students, but between indicators. In addition to normality, we can estimate Equation (A.7.5) using parametric maximum likelihood under the following assumptions:

Assumption A.7.1 (No Peer Effects – Students). *Unobserved preferences of child i do not affect the unobserved preferences for child i' conditional on the teacher relationship skills and school component, $Cov(\rho_{i,j,k}, \rho_{i',j,k} | \theta_{j,k}, \phi_k) = 0$, $i \neq i'$.*

Assumption A.7.2 (No Peer Effects – Teachers). *Teacher relationship skills in class j do not affect teacher relationship skills in class j' conditional on the school component, $Cov(\theta_{j,k}, \theta_{j',k} | \phi_k) = 0$ for $j \neq j'$.*

Note that these assumptions are different from the assumptions of the leave-out-mean. Table A.7.29, Table A.7.30, and Table A.7.31 present results using this approach for first, second, and third grade, respectively.

When we use predicted variables, the standard errors will be incorrect because they do not account for the fact that variables have been predicted. We, therefore, apply a clustered residual wild bootstrap procedure. We draw 1,000 bootstrap samples of the original data and apply the estimation procedure to each pseudo-sample. We cluster and resample at the school level so that each bootstrap sample includes both classes.

Table A.7.29. The Effect of Teacher Relationship Skills on Academic and Social-Emotional Skills and the End of First grade (G1)

	(1)	(2)	(3)	(4)	(5)
Panel A: Academic Skills at the End of First Grade (G1)					
Math	0.075**	0.064**	0.068**	0.068**	0.070**
	(0.020)	(0.020)	(0.020)	(0.019)	(0.018)
Adjusted R^2	0.096	0.165	0.377	0.377	0.377
Observations	5,610	5,610	5,610	5,610	5,610
Literacy	0.054**	0.042*	0.037*	0.036*	0.040*
	(0.020)	(0.020)	(0.018)	(0.017)	(0.016)
Adjusted R^2	0.109	0.194	0.442	0.443	0.444
Observations	5,637	5,637	5,637	5,637	5,637
School-Specific Intercept	✓	✓	✓	✓	✓
Child and family Background		✓	✓	✓	✓
Initial Skills (G0)			✓	✓	✓
Peer Composition				✓	✓
Teacher and Classroom					✓

Continued on next page

Table A.7.29 *Continued from previous page*

	(1)	(2)	(3)	(4)	(5)
Panel B: Social-Emotional Skills the End of First Grade (G1)					
Reading Interest	0.088**	0.087**	0.084**	0.085**	0.089**
	(0.021)	(0.021)	(0.021)	(0.019)	(0.020)
Adjusted R^2	0.142	0.153	0.156	0.157	0.158
Observations	5,630	5,630	5,630	5,630	5,630
Self-Concept	0.054**	0.053**	0.050**	0.050**	0.056**
	(0.017)	(0.017)	(0.016)	(0.017)	(0.017)
Adjusted R^2	0.102	0.108	0.133	0.133	0.133
Observations	5,636	5,636	5,636	5,636	5,636
School-Specific Intercept	✓	✓	✓	✓	✓
Child and family Background		✓	✓	✓	✓
Initial Skills (G0)			✓	✓	✓
Peer Composition				✓	✓
Teacher and Classroom					✓

Notes. This table reports the point estimates from an OLS regression using the control function approach instead of the leave-out-mean. School-clustered bootstrapped standard errors in parentheses (1,000 bootstrap samples). Family background includes sex, birth month, birth year, the number of siblings, dummies for mother's education, family reading disability, non-Western immigrant, and family income (quartic family income polynomial). Initial skills (G0) includes the scores for math and reading measured at the *start* of first grade (G0). We also include a cubic polynomial. Peer composition consists of all family background variables and initial skill level variables specified as leave-out-means. Teacher and classroom variables include the teacher's sex, education, experience, and class size. All models include an indicator for treatment status and variables that predict missingness (see Appendix A.2 for details).
$^{+}$ $p < 0.10$, * $p < 0.05$, and ** $p < 0.01$ (two-tailed).

Table A.7.30. The Effect of Teacher Relationship Skills on Academic and Social-Emotional Skills and the End of Second Grade (G2)

	(1)	(2)	(3)	(4)	(5)
Panel A: Academic Skills at the End of Second Grade (G2)					
Math	0.035	0.031	0.032	0.042*	0.048*
	(0.020)	(0.021)	(0.020)	(0.019)	(0.019)
Adjusted R^2	0.090	0.153	0.335	0.335	0.337
Observations	5,319	5,319	5,319	5,319	5,319
Literacy	0.045*	0.040*	0.034*	0.036*	0.034*
	(0.019)	(0.018)	(0.015)	(0.016)	(0.016)
Adjusted R^2	0.108	0.186	0.409	0.408	0.409
Observations	5,345	5,345	5,345	5,345	5,345
School-Specific Intercept	✓	✓	✓	✓	✓
Child and family Background		✓	✓	✓	✓
Initial Skills (G0)			✓	✓	✓
Peer Composition				✓	✓
Teacher and Classroom					✓

Continued on next page

Table A.7.30. *Continued from previous page*

	(1)	(2)	(3)	(4)	(5)
Panel B: Social-Emotional Skills the End of Second Grade (G2)					
Reading Interest	0.034	0.034	0.031	0.030	0.039^{+}
	(0.021)	(0.022)	(0.021)	(0.021)	(0.021)
Adjusted R^2	0.055	0.074	0.089	0.089	0.091
Observations	5,334	5,334	5,334	5,334	5,334
Self-Concept	0.042^{*}	0.045^{*}	0.041^{*}	0.044^{*}	0.038
	(0.020)	(0.021)	(0.020)	(0.020)	(0.021)
Adjusted R^2	0.033	0.049	0.088	0.088	0.090
Observations	5,348	5,348	5,348	5,348	5,348
School-Specific Intercept	✓	✓	✓	✓	✓
Child and family Background		✓	✓	✓	✓
Initial Skills (G0)			✓	✓	✓
Peer Composition				✓	✓
Teacher and Classroom					✓

Notes. This table reports the point estimates from an OLS regression using the control function approach instead of the leave-out-mean. School-clustered bootstrapped standard errors in parentheses (1,000 bootstrap samples). Family background includes sex, birth month, birth year, the number of siblings, dummies for mother's education, family reading disability, non-Western immigrant, and family income (quartic family income polynomial). Initial skills (G0) includes the scores for math and reading measured at the *start* of first grade (G0). We also include a cubic polynomial. Peer composition consists of all family background variables and initial skill level variables specified as leave-out-means. Teacher and classroom variables include the teacher's sex, education, experience, and class size. All models include an indicator for treatment status and variables that predict missingness (see Appendix A.2 for details).
$^{+}$ $p < 0.10$, * $p < 0.05$, and ** $p < 0.01$ (two-tailed).

Table A.7.31. The Effect of Teacher Relationship Skills on Academic and Social-Emotional Skills and the End of Third Grade (G3)

	(1)	(2)	(3)	(4)	(5)
Panel A: Academic Skills at the End of Third Grade (G3)					
Math	0.032	0.030	0.031	0.027	0.031
	(0.023)	(0.022)	(0.021)	(0.021)	(0.021)
Adjusted R^2	0.094	0.143	0.275	0.277	0.276
Observations	5,050	5,050	5,050	5,050	5,050
Literacy	0.034^+	0.032^+	0.032^+	0.037^*	0.043^*
	(0.020)	(0.019)	(0.017)	(0.017)	(0.017)
Adjusted R^2	0.106	0.182	0.357	0.357	0.356
Observations	5,093	5,093	5,093	5,093	5,093
School-Specific Intercept	✓	✓	✓	✓	✓
Child and family Background		✓	✓	✓	✓
Initial Skills (G0)			✓	✓	✓
Peer Composition				✓	✓
Teacher and Classroom					✓

Continued on next page

Table A.7.31. *Continued from previous page*

	(1)	(2)	(3)	(4)	(5)
Panel B: Social-Emotional Skills the End of Third Grade (G3)					
Reading Interest	0.007	0.004	0.001	−0.006	0.000
	(0.024)	(0.024)	(0.024)	(0.023)	(0.023)
Adjusted R^2	0.065	0.098	0.114	0.113	0.113
Observations	5,079	5,079	5,079	5,079	5,079
Self-Concept	−0.028	−0.026	−0.027	−0.032	−0.033
	(0.021)	(0.020)	(0.020)	(0.021)	(0.021)
Adjusted R^2	0.026	0.055	0.110	0.108	0.109
Observations	5,096	5,096	5,096	5,096	5,096
School-Specific Intercept	✓	✓	✓	✓	✓
Child and family Background		✓	✓	✓	✓
Initial Skills (G0)			✓	✓	✓
Peer Composition				✓	✓
Teacher and Classroom					✓

Notes. This table reports the point estimates from an OLS regression using the control function approach instead of the leave-out-mean. School-clustered bootstrapped standard errors in parentheses (1,000 bootstrap samples). Family background includes sex, birth month, birth year, the number of siblings, dummies for mother's education, family reading disability, non-Western immigrant, and family income (quartic family income polynomial). Initial skills (G0) includes the scores for math and reading measured at the *start* of first grade (G0). We also include a cubic polynomial. Peer composition consists of all family background variables and initial skill level variables specified as leave-out-means. Teacher and classroom variables include the teacher's sex, education, experience, and class size. All models include an indicator for treatment status and variables that predict missingness (see Web Appendix B for details).
$^+ p < 0.10$, $^* p < 0.05$, and $^{**} p < 0.01$ (two-tailed).