Individual Differences in Skill Development: Toward a Causal Explanation

by

Rafael Quintana

B.A., Universidad Nacional de Colombia, 2009

M.A., École Normale Supérieure, 2013

Submitted to the Graduate Faculty of the

School of Education in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2020

UNIVERSITY OF PITTSBURGH

SCHOOL OF EDUCATION

This dissertation was presented

by

Rafael Quintana

It was defended on

April 29, 2020

and approved by

Steven Finkel, Daniel Wallace Professor, Political Science

Lindsay Page, Associate Professor, Psychology in Education

Elizabeth Votruba-Drzal, Professor, Psychology

Dissertation Director: Richard Correnti, Associate Professor, Learning Sciences and Policy

Individual Differences in Skill Development: Toward a Causal Explanation

Rafael Quintana, PhD

University of Pittsburgh, 2020

Describing and explaining individual differences in skill development is a fundamental component of educational research. However, given the highly interdisciplinary nature of the field and the various theoretical and methodological approaches involved, studies on skill development often fail to provide a coherent and cumulative body of research. In this three-study dissertation, I discuss foundational conceptual and methodological issues in skill development, and show how different approaches can be integrated in a principled and cumulative fashion. The first study presents a general framework, referred to under the label of "academic mobility", for describing the development of educational inequalities using student learning outcomes. In this study, I discuss ways of operationalizing the concept of educational inequality, and measure academic mobility at a national level using five mobility metrics. While the first study is descriptive in nature, the second and third study intend to shed some light into why individual differences might arise. The second study presents an approach for establishing the explanatory relevance of different predictors based on distal and proximal considerations. For this purpose, I implement several causal search algorithms and find that, consistent with my research hypotheses, previous achievement and executive functions are proximal mechanisms of both reading and math achievement. Finally, in the third paper I describe the relationship between executive functions and academic achievement by implementing several within-person methodological strategies.

Table of Contents

1.0 Introduction 1

1.1 Overview 2

1.2 On the use of causal language 5

1.3 The concept of academic mobility 8

1.4 The structure of academic achievement 10

1.5 Identifying proximal mechanisms 11

1.6 Thinking within-persons 13

1.7 General contribution 14

2.0 Study 1: The concept of academic mobility: Normative and methodological considerations 15

2.1 Introduction 15

2.2 The normative dimensions of academic mobility 20

2.2.1 Defining equality of opportunity 20

2.2.2 Measuring educational inequality: Normative considerations 23

2.2.3 Between versus within-person variation 23

2.2.4 Relative versus absolute measures of change 26

2.2.5 Defining academic mobility 28

2.3 Measuring educational inequality: Methodological considerations 30

2.3.1 Using growth models to measure relative change 30

2.3.2 Using achievement gaps to measure relative change 33

2.3.3 Mobility metrics and their advantages for studying relative change 36

2.3.4 Dimensions of academic mobility 37

2.4 Data 37

2.5 Five mobility metrics 39

2.5.1 Linear rank-rank measures 39

2.5.2 Measures of the amplitude of academic mobility 40

2.5.3 Transition probabilities 42

2.5.4 Measuring stability and change 44

2.5.5 Measuring group differences over time 46

2.6 Estimates of academic mobility across racial groups in the US 47

2.6.1 Overall degrees of academic mobility 47

2.6.2 Amplitude of academic mobility 50

2.6.3 Transition matrices and rank reversal probabilities 51

2.6.4 Stability and change 55

2.6.5 Group differences over time 57

2.7 Discussion 58

2.7.1 Complementing achievement gap analyses with academic mobility metrics 59

2.7.2 Limitations and conclusion 63

3.0 Introduction to Studies 2 and 3 64

4.0 Study 2. The structure of academic achievement: Searching for proximal mechanisms using causal discovery algorithms 68

4.1 Introduction 69

4.2 The challenges of causal inference: The case of academic achievement 71

4.3 Causal structures and proximal mechanisms 74

4.3.1 Defining proximal mechanisms.....76

4.4 Searching for the proximal mechanisms of academic achievement 79

4.4.1 Previous achievement 80

4.4.2 Executive functions 81

4.4.3 Motivation.....82

4.5 Using causal discovery algorithms to search for the proximal mechanisms of academic achievement..... 83

4.5.1 Causal learning algorithms 86

4.6 Data..... 90

4.6.1 Child-level variables 92

4.6.2 Family-level variables 94

4.6.3 School-level variables.....95

4.6.4 Neighborhood-level variables.....96

4.6.5 Demographic variables 96

4.7 Empirical analysis 98

4.7.1 Skeleton identification 99

4.7.2 Edge orientation 100

4.8 Results..... 101

4.8.1 Reading achievement 101

4.8.2 Math achievement 107

4.8.3 Reading and math achievement combined 111

4.8.4 Using the estimated proximal mechanisms to predict the value of academic achievement 113

4.9 Discussion 116

5.0 Study 3: Thinking within-persons: Using unit fixed-effects models to describe the causal relationship between executive functions and academic achievement 120

5.1 Introduction 121

5.2 Basic concepts in causal inference 123

5.2.1 Causal inference and the purposes of statistical modelling 123

5.2.2 Defining causal effects 126

5.2.3 Causal identification 127

5.2.4 Using longitudinal data for causal identification 131

5.2.5 Estimating FE models 133

5.3 Describing causal mechanisms at the within-person level 136

5.3.1 The advantages of thinking "within-persons" 136

5.3.2 Within-person asymmetrical causation 137

5.3.3 Within-person interactions 140

5.3.4 Estimating between-person differences in within-person effects 142

5.3.5 Modelling within-person reciprocal causation 143

5.4 Data 148

5.4.1 Measures 148

5.5 Results 150

5.5.1 Within and between-person effects 150

5.5.2 Within-person asymmetric effects 152

5.5.3 Within-person interactions.....153

5.5.4 Within-person random slopes.....153

5.5.5 Within-person reciprocal causation.....154

5.6 Discussion..... 157

6.0 General discussion.....160

6.1 Synthesis across studies..... 165

6.1.1 Theoretical contributions.....165

6.1.2 Methodological contributions.....167

6.2 Supporting value judgements in educational research..... 168

6.2.1 Value judgments in descriptive research.....168

6.2.2 Value judgments in explanatory research.....169

6.3 Limitations and future research..... 171

Appendix A : Supplemental materials for study 1.....175

Bibliography.....180

List of Tables

Table 1. Rank reversal probabilities by racial group 54

Table 2. Estimated parameters and model fit statistics of the stable trait, autoregressive trait, and state (START) model by race 56

Table 3. High and low-confidence directed and undirected edges using the reading dataset 104

Table 4. High and low-confidence directed and undirected edges using the math dataset 108

Table 5. High and low-confidence directed and undirected edges using the combined dataset 112

Table 6. Summary of cross-validation results. 115

Table 7. Estimated within-person effects of executive functions on reading and math achievement using eight different models. 151

Table 8. Fit statistics of the full and constrained General Cross-Lagged Panel Model. 156

Appendix Table 1. Sample size and descriptive statistics of percentile rank scores in reading achievement by race. 175

Appendix Table 2. Rank-rank estimates by racial group. 176

Appendix Table 3. Indicators of the amplitude of intraindividual mobility by race. 176

Appendix Table 4. National quartile transition matrix (%). 176

List of Figures

Figure 1. Six patterns representing the development of inter and intra-individual differences among four individuals belonging to two different groups 35

Figure 2. Scatter plots representing the relationship between students' achievement rank in eighth grade and students' achievement rank in kindergarten (Panel A) and the relationship between students' rank achievement in eighth grade and in kindergarten by race (Panel B). 49

Figure 3. Transition bar chart comparing Whites' and Blacks' positional mobility. 53

Figure 4. Coefficient of determination (R^2) of achievement predicted by race across grades. 57

Figure 5. Estimated DAG with the parents and non-parents of reading achievemen..... 106

Figure 6. Estimated DAG with the parents and non-parents of math achievement..... 110

Figure 7. Estimated DAG with the edges directly related to academic achievement. 113

Figure 8. Three types of bias with their respective solutions. The square represents that the variable has been conditioned on..... 129

Figure 9. Causal graph representing the data-generating process implied by traditional FE models..... 143

Appendix Figure 1. The stable trait, autoregressive trait, and state (START) model. 177

Appendix Figure 2. Mean negative (1) and positive (2) ocassion-to-occasion achievement difference by race. 178

Appendix Figure 3.Transition bar charts representing systems with no mobility (1) and complete mobility (2). 179

1.0 Introduction

It is generally recognized that a fundamental goal of education is to develop the skills and knowledge required for participation in social, political, economic, and cultural life (e.g., National Research Council, 2012; OECD, 2019). In particular, reading and mathematics are widely considered "basic" or "foundational" skills, in the sense that they are required across different disciplines, and are predictive of a range of educational, economic and health outcomes (e.g., Cutler & Lleras-Muney, 2010; Duncan et al., 2007; Levy & Murnane, 2004). Describing and explaining individual differences in cognitive abilities in general, and reading and mathematics in particular (commonly considered the key constituents of "academic achievement") has been, then, a major component in educational research. The wide-ranging and interdisciplinary literature on this topic can be categorized under the umbrella term of "skill development research."

Yet the extent to which skill development research constitutes a monolithic and coherent field of study can be put into question, given the vastly different theoretical and methodological approaches employed, and the various disciplinary traditions involved (e.g., education, psychology, sociology or economics). The difficulty of unifying these approaches and traditions often derives from discrepancies in key underlying assumptions that are frequently overlooked. For example, while sociologists might focus on social-level phenomena to explain differences in academic achievement, psychologists might focus on individual-level phenomena; and without any bridging principles or some general framework connecting these underlying assumptions, we might end up with parallel bodies of research that are fundamentally disconnected from each other. The general purpose of this dissertation is to discuss conceptual and methodological tools that

might be useful for investigating skill development in a principled, systematic and coherent fashion.

1.1 Overview

This dissertation discusses foundational methods and assumptions in skill development research: the first study examines methods and assumptions that are commonly made in descriptive research, while the second and third studies consider methods and assumptions typically made in explanatory research. More specifically, the first study discusses normative and methodological principles for describing educational inequalities using academic achievement outcomes; the second study discusses conceptual and methodological tools for establishing the relevance of explanatory variables in the context of skill development; and the third study discusses modelling strategies for estimating causal effects using achievement outcomes. Even if the three studies address different issues in skill development, they have important features in common.

Studies of change. The three studies reflect on conceptual and methodological issues related to the study of change in skill development: the first study discusses issues related to within and between-person variation as well as absolute and relative change; the second study presents a method that takes advantage of the longitudinal structure of the data to impose constraints on the causal search algorithms (e.g., by forbidding variables measured at some time point to have a causal effect on variables measured at prior time points); and the third study examines different modelling strategies that focus on within-person variation to estimate causal effects.

Repeated measures. More generally, in this dissertation I argue that skill development research often fails to capitalize on the opportunities that panel data offer. From a descriptive

perspective, researchers have often summarized the development of individual or group-level differences in achievement using mean-based measures (e.g., achievement gaps). Even if these measures can be informative, they do not capture a wide-range of distributional information that can reveal policy and socially relevant aspects, e.g., the extent to which students beginning at the corners of the distribution remain in their starting positions, or the overall mobility of the system.

From an explanatory perspective, I show how some of the modelling techniques that are commonly used to estimate the effect of a particular predictor on academic achievement fail to capitalize on the longitudinal structure of the data by not disaggregating within and between-person variability. In the third study, I show that panel data offers important advantages for causal inference, as it allows to control for stable confounders. However, this is achieved only if the statistical model implemented properly distinguishes within and between-person variation.

Early childhood. The three studies focus on skill development in early childhood. In particular, the three studies include an empirical application of a different method using the Early Childhood Longitudinal Study samples (both the 1998 and 2011 samples). The focus on early childhood is based on the importance of this period for understanding the development of individual differences (see e.g., Heckman, 2011).

Progression from descriptive to explanatory models. The three studies in this dissertation follow a common progression in scientific research, from descriptive to explanatory models. As several authors have noted (e.g., Berk, 2004; Hernán, 2018; Pearl, 2019; Shmueli, 2010; Spirtes & Zhang, 2016), it is important to distinguish between the different purposes of statistical modelling – and, in particular, between descriptive, predictive and causal purposes. Even if conflation between these concepts is common, these three purposes correspond to substantially different

claims that have different standards of success, uses and supporting assumptions. Understanding this distinction is essential, then, to properly conduct, interpret and assess quantitative research.

In a nutshell, the goal of *descriptive* models is to characterize or synthesize the observed patterns and relationships in the data, and their standards of success are related to their informativeness and parsimony. The goal of *predictive* models is to maximize the predictive accuracy of a target variable using some measured variables, and their standards of success are related to the extent to which they minimize the prediction error. Finally, the goal of *causal* models is to understand the processes in the world that generated the data (i.e., understand how the world works), and their standards of success are related to the extent to which they adequately explain or "mirror" the mechanisms underlying the system under investigation. In other words, causal models assume that the mechanisms in the world are faithfully represented in the model (which can be, as I will explain, a very strong and often untestable assumption); however, these models allow us to understand how the world works as well as predict the effect of interventions, and this is why they are considered in many cases the most useful.

The first study is entirely descriptive, as it provides different metrics that intend to synthesize or summarize the observed patterns in the data. Importantly, no attempts are made at explaining why these patterns emerge nor make any future predictions. The argument that I put forward –following the definition of descriptive models provided above– is that the metrics presented are more *informative* and *parsimonious* compared to traditional methods. I also argue that descriptive analyses are a fundamental component of scientific research, as they define the phenomena that subsequent research will try to explain. Thus, I show how the metrics presented can be used to investigate important issues that cannot be easily studied using traditional methods, e.g., how does the academic mobility in a particular educational system such as a district or a

country compare with the academic mobility of other educational systems, or the extent to which mobility patterns vary across the achievement distribution.

In the second and third studies, I move to an explanatory framework (even though I show that the methods presented in the second study can also be used for predictive purposes); that is, the aim of these studies is to shed some light into why individual differences in academic achievement might arise. The second paper discusses and implements a methodological strategy for establishing the relevance of explanatory variables, and the third paper discusses and implements methodological strategies for estimating the magnitude of the relationship between a predictor and academic achievement. In other words, in the second study I provide methods for justifying the selection of explanatory variables, and in the third study I estimate the causal effect of such variables on achievement.

1.2 On the use of causal language

Researchers frequently avoid the use of causal language when relying on observational data, given that observational studies are typically incapable of supporting causal inferences (e.g., Morgan & Winship, 2014). Instead, the general advice is to use other terms such as "predicts", "influences", "impacts", "affects", "benefits", "is associated with", etc. In this dissertation, I join the researchers (e.g., Bollen & Pearl, 2012; Grosz, Rohrer & Thoemmes, 2020; Hernán, 2019) who argue that, even if this practice is based on logical concerns, the unqualified avoidance of causal terminology in observational studies might do more harm than good. In particular, precluding the use of the c-word prevents a clear understanding and discussion of the broader *goals* of quantitative research.

As I explain in the third study, the driving questions of most programs of empirical research in education and psychology are causal in nature; for example, following a wide-ranging body of empirical research, the goal of the third study is to estimate the causal effect of executive functions on academic achievement. A central motivation behind this line of inquiry is to help us understand whether improving students' executive functions can affect their academic achievement. Clearly, the goal of this body of research is causal in nature (as it examines the extent to which changes in one variable determine changes in other variable), and avoiding the c-world could only obscure the true aim of these investigations.

Researchers often justify avoiding causal terminology to describe the aims of their investigation given the well-known threats to internal validity faced by observational studies. However, as Hernán (2019, p.618) explains, "this argument simply conflates the aims and the methods of scientific research (…) without causally explicit language, the means and ends of much of observational research get hopelessly conflated." Specifically, avoiding causal terminology impedes recognizing that the goal of a large portion of the empirical research in social and behavioral sciences is causal in nature[1].

As explained above, clarifying the aims of scientific research is essential for properly designing, conducting and assessing an investigation. The aims of a research study determine, among other things, the kinds of designs, data, assumptions and standards of success suited to the investigation. For example, by explicitly stating that the *goal* of study three is to investigate the

[1] This does not imply, however, that empirical research in social and behavioral sciences *should* be causal in nature. In fact, in study one I suggest that descriptive research is often neglected or underestimated, and some researchers have put forward similar arguments regarding predictive research (e.g., Shmueli, 2010).

causal effect of executive functions on academic achievement, I imply, among other things, that (1) a central aim is to obtain unbiased parameters, which will be endowed with a causal interpretation; (2) the estimated parameters might be biased due to several reasons (e.g., confounding, selection or overcontrol bias); and (3) statistical control and weighting are likely to be needed in order to address some of these biases. It is worth noting that none of these issues would apply if the goal of the investigation was entirely descriptive or predictive. One can see, then, how acknowledging the goal of an investigation can have important implications for the transparency and testability of empirical research (e.g., Hernán, 2019; Pearl, 2019). In addition, one can see how explicitly using causal language can actually help restrict excesses in interpretation (e.g., by clearly implying that the presence of unobserved confounding might impede the interpretation of the estimated coefficients).

The crucial dilemma, however, is that while the aims of most of the research in social and behavioral science is causal in nature, only relatively few well-conducted experimental or quasi-experimental designs can justifiably refer to their findings as causal. I have argued (following other researchers, e.g., Bollen & Pearl, 2012; Hernán, 2019), that the upshot of this predicament should not be the proscription of causal terminology in all the other non-experimental (or quasi-experimental) designs. On the contrary, as I argue in study three, if the goal is to answer a causal question using observational data, then the study can only benefit by explicitly adopting a causal inference framework.

The contention, rather, is that one should explicitly recognize that a particular line of investigation is causal, while properly acknowledging the limitations of observational data. As Hernán (2019) remarks, this implies that "[t]he only part of the article in which the term 'causal effect' has no place is the Results section, which should present the findings without trying to

interpret them." That is, one should avoid interpreting the estimated coefficients in an observational study, given the likelihood of unobserved confounding and other sources of bias. Instead of focusing on interpreting the coefficients, one can focus on weaker but more credible claims (e.g., that some mechanisms appear to be more proximal than others; see study two), or in recognizing and trying to reduce remaining sources of bias (see study three).

In conclusion, the aim of studies two and three of this dissertation is to contribute to explanatory (i.e., causal) research in skill development. One can easily ratify this statement by realizing what these studies do *not* aim at: for example, they do not intend to summarize the observed patterns in the data in particular ways (descriptive research), nor implement methods for maximizing the predictive accuracy of a target variable (predictive research). By clearly stating that the goal of these studies is causal, I indicate, among other things, that the risk of confounding (which is a causal concept) needs to be properly recognized. As suggested above, I do this by (1) avoiding causal terminology in the Results section; (2) avoid interpreting the estimated coefficients; (3) focusing on weaker but more credible claims; and (4) clearly recognizing remaining sources of bias.

The three studies in this dissertation follow, then, a common progression in scientific research, from descriptive to explanatory models. Below, I discuss in more detail how each of the three studies contributes to the existing literature on skill development.

1.3 The concept of academic mobility

The first study presents a general framework, referred to under the label "academic mobility", for describing the development of educational inequalities using student learning outcomes. For this purpose, the paper (1) discusses some general normative considerations that

can be adopted to operationalize the concept of educational inequality; (2) reviews some concepts of change (e.g., absolute and relative change) which are important to consider in order to measure educational inequality; and (3) discusses and implements five mobility metrics that have been used in different fields (e.g., economics and psychology) and which can be adopted to measure educational inequality by investigating academic mobility.

This study contributes to the literature in several respects. Conceptually, it clarifies some notions that are important to consider for describing educational inequalities using student learning outcomes. In particular, it explains that educational outcomes have a positional dimension, and that in order to measure the development of inequalities one needs to consider the extent to which individuals' *relative* position changes over time. This perspective differs from traditional approaches, typically intended to measure "growth" (i.e., natural change) or "learning gains" rather than changes in the individuals' relative position in the achievement distribution.

Methodologically, the study discusses some of the limitations of traditional methods for measuring relative change, and presents five mobility metric that overcome some of these limitations. These metrics can be used to answer several questions that cannot be easily answered with traditional mean-based methods, such as: (1) What is the overall degree of mobility of a particular education system? (2) How mobile or persistent are the individuals at the bottom and at the top of the distribution? (3) Are there differential mobility patterns across groups (e.g., racial groups)? (4) How does academic mobility change throughout schooling? And (5) To what extent does schooling serve as an equalizer? The mobility metrics presented shed some light on all of these questions.

Finally, the metrics presented also allow us to relax some important assumptions of traditional methods, in particular related to the interval scale of the underlying test scores and to parametric assumptions related to the functional form of growth processes.

1.4 The structure of academic achievement

There is a vast literature examining the contextual factors affecting academic achievement. The factors investigated range from individual-level constructs such as motivation and so-called non-cognitive skills, to a wide-range of contextual influences, including peer, teacher, classroom, home, school, neighborhood, or policy-level effects. Typically, the goal of these studies is to identify the effect size associated to a particular factor, which would indicate whether the factor under investigation is causally associated with academic achievement, as well as how strong or weak the causal relationship is between each factor and student achievement. This study introduces a principled way for organizing this body of research; justifying the choice of a particular explanatory variable; and examining causal relationships in a "structural" or "holistic" rather than "bivariate" or "atomistic" fashion.

Responding to the limitations of bivariate associations that I describe in the paper, this study explains how causal graphs can be useful tools for (1) synthesizing, interpreting and using research findings; (2) assessing causal modelling assumptions; (3) generating accurate predictions with a minimal set of predictors; and (4) supporting claims regarding the explanatory relevance of particular variables. Regarding this final point, I argue that a principled and effective strategy to analyze complex organism-environment interactions –such as the contextual effects on academic achievement– is to identify the immediate causes or proximal mechanisms of the variable of

interest. Given the complexity of the contextual effects on academic achievement, I argue that these methods are relevant and useful when applied to skill development research.

Even if many researchers agree on the importance of causal structural knowledge, it is not clear how to construct such graphs. In this study I implement several causal search algorithms to identify the causal graph using the observed conditional and unconditional independencies in the data. Consistent with previous research, the algorithms identified prior achievement, executive functions (in particular working memory, cognitive flexibility and attentional focusing) as well as motivation as direct causes of academic achievement.

I also discuss the assumptions and limitations of the causal search algorithms employed in this study. In particular, I explain that these algorithms assume the absence of hidden confounders (the so-called causal sufficiency assumption). This is an important limitation, given that in observational studies it is almost certain that important confounders have been omitted. However, the estimated graphs can provide valuable insights even if we drop the causal sufficiency assumption. In particular, the results can be used to identify proximal and distal relationships among the measured variables, reveal which observed variables can be potential confounders, mediators or moderators, generate causal hypotheses, and indicate which causal relationships do not exist.

1.5 Identifying proximal mechanisms

In many causal studies mechanistic hypotheses play a secondary role. Typically, researchers are mainly interested in the causal effect of some particular variable; after this causal effect has been identified, researchers might examine mechanistic hypotheses in order to

investigate potential effect heterogeneity (e.g., through interaction analyses), or potential mediating paths using mediation analysis (e.g., in order to explain or provide additional support to the main findings). In the context of mechanism-based explanations, this "outside-in" approach can be used to identify instruments that can be employed to identify the causal effect of the remainder of the causal chain (Knight and Winship 2013).

An alternative "inside-out" strategy is represented by the study of proximal mechanisms. This approach has several advantages compared to the "outside-in" method. As I explain in the paper, identifying proximal mechanisms would help investigate and explain the phenomena under investigation; guide effective interventions; and obtain stable and accurate predictions of the target variable.

A consistent finding in this study is that, in the panel dataset considered (ECLS) and employing several different model specifications (e.g., with or without demographic variables; with or without prior achievement; with or without the two achievement outcomes; and using different search procedures with several bootstrapped samples), executive functions (in particular working memory, cognitive flexibility and attentional focusing) are directly related to academic achievement. It is important to note that these results intend to shed some light on the causal structure of academic achievement, i.e., they intend to indicate how the observed variables might (or might not) be causally related. In particular, they suggest that manipulating the value of executive functions affects the value of academic achievement. However, typically we are not only interested in whether two variables are related, but also on how strong or weak the relationship is. How much would achievement change if we manipulated the value of executive functions? Put differently, what is the magnitude of the coefficient describing this relationship? This is the question addressed in the third study of this dissertation.

1.6 Thinking within-persons

Even though numerous studies have estimated the effect of executive functions on academic achievement, the literature presents contradictory results. In addition, there is a lack of clarity regarding what modelling strategies should be preferred, and the extent to which the estimated coefficients can be interpreted as "causal." In this study, I discuss some of the possibilities and limitations of describing causal mechanisms using observational data –in particular longitudinal data.

The paper begins with the premise that educational and psychological research is mainly concerned with causal questions, and as a consequence can benefit from explicitly adopting a causal inference framework. Consequently, I review some basic concepts in the causal inference literature related to the definition, identification and estimation of causal effects. I also explain that the main contribution of longitudinal analyses, from a causal perspective, is the ability to control for time-invariant unobserved heterogeneity, which can be achieved by focusing exclusively on within-person variation. I review different procedures to estimate within-person effects, as well as different modelling strategies that can be used to test substantive hypotheses regarding within-person asymmetrical causation, moderation, effect heterogeneity, and reciprocal causation.

I provide an empirical illustration of the methods presented by estimating the within-person effect of executive functions on academic achievement. The results of the most robust model implemented suggest that there is a statistically significant effect of executive functions on math but not on reading achievement. More generally, this study illustrates how different models generate substantially different results, which highlights the importance of clarifying the possibilities and limitations of causal inferences using observational data.

1.7 General contribution

Due to the complexity and interdisciplinary nature of the topic, skill development research appears to be a fragmented field, composed of multiple –and often unconnected– bodies of research, and a dizzying range of theoretical and methodological approaches (see, e.g., Hattie, 2009; Pfost, Hattie, Dörfler & Artelt, 2014; Smithers et al., 2018). The purpose of this dissertation is to discuss some conceptual and methodological tools for investigating skill development in a principled, systematic and coherent fashion, following a common progression in scientific research. The first paper provides an original framework for describing individual differences in skill development. The purpose of this paper is to identify relevant information in the data that further explanatory work should try to explain. The second paper presents a framework for justifying the selection of explanatory variables, and in the third paper I estimate the causal effect of such variables on achievement.

2.0 Study 1: The concept of academic mobility: Normative and methodological considerations[2]

Most of the literature on the development of educational inequality has operated under the achievement gaps paradigm, often assuming that the underlying normative and methodological foundations related to equality and justice in education are a settled matter. In this paper, we argue that important normative dimensions are overlooked with traditional mean-based measures, and that metrics that capture students' academic mobility as they progress through school can provide the informational base needed to describe and evaluate these policy and socially-relevant aspects. We discuss some key normative principles and methodological dimensions related to academic mobility, and provide an empirical example of the mobility metrics presented using a nationally representative dataset.

2.1 Introduction

The idea that all individuals should have equality of opportunity regardless of their social origin is a fundamental principle in modern democratic societies. In addition, it is commonly believed that education plays a key role in achieving this goal, as schooling processes should provide equal opportunities to all individuals, allowing them to move up or down the social ladder

[2] Manuscript published in the American Educational Research Journal; see Quintana & Correnti, 2020. The final version is available at https://journals.sagepub.com/home/AER).

based on their efforts and accomplishments, rather than their family background and other factors that lie beyond the individuals' control. Academic and policy discussions on education regularly examine educational inequalities, often assuming that these normative[3] foundations, as well as the metrics appropriate to measure educational systems on these dimensions, are a settled matter. In particular, within-nation analyses on educational inequality have been to a large extent dedicated to measuring the so-called "achievement gaps", which refer to the differences in mean academic performance between different groups of students (defined, for example, by race, gender or socioeconomic status). The implicit normative implication seems straightforward: the larger the gap, the less equality of opportunity among groups of individuals. Discussing or making explicit these normative assumptions can be regarded as stating the obvious, and as a consequence many researchers focus instead on methodological variations of measuring achievement gaps, which are assumed to be the main metric of educational inequality.

However, the normative implications related to the concept of equality of opportunity in general, and the concept of equality of educational opportunity in particular, are far from settled. For a long time, researchers working on issues of justice and inequality have called attention to the tensions and ambiguities associated with egalitarian ideals. These discussions have been held primarily by philosophers (e.g., Anderson, 1999; Cohen, 1989; Rawls, 1971) and economists working on issues of inequality (e.g., Romer, 1998; Romer & Trannoy, 2015; Sen, 2009).

[3] Normative statements can be divided in two kinds (see Thomson, 2008): (1) evaluative judgments, which make claims about the value (or disvalue) of a state of affairs (e.g., equality in achievement outcomes is desirable/not desirable); and (2) directive judgments, which state that something or someone ought (or ought not) do something (e.g., educational systems should/should not promote equality of opportunity). In this paper, we use the term "normative" to refer to "value" judgments broadly defined, i.e., as encompassing these two kinds of statements.

Researchers in the field of education (e.g., Coleman,1966; Jencks, 1988) have also distinguished different meanings associated with the concept of educational opportunity, and noted some prominent dilemmas that arise from this ideal. This concept can be understood in different ways, implying different ideals of equality and opportunity, and therefore different metrics to examine the distribution of educational outcomes. These underlying conceptual disagreements, as well as their methodological and practical implications, might not be immediately evident to many researchers and policy makers in the field of education. As Jencks (1988, p.518) puts it, "the enduring popularity of equal educational opportunity probably derives from the fact that we can all define it in different ways without realizing how profound our differences really are."

A central idea underlying the previous discussion is that, as many researchers in the field of economic inequality have recognized, measurement practices are inevitably entangled with normative considerations. Several scholars have argued that, when dealing with social science theories (which normally involve thick ethical concepts such as fairness and inequality), one cannot separate a purely descriptive part and a purely evaluative part, as the researcher simultaneously describes and evaluates (e.g., Anderson, 2004; Putnam, 2002; Schumpeter, 1949).[4]

[4] Similar considerations apply to psychological and educational assessments, which should not be conceived as neutral descriptions of the examinees' behaviors, but an entanglement of the latter with the normative expectations of the test designer. As Messick (1994, p. 13) remarks, "validity, reliability, comparability, and fairness are not just measurement issues, but social values that have meaning and force outside of measurement wherever evaluative judgments and decisions are made." This is the reason why test designers should make an effort to explicitly state and justify their value judgments. A useful framework in this regard is provided by the evidence-centered design approach, which indicates that a measurement procedure should be based on an explicit cognitive model of the construct that is being measured (Mislevy, Steinberg & Almond, 2003).

Furthermore, as Putnam (2002, p.112) notes, there are facts "which only come into view through the lenses of an evaluative outlook." What this implies, then, is that social science theories should be regarded as an inextricable combination of a normative framework and a measurement model, and that both of these components should be explicitly stated and justified (e.g., Anderson, 2004; Feuer, 2015). In the case of educational inequality, this implies answering a series of related questions, notably: What normative principles should we adopt to describe and evaluate the state of our educational systems? And, in view of these principles, what features of the distribution of educational outcomes become important to measure, and how can we measure them?

In this study, we argue that the notion of achievement gaps provides a limited informational base to assess educational inequality, and that the concept of academic mobility, which refers to changes in individuals' rank position over time, can help us overcome some of these limitations. Policy and socially-relevant aspects that are not captured by average-based differences include: (1) achievement patterns at particular segments of the achievement distribution (e.g., the upper and lower ends); (2) existing gaps and distinct mobility patterns within groups; and (3) differences and similarities in patterns of within-person variation. Additional limitations include a restricted generalizability (as they are only interpretable by reference to another group), and susceptibility to problems of scale. We will explain these points in detail, and show how academic mobility metrics can help us overcome these limitations.

The objective of the study is to discuss appropriate ways of describing and evaluating inequalities in an education system. The purpose then is not to explain why educational inequalities occur, or to identify factors that predict differences in educational outcomes. In other words, the aim of the present study is entirely descriptive (rather than, for example, explanatory or confirmatory). In addition, it is worth noting that the "education system" we refer to in our theoretical and empirical sections is the national education system. The academic mobility metrics that we present can belong

then to what Reeves (2015) calls the "measure of a nation." At the same time, these metrics can be used to describe and evaluate other systems (e.g., particular schools), and they can be adapted to account for hierarchical structures (e.g., patterns of educational inequalities within and between classroom, schools, districts, etc.). However, for the purpose of simplification we focus on the national education system monolithically conceived.

Finally, our objective is to examine educational inequalities in skill development, and our empirical application considers reading achievement. Most of the empirical research on educational inequality has focused on access and completion (e.g., Lucas, 2001; Bailey & Dynarski, 2011), school and college transitions (e.g., Breen & Jonsson, 2000; Page & Scott-Clayton, 2015), track or curriculum stratification (e.g., Lucas & Berends, 2002; Van de Werfhorst & Mijs 2010), and intergenerational mobility (e.g., Breen & Jonsson, 2005). Researchers have also investigated inequalities related to learning outcomes (generally recognized as an important –or even the main– outcome of educational processes), mostly by conducting achievement gap analyses. In this study, we reflect on the appropriateness and limitations of achievement gaps and other measures commonly used to examine learning outcomes to describe and evaluate educational inequalities in skill development.

The paper is structured as follows. First, we argue that academic mobility provides an adequate normative framework for describing and evaluating educational inequalities. Second, we present methodological considerations along with different metrics that are appropriate for measuring academic mobility. Finally, using a nationally representative sample, we provide an empirical example of the metrics by conducting a comparative analysis of academic mobility based on racial groups.

2.2 The normative dimensions of academic mobility

2.2.1 Defining equality of opportunity

Egalitarian concerns have been at the center of modern social and political debates since John Rawls's published his *Theory of Justice* in 1971. In his groundbreaking book, Rawls challenged the dominant ethical view of utilitarianism and argued that, in a just society, individuals should be treated as equals in some specific respects: everyone should have equal basic liberties and fair equality of opportunity (Rawls, 1971). The precise meaning and acceptability of these principles of justice have been widely discussed in the literature (see, e.g., Anderson, 1999; Cohen, 1989; Romer, 1998). However, most researchers working in the egalitarian tradition agree on a fundamental point, namely that one should distinguish between morally acceptable and unacceptable forms of inequality (Romer & Trannoy, 2015). In particular, many authors have moved from expecting equality of outcomes to expecting equality of opportunities, generally defined as "chance[s] of getting a good if one seeks it" (Arneson, 1989, p.85). The rationale for this idea is that while achieving equality of outcomes would presumably imply severe limitations on individual liberties, equality of opportunity serves the egalitarian ideal and at the same time preserves individual freedom and responsibility (Cohen, 1989).

Theories of equality of opportunity have generally tried to specify which are the tolerable and the intolerable sources of inequality. In an influential account, Roemer (1998) differentiates between "circumstances" (i.e., exogenous factors that lie beyond the individual's responsibility such as race or gender) and "accountable effort" (i.e., endogenous factors deriving from the individual's choices), and argues that only the inequalities deriving from accountable effort should be tolerated. Even though this approach provides a powerful framework for defining and

measuring educational inequalities, it also faces considerable challenges (e.g., Kanbur & Wagstaff, 2016). For example, it is unlikely that researchers can identify and measure all relevant circumstances, and it is not clear at what point we should begin attributing disparities in outcomes to differences in accountable effort (as differences among infants and young children mainly reflect the parents' circumstances and effort; see, e.g., Bodovski & Farkas, 2008; Lareau, 2003).

These complications point to the difficulty of articulating a positive theory of justice in general, and of equality of opportunity in particular (see, e.g., Jencks, 1998). Providing a normative framework that establishes in a clear-cut fashion the ideal distribution of resources based on individuals' circumstances and personal choices (or other exogenous and endogenous factors) may be both theoretically and practically unattainable. In order to overcome this conundrum, Sen (2009) argues in favor of a "relational" approach to justice (in opposition to the "transcendental" view espoused by many egalitarians), according to which the aims of social justice should consist in removing the manifest injustices rather than in conceiving an abstract ideal of justice. That is, rather than advancing a positive theory of justice, we should focus on the negative aims of egalitarian justice (Anderson, 1999). By assigning epistemological priority to social inequalities, the central goal becomes to bring attention to injustices which any reasonable egalitarian theory would condemn. As Sen explains, instead of speculating on what a perfectly just society would look like, a relational approach produces comparative assessments intended to identify clear cases of social asymmetry and discrimination. One advantage of this approach is that, in contrast to the distant ideals offered by the transcendental tradition, the identification of clear cases of manifest injustice can lead to more targeted policy interventions, as well as a clearer diagnosis of the current situation.

The goal of this paper is to reflect on the comparative assessments that are appropriate for describing and evaluating educational inequalities, and which might allow us to identify cases of manifest injustice. The literature on educational inequality has mostly operated under the achievement gaps paradigm, comparing averaged-based measures between groups at different stages of development and using a variety of metrics and research designs (see Reardon, Robinson-Cimpian, & Weathers, 2015, for a survey of the achievement gap literature). Frequently, however, researchers do not explicitly state their normative assumptions and implications for their analyses. Value judgements regarding features of the educational system that can be considered fair or unfair are often embedded in the methodological sections, and are not explicitly recognized and justified. Thus, it is not clear how to interpret –from an ethical standpoint– achievement differences between groups, and what other ethically-relevant aspects are not captured by analyses within this paradigm.

In order to shed light on these issues, we discuss key normative principles that can help us distinguish morally acceptable and unacceptable forms of educational inequality. Similar to other theories of equality of opportunity, we start from the premise that a completely unequal society is one in which circumstances (e.g., race or socioeconomic status) completely determine individuals' outcomes. Unlike many of those theories, however, we do not espouse a particular theory of justice nor a particular distributional ideal. Instead, we adopt Sen's relational view of justice, and focus on comparative assessments to help us identify manifest disadvantages. Identifying these disadvantages requires moving beyond average-based differences between groups, and implementing fine-grained comparative analyses of how students traverse through the achievement distribution over time. We argue, then, that the concept of academic mobility can provide an appropriate normative framework for describing and evaluating educational

inequalities, as well as offering a suitable operationalization of our intuitive ideal of "equality of opportunity."[5]

2.2.2 Measuring educational inequality: Normative considerations

In this section, we argue that the concept of academic mobility, which has not been utilized in educational research (with very few exceptions, notably Feinstein [2003], Kerckhoff [1993], McDonough [2015] and Sohn [2010]), provides a valuable normative framework for evaluating and describing potential inequalities within education systems. We support this claim by showing how the concept of academic mobility is sensitive to particular forms of inequality that can be regarded as more concerning and unacceptable than others. In particular, we discuss the normative implications of two well-established dichotomies in the literature: between versus within-person variation, and relative versus absolute change.

2.2.3 Between versus within-person variation

Many researchers and practitioners in the field of education recognize that educational processes are better described by metrics that capture students' performance over time rather than "status" or cross-sectional measures. That is, instead of focusing on the differences in performance

[5] It is worth stressing that our claim is not that average-based differences are somehow invalid or ineffective, and that academic mobility metrics should be used instead. Our intention is rather to discuss an approach that can be used to complement these measures, which have been widely and successfully employed to measure educational inequalities.

of an individual (or group of individuals) at a single point in time (commonly referred to in the education literature as "gaps"), one needs to consider how these differences change (or do not change) over time. An important normative reason behind this perspective is that, in order to examine educational inequality, one needs to understand to what extent inequalities increase or decrease throughout the schooling process, above and beyond preexisting differences. Cross-sectional or "status" measures are, then, inherently biased, as they reflect both schooling effects (or, more generally, inequalities that increase or decrease throughout schooling) as well as preexisting inequalities.

For purposes of describing and evaluating educational inequalities, it is important to examine, then, the magnitude and direction of individuals' deviation from their starting position. The average amplitude of these changes in the population would provide a measure of the overall degree of mobility (or stratification) in the education process (or, put differently, the degree to which individuals' final outcome is determined by their initial position). For example, a system where all individuals remain in the position they started in throughout the schooling process would be completely stratified, whereas a system in which the individuals' initial and final positions are fully uncorrelated will have total mobility. Furthermore, given that these measures control for the individuals' initial status, differential patterns of mobility (e.g., groups of individuals who tend to move downward) can be particularly concerning and could indicate clear cases of social asymmetry and discrimination.

The variability describing individuals' deviation from their own baseline levels, regardless of the baseline values *per se*, is generally referred to in the literature as within-person (or intraindividual) variation (e.g., Hoffman, 2015). On the other hand, the variability describing differences between individuals, independently of the degree to which the individuals themselves

vary over time, is generally referred to as between-person (or interindividual) variation. Researchers have long argued that we should clearly distinguish between these two kinds of variation, as they represent different phenomena, and inferences from one level of analysis to the other are seldom warranted (e.g., Hamaker, 2012; Molenaar, 2004). In particular, group generalization can obscure genuine individual differences (Fscher, Medaglia & Jeronimus, 2017), and group-derived estimates can misrepresent individual-level processes (commonly referred to as the "ecological fallacy"; e.g., Curran & Bauer, 2011). For example, based on six studies with a repeated-measure design, Fischer et al. (2017) found that the variance within individuals was two to four times larger than the variance within groups. This suggests that we might fail to identify important differences in individuals' variation if we rely exclusively on group-level estimates.

Many researchers have defended the priority of within-person variation for adequately describing processes of change (e.g., Hamaker, 2012; Nesselroade & Molenaar, 2010). As explained above, examining differences in intraindividual variability (i.e., the degree to which individuals' deviate from their initial position), is also essential for descriptive efforts related to educational inequality. The average-based differences commonly used in the achievement gap literature represent useful metrics for summarizing achievement disparities between groups. However, mean-based measures do not capture patterns of intraindividual variation, and as a consequence can obscure important individual-level processes. A comprehensive description of the development of educational inequalities should address within and between-level phenomena, and explain how processes that take place at the within-person level differ across individuals (and groups of individuals). Some of the academic mobility metrics that we discuss below are sensitive to within-person variation, while others to between-person variation; a comprehensive application

of different metrics can yield, then, a better understanding of the development of individual and group differences in academic achievement.

2.2.4 Relative versus absolute measures of change

In the context of educational processes, the dichotomy between absolute and relative change is based on a fundamental difference between the properties that are being measured: while relative measures refer to relational properties, e.g., percentile scores, absolute measures refer to intrinsic properties, e.g., a particular ability θ. This distinction mirrors the difference in the psychometric literature between norm-referenced reporting, in which student's performance is described relative to a reference group (i.e., in terms of their position with respect to the group), and criterion-referenced reporting, in which student's performance is described according to some fixed level of performance, independently of their position in the group (Hamilton, 2003). Since Glaser (1963) argued in favor of criterion-referenced reporting for describing students' achievement, this kind of reporting has become increasingly common in education, and performance standards have been the cornerstone of educational reform (Hamilton, 2003; Shepard, Hannaway & Baker, 2009).

"Pure" measures of absolute change are provided by studies assessing growth trends in particular competencies using vertically scaled assessments. For example, using nationally representative assessments, Hill et al. (2008) computed the difference of mean scale scores in adjacent grades, in order to characterize the natural developmental progress in achievement made by students from one year to the next. According to the authors, these differences can serve as normative expectations and benchmarks for achievement effect sizes in educational interventions. This goal is justified, inasmuch as one is interested in providing a measure of pure absolute change;

that is, a description of "natural" (in the sense of "shared" or "structural") growth, with normative implications that emerge from these common stages of development.

For many purposes it is useful to use absolute measures, as researchers and practitioners are frequently interested in the knowledge, skills and dispositions that children have acquired at particular moments in their education (see Pallas [2000] for a review on the many valuable non-positional dimensions of educational outcomes). It might also be important to attach normative implications to absolute measures, for example by defining proficiency cut scores. However, even though absolute measures can provide the foundation for certain normative claims, in order to measure educational inequality, it is essential to consider the relational or positional aspects of academic achievement.

A good has a positional dimension if an individual's relative position in the distribution of the good affects the value of the good (Brighouse & Swift, 2006). Education is often considered a paradigmatic case of a positional good, as the value of one's education *partly* depends on how well-educated other people are (Anderson, 2007; Brighouse & Swift, 2006; Nikolaev, 2016; Solnick & Hemenway, 1997). Health outcomes, in contrast, are often considered non-positional goods, as the conception of one's health is largely independent of the health of others. As Brighouse and Swift (2006) explain, the positional aspect of education is derived from its competitive component: our K–12 education system is instrumental in accessing a range of valuable goods (e.g., a good higher education, an interesting job, etc.), and our competitive success in these markets depends to a large extent on our relative position in the overall distribution. Furthermore, schooling processes play a critical role in sorting individuals into different positions, exacerbating or compensating social inequalities, and shaping students' identities as well as their

future educational and non-educational experiences (see, e.g., Domina, Penner, & Penner, 2017; Downey & Condron, 2016).

Acknowledging the positional aspect of academic achievement has important implications for the study of educational inequality. Primarily, it implies that the value of one's education is not only determined by the specific skills and knowledge one has acquired (or other non-positional aspects), but also by one's relative position in the overall distribution. In addition, different normative concerns apply to different parts of the distribution. Following Rawls' (1971) theory of justice, priority should be given to promoting the opportunity of the worst-off group. That is, analyses of educational inequality should focus on the lowest segments of the distribution, and examine to what extent some groups face particular barriers that prevent them from moving up, and persist in disadvantage. At the same time, other important normative concerns apply at the top of the distribution. Following Anderson (2007), one can argue that democratic elites should be drawn from all sectors of society, including the less advantaged. Analyses of educational inequality should examine, then, to what extent some groups are underrepresented at the top of the distribution (which comprises the individuals who have higher chances of getting accepted to the best higher education schools and the highest positions).

2.2.5 Defining academic mobility

We have argued that in order to more precisely examine educational inequality, researchers should consider metrics that fulfill three key conditions. First, the measures used need to reflect how the initial differences between individuals (or groups of individuals) change over time. Second, these metrics should capture individual and group differences in intraindividual variability. Third, these metrics need to be sensitive to the individuals' relative position in the

overall distribution, rather than to absolute properties reflecting natural or shared growth. We will refer to the processes described by these metrics as "academic mobility". In a nutshell, these metrics describe the change in individuals' relative position in the overall achievement distribution over time.[6]

Based on the previous discussion, one can identify additional desirable features of metrics describing academic mobility. First, one should be able to identify mobility patterns at different segments of the academic distribution, and in particular at the extremes of the distribution –this is due to the importance of monitoring the composition of democratic elites (i.e., the top of the distribution) as well as the worst-off group (i.e., the bottom of the distribution). Second, one should be able to make comparisons across time and between different groups and populations. This is supported by the relational theory of justice and opportunity presented above, in which investigations of inequality should be regarded as an essentially comparative exercise intended to identify cases of social asymmetry and discrimination. Third, given that the ultimate goal is to provide a clear diagnostic of a relevant issue, as well as helping policy makers make informed decisions for improving educational equality, the metrics should be transparent and easy to understand. This is related to the intended use of measurement procedures, and the key role of

[6] The reason for using the term "mobility", rather than more commonly employed terms in education such as "growth", is that, contrary to the latter, this term connotes relative position with respect to the overall distribution. In addition, the methodological approaches that we present below are mostly used in studies of social and economic mobility.

consequential validity (see e.g., Shepard, 1997).[7] Finally, it is worth noting that given that academic mobility is multidimensional, one should consider different metrics that are sensitive to different aspects of mobility. In the next section, we will describe the methodological aspects that allow mobility metrics to fulfill these conditions.

2.3 Measuring educational inequality: Methodological considerations

2.3.1 Using growth models to measure relative change

Even though in theory one can differentiate between pure measures of relative and absolute change, in practice many measures combine in intricate ways the two dimensions. It is useful, however, to keep in mind this dichotomy, as it represents different kinds of processes with distinct normative implications, which require different measures and modelling strategies. The objective of absolute measures is to quantify the magnitude of change of a particular construct over time, generally referred to in the educational literature as "growth". The construct of interest is assumed to follow a common developmental trend, but some individuals "grow" faster or slower than others. Given that all individuals follow a common developmental pattern, it is sensible to define a "normal" growth process, as well as "abnormal" departures from the common trend. A classic example of absolute measures of change are the height and weight charts used by pediatricians to

[7] In this case, consequential validity might be related, for example, to the capacity of using these metrics for identifying critical areas of intervention, or being able to monitor the effects of particular policies on educational inequality.

monitor infant health and development. As this example illustrates, absolute measures have the following properties: (1) they assume that most individuals grow, albeit at different rates (i.e., there is a "main effect" of time); (2) they support normative claims related to what is "normal" or "abnormal" at a particular developmental stage (e.g., a normal height for a 5-year old might be abnormal for a 10-year old); and (3) they require an underlying score scale with interval properties (e.g., the difference between 30 and 40 inches is the same as the difference between 40 and 50 inches).

The purpose of relative measures, on the other hand, is to describe the change in individuals' position in some particular distribution over time. Contrary to absolute measures, relative measures (1) do not describe common developmental patterns (or growth), and are thus unconcerned about any "main effects" of time; (2) do not support normative claims regarding what is "normal" or "abnormal", but rather about the advance or decline of equality and justice; and (3) do not have to rely on interval scales, as the purpose is not to quantify the magnitude of growth but to describe changes in the individuals' relative position (which only requires an ordinal scale).

Shared or structural development processes can be adequately described by traditional growth models (e.g., hierarchical models or latent growth curve models), as in this case one can assume that all individuals grow according to the same function but their growth varies in magnitude (Raudenbush, 2001). In other words, these models are appropriate for measuring absolute change, as they assume that all individuals in the population follow a similar functional form of development, and that the variance of the growth factors are sufficient to capture interindividual differences in change across time (Nagin, 2005).

Traditional growth models can also be used to measure relative change, for example by examining the intercept-slope covariance (see, e.g., Pfost, et al. [2014]). However, traditional

growth models are better suited for describing absolute processes of change, and the inferences one can make regarding mobility are limited. The reason for this is that relative change does not represent an aggregate process that can be described by a common function, but rather a mixture of movements in the rank ordering of individuals, net of any common developmental process (i.e., of any main effect of time). Relative processes imply that if someone moves up in the rank order, then another person needs to move down; consequently, one cannot describe relative changes by a common developmental function, as these changes involve opposite developmental patterns, and the common function describes the average change. In other words, absolute measures are not necessarily sensitive to mobility patterns in the population, as the average gains can remain the same independently if there is high or low mobility. Furthermore, mobility studies are often interested in obtaining descriptions of the movement occurring in the edges of the distribution, whereas traditional growth models are well-suited for describing average trends.

Traditional growth models quantify the magnitude of change by relying on a measurement scale with interval properties (i.e., they require vertically scaled assessments). However, in the context of skill development this assumption is often untenable (Briggs & Betebenner, 2009; Ho & Haertel, 2006). In addition, as some studies have demonstrated (Bond & Lang, 2013; Ho & Haertel, 2006), the estimated statistics can change dramatically and even change sign under some monotonic (i.e., order-preserving) transformations –which could represent the "true" underlying scale. In order to deal with this issue, various modelling strategies have been proposed that do not require vertically scaled assessments (see, e.g., Castellano & Ho, 2013).[8]

[8] A widely-used model that does not assume interval scale properties is Betebenner's (2009) Student Growth Percentile (SGP), which compares students' current achievement scores to the scores of students with a similar

2.3.2 Using achievement gaps to measure relative change

Educational research and policy have focused primarily on student growth –which is demonstrated by the emphasis on criterion-referencing reporting and the extensive use of growth models of different kinds. On the other hand, the most common measures employed in the literature to assess relative change (i.e., the change in individuals' position over time) are the average-based differences coming from the achievement gap literature (e.g., Fryer & Levitt, 2004; Reardon & Galindo, 2009).[9] These studies normally use standardized score differences, obtained by dividing the difference in mean achievement scores between two groups by the pooled standard

academic history. This model belongs to a wider class of models that quantify growth based on the individuals' deviation from some predefined expectation or conditional distribution (Castellano & Ho, 2013). In the case of SGPs, the expected test performance is defined using the students' past test scores. Contrary to traditional growth models, SGPs do not report learning gains but rather percentile ranks that compare students to their academic peers. Still, as the name suggests, the main objective of SGPs is to quantify student growth, and as a consequence can be regarded as a growth measure (Castellano & Ho, 2013). Thus, as Betebenner (2009) remarks, the purpose of SGPs is to support normative claims regarding normal or abnormal growth (rather than, e.g., to support claims related to the equality of an education system).

[9] Other alternatives for examining relative change include metric-free measures (see Ho & Haertel, 2006) and group-based or mixture modelling (see Nagin, 1999, 2005; Muthen, 2004). However, these methods have some limitations. On the one hand, the former can be used only to measure between-group (rather than within-group) differences, and these measures can be difficult to interpret; the latter, on the other hand, can be used to identify clusters of individuals with similar trajectories, but they do not provide clear metrics at the population or sub-population level that can be used to compare different patterns of academic mobility across groups or time.

deviation of those groups. These measures have the advantage that they are easily computed and interpretable.

Average-based measures, however, have the following limitations. First, by standardizing the mean differences, one confounds the difference in means with the variation within groups (Reardon & Galindo, 2009). Second, in most applications average-based measures depend on the interval properties of the scale for the original score (which, as indicated above, is often an untenable assumption). Third, average-based measures do not take into account the individuals' position in the overall distribution (see Figure 1, Patterns A and B). However, as noted above, it is not only important to consider the mean difference in achievement between two groups, but also the position of the individuals or group of individuals in the overall distribution. In particular, one should be able to conduct fine-grained analyses at different segments of the achievement distribution, and examine the extent to which some subgroups are more or less mobile (e.g., at the extremes of the distribution; see Figure 1, Patterns D and E).

The fourth limitation of average-based measures is that, given that they are mainly sensitive to between-person variation, they provide very little information regarding differences in within-person variability (see Figure 1, Pattern C). We know, however, that there is considerable heterogeneity within racial groups (see, e.g., Davis-Kean & Jager, 2014). Thus, it is important to consider more fine-grained analyses of change within groups. For this purpose, it is beneficial to move from the "changes-in-gaps" paradigm, based on changes in differences in group means, to a within-person study of academic mobility, which describes who moves, by how much, from where, and when does the movement occur.

Finally, average-based differences have a limited generalizability, as they are only interpretable by reference to another group. However, one might be interested in knowing the

degree of academic mobility across all groups or at the population level (e.g., in order to conduct historical comparisons within countries, or between-country comparisons).

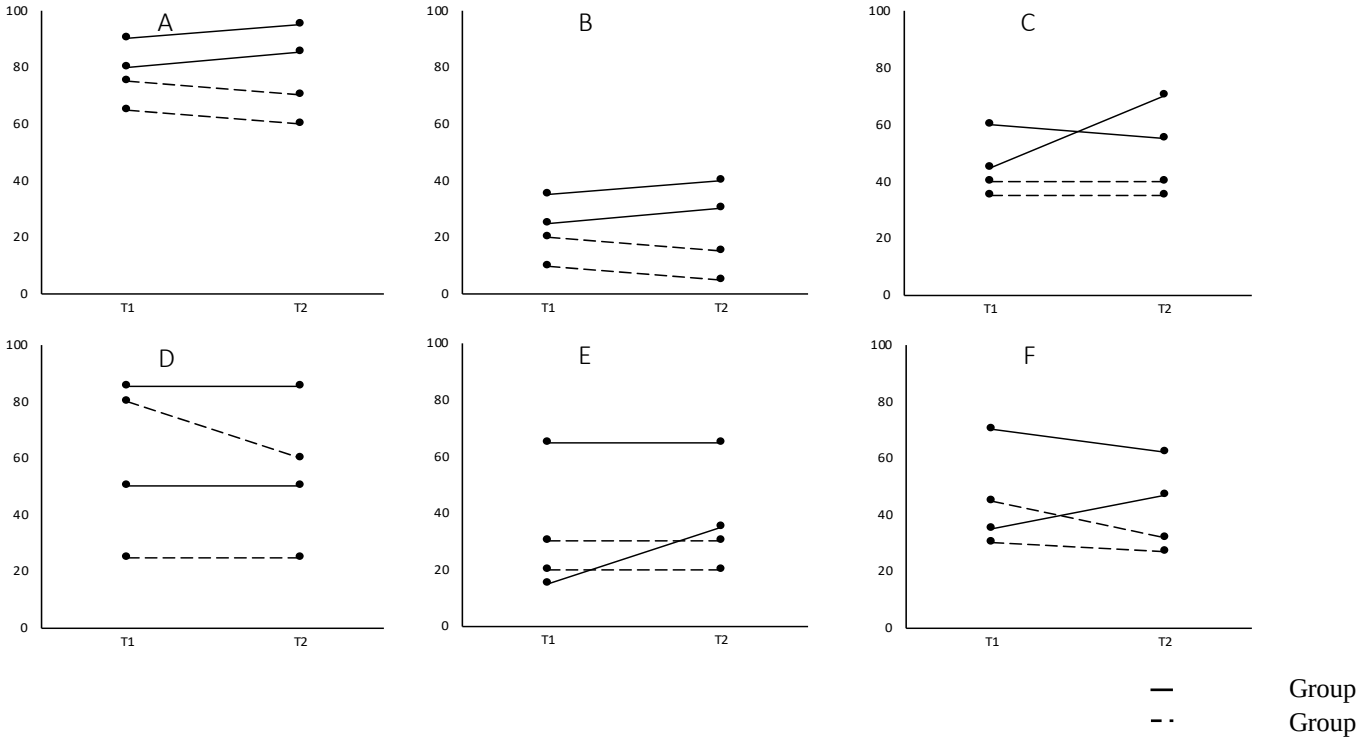


Figure 1. Six patterns representing the development of inter and intra-individual differences among four individuals belonging to two different groups

***Note*. The Y-axis represents the individuals' percentile rank in the overall distribution, and the X-axis represents two measurement occasions. In each pattern, the difference (or gap) between Group 1 and Group 2 is 15 percentile points in T1, and 25 percentile points in T2 –i.e., the gap increases by 10 percentile points. By comparing these patterns, one can perceive important dimensions that are not captured by mean-based differences. Patterns A and B show that mean-differences are not sensitive to the individuals' position in the overall distribution. Pattern C shows that mean-differences are not necessarily sensitive to within-group mobility. Patterns D and E show that mean-differences are not necessarily sensitive to downward or upward within-person mobility from different points in the distribution. Pattern F illustrates a more realistic state of affairs, where there is both within and between-group mobility. While the lines in the Figure are associated to individuals, they could also represent groups.**

2.3.3 Mobility metrics and their advantages for studying relative change

The previous discussion indicates that the majority of studies in education have focused on absolute measures of change (or, more generally, measures describing student growth), limiting their ability to identify patterns of relative change. On the other hand, studies more explicitly interested in relative change have used measures that have important disadvantages or limitations. In the next section, we present several mobility metrics that have been predominantly used in other domains (particularly in studies on income mobility and personality psychology), and which can be readily adopted to measure academic mobility. In contrast to the measures of change described above, these metrics exhibit important advantages. First, they are only sensitive to changes in the rank ordering of individuals, and are therefore unaffected by structural growth processes. This implies that the metrics are comparable across time (i.e., they are "intertemporally scale invariant"). This also implies that the results do not rely on –or, depending on the metric, are less conditioned by– functional form assumptions, which have become a hindrance for modelling developmental trajectories, as researchers try to find the best shape (e.g., complex polynomial or nonlinear trajectories; see Cameron et al., 2015, Hoffman, 2015) that fit the observed patterns of change. Second, some of the measures are "metric-free" or "scale-invariant", meaning that they are not altered by any monotonic transformation in the values of achievement, as well as changes in the marginal distributions. Thus, the metrics are comparable across assessments. Third, some metrics are sensitive to within-person variation and others to between-person variation, offering researchers different metrics that provide a comprehensive picture of the development of individual differences in achievement. Fourth, the metrics are easily computable and interpretable, which are key aspects for consequential validity.

2.3.4 Dimensions of academic mobility

Academic mobility is multifaceted, and the adequate measure depends upon one's normative objective. Based on the normative considerations presented in the first section, some fundamental questions are the following: (1) What is the overall degree of mobility of a particular education system? (2) How mobile or persistent are the individuals at the bottom and at the top of the distribution? (3) Are there differential mobility rates across groups (e.g., racial groups)? (4) How do mobility rates change throughout schooling? (5) To what extent does schooling serve as an equalizer?[10] Below we present several metrics that are adequate to measure each of these dimensions.

2.4 Data

In order to provide an empirical example of the mobility metrics presented below, we use the Early Childhood Longitudinal Study kindergarten cohort (ECLS-K), which is a nationally representative sample of 21,409 American children who entered kindergarten in 1998 (see Tourangeau et al., 2009, for more information regarding this study). The ECLS-K study followed

[10] An important question missing in this list is: What factors explain different mobility rates across individuals and groups? We did not include this question, given that the main purpose of the present study is to introduce the metrics and provide a description of academic mobility using the ECLS data. However, the metrics considered can be used as dependent variables in explanatory studies.

these children from fall of kindergarten to spring of eighth grade in 2007, providing a comprehensive picture of children's academic development until secondary school.

The choice of this dataset is driven by two considerations. First, it is to date the nationally representative sample that covers the longest time period in student's schooling experience since kindergarten. Second, there have been a large number of studies using this dataset, so one can identify the consistencies with previous results as well as inferences regarding the development of inequalities that could not have been made using traditional measures (including both measures of growth as well as those attempting to examine relative change).

In the present study, we used students' reading achievement as the main dependent variable. In particular, we used student's percentile rank in the overall distribution in six waves: fall-kindergarten; spring-kindergarten; spring-first grade; spring-third grade; spring-fifth grade; and spring-eighth grade. The appropriate longitudinal weight provided by ECLS was used in all calculations. The use of this weight allows us to (1) provide population estimates; (2) adjust for differential selection (e.g., oversampling); and (3) reduce bias associated with missing data. The percentile ranks were created using only the analytic subset (i.e., considering individuals with a non-zero and non-missing weight). Table A1 in the Appendix presents descriptive statistics of our main outcome disaggregated by race. In order to obtain adequate standard errors for the estimates, we employed in each computation the paired jackknife method utilizing appropriate replicate weights provided by ECLS.

We use students' percentile rank scores in order to stress the importance of considering individuals' relative position in the overall distribution (e.g., the top or bottom 25%). In addition, given that these scores are purely relational, they are not affected by any structural growth processes (i.e., the distribution is stationary). Importantly, however, even if some of the metrics

presented do not rely on parametric assumptions (in particular transition probabilities), the other metrics can assume linearity or can be affected by the distributional form. Consequently, studies using these metrics should test their results using other scores (e.g., T or theta-scores). A full exploration of the sensitivity of these metrics to various parametric assumptions is beyond the scope of this study.

2.5 Five mobility metrics

2.5.1 Linear rank-rank measures

A commonly used measure of mobility is obtained using a linear rank-based approach (see Chetty et al., 2014, 2018), which provides a parsimonious summary of the overall degree of mobility at the population and sub-population levels. In this approach, mobility estimates are obtained by regressing the child's rank in the national distribution at some time point on his or her rank at a previous time point. In the present application, this can be expressed as:

$$R_{i,8} = \alpha_i + \beta_i R_{i,k} + \varepsilon_i \quad (1)$$

where R_{i8} represents the achievement percentile rank of individual *i* relative to all other individuals in 8th grade, and $R_{i,k}$ the percentile rank of the same individual in kindergarten. By combining the two parameters describing the linear function (α_i and β_i), one can estimate the expected rank of children in 8th grade based on their rank at the national level in kindergarten. The intercept α_i can be used to measure the overall *direction* of academic mobility across subgroups

(either upward or downward mobility), and the slope can be used to measure the *degree* of academic mobility at the population or sub-population levels (e.g., if β = 1, then the expected mobility is zero, as the percentile rank in kindergarten would be a perfect predictor of the percentile rank in 8th grade; on the other hand, if β is close to zero, then there is high relative mobility, as the relative position in kindergarten would not be a strong predictor of the children's' relative position in 8th grade).

Provided that the relationship between children's mean ranks in 8th grade and their ranks in kindergarten is well approximated by a linear function, this approach has several advantages: (1) it provides a small set of statistics capturing positional mobility estimates at the population and sub-population levels; (2) the estimated statistics control for any structural patterns in growth (given that the marginal distribution for both R_k and R_8 are uniform distributions); and (3) as Mazumder (2015) remarks, by using percentile ranks at the population level this specification can be used to measure mobility differences across subgroups with respect to the national distribution (while using scale scores would allow us to estimate mobility differences with respect to the subgroup mean).

2.5.2 Measures of the amplitude of academic mobility

The mobility estimates obtained using the rank-rank approach only consider the mobility between two time periods, and do not distinguish between and within-person changes. In this section, we present two raw-score metrics for quantifying the amplitude of within-person mobility: the intraindividual standard deviation, representing the mean intraindividual mobility; and the means square successive difference, representing the mean intraindividual mobility from one time point to the next (Wang, Hamaker & Bergeman, 2012). These indicators can provide valuable

information regarding the magnitude, direction and timing of intraindividual mobility. As explained above, these individual-level processes can be obscured or misrepresented by group-level estimates.

The intra-individual standard deviation (ISD) is often used in psychology as a measure of intraindividual variability (Wang, Hamaker & Bergeman, 2012). In the current application, it is computed as follows:

$$ISD_i = \sqrt{\frac{\sum(R_{i,t} - R_i)^2}{T_i - 1}} \quad (2)$$

where $R_{i,t}$ represents the rank score of individual *i* at occasion *t*; and R_i represents the mean rank of individual *i* over the number of measurement occasions of that individual (T_i). The ISD_i is interpreted as the average mobility of individuals across the entire period (net of any systematic growth over time).

Another useful statistical indicator for quantifying intraindividual variability is the means square successive difference (MSSD), which measures the mean occasion-to-occasion mobility (Jahng, Wood & Trull, 2008). Following Jahng, Wood & Trull (2008), this statistic can be computed as follows:

$$MSSD_i = \frac{1}{T_i - 1} \sum_{t=1}^{T_i - 1} \left(R_{i,t+1} - R_{i,t}\right)^2 \quad (3)$$

where $R_{i,t}$ represents the rank score of individual *i* at occasion *t*, and $R_{i,t+1}$represent the next measurement occasion for the same individual. As Wang et al. (2012) explain, the MSSD measures both the amplitude of mobility (i.e., the ISD), as well as time dependencies, i.e., the extent to which achievement at one point is determined by previous achievement. This metric can

be further adapted in order to obtain more fined-grained information; for example, the amount of occasion-to-occasion mobility across groups in the upward or downward direction.[11]

2.5.3 Transition probabilities

One way of obtaining more fine-grained and easily interpretable estimates of directional mobilities is by using transition probabilities. These metrics measure the probability that a particular group of children will finish in certain position conditional on their original rank. Thus, these metrics are commonly regarded as measuring origin independence (or dependence), as they indicate the extent to which children's destination is related to their original position.

A common way of gauging positional movement is by constructing a "transition" or "mobility" matrix, which classifies individuals according to fixed and equal-sized categories (e.g., quartiles), with initial-period categories determining the row and final-period categories determining the column (Fields, 2006). As Fields (2006) explains, if the system is perfectly stable, then all the values will lie along the principal diagonal, and thus the mobility matrix would be an identity matrix. On the other hand, assuming a quartile partition, in a system with complete mobility 25% of the values in each initial quartile will be placed in each final quartile.[12]

[11] It is worth noting that the ISD and MSSD can have poor reliability and are sensitive to insufficient measurement occasions (Estabrook, Grimm & Bowles, 2012). Depending on several factors (e.g., the test reliability), the ISD can have an appropriate reliability (e.g., above .80) with less than 10 assessments or with more than 20 assessments (see Wang & Grimm, 2012 for an in-depth discussion on this topic).

[12] Transition matrices have the advantage that do not rely on functional form assumptions and are not susceptible to problems of scale. However, transition matrices can be based on arbitrary cutoffs that can distort to

Using the 1998-99 Early Childhood Longitudinal Study (ECLS-K), McDonough (2015) used transition matrices to estimate the staying probabilities and directional rank mobilities of academic achievement for Black and White students. In this study, we present one kind of transition probabilities not covered by McDonough (2015), and which have important normative implications. These statistics measure the probability of ending up in the opposite side of the distribution, and are thus called "rank reversal" probabilities (Jäntti & Jenkins, 2013). In particular, we consider the two "extreme" rank reversal probabilities: the chances of ending in the highest quartile after beginning in the lowest quartile (the upward rank reversal), and the chances of ending in the lowest quartile after beginning in the highest quartile (the downward rank reversal). The upward rank reversal (URR) probability can be represented as follows:

$$URR_i = \Pr\left(R_{i,8} > 75^{th}\ percentile \,\middle|\, R_{i,K} \leq 25^{th}\ percentile\right) \qquad (4)$$

and the downward rank reversal (DRR) probability as follows:

$$DRR_i = \Pr\left(R_{i,8} < 25^{th}\ percentile \mid R_{i,K} \geq 75^{th}\ percentile\right) \qquad (5)$$

some extent the mobility patterns. For illustration purposes, we used a quartile partition in our empirical application. However, one can use a more principled partition and solve the "ordinality problem" by (1) relating the test scores to a desired or undesired outcome –such as a college admission cutoff score or some adult outcome (e.g., Cunha & Heckman, 2008)–, or by (2) defining levels of achievement or proficiency using a standard-setting process (e.g., Hambleton & Pitoniak, 2006).

These extreme forms of positional mobility reveal complete origin independence, and can serve useful normative purposes. The URR statistic is a rough indicator of a meritocratic society, as it represents the probability of rising from the bottom to the top quartile (thus, Chetty et al., 2014 refer to this probability as the "American dream statistic"). On the other hand, the DRR indicates a society that allows severe setbacks, where students who begin among the highest achieving in the nation end up at the bottom of the distribution. It is worth noting that these "corner probabilities" are usually undetectable using average-based measures.

2.5.4 Measuring stability and change

Most of the mobility metrics presented so far consider positional change between the origin and the destination. In order to model this positional change as well as temporal dependencies simultaneously, one can estimate the percent of the variance that (1) remains stable, (2) changes systematically over time, or (3) is due to idiosyncratic circumstances, using the stable trait, autoregressive trait, and state model (START) presented by Kenny and Zautra (1995, 2001). This model captures both time-invariant and time-varying dimensions of positional mobility, by disentangling the longitudinal structure of achievement in terms of a completely stable (or "trait") factor, and a systematically-varying (or "state") factor, as well as an idiosyncratic (or error) component.[13] Figure A1 in the Appendix presents a path diagram of this model. As Newsom

[13] We use the "trait" and "state" terms, which are normally used in the context of these models (mostly applied in personality psychology). In the present study, however, these terms should not be interpreted as -exclusively- psychological attributes, as they reflect a wide range of environmental and social influences.

(2015) explains, the START model can be described using two equations. The first equation indicates that each observed measure of achievement is a function of three sources of variance:

$$Var(R_{it}) = Var(\eta) + \lambda_{it}^2(\eta_t) + Var(\varepsilon_{it}) \quad (7)$$

where $Var(\eta)$ represents completely stable variance; $Var(\eta_t)$, referred to as state variance, represents systematic variance that changes over time; and $Var(\varepsilon_{it})$, referred to as idiosyncratic variance, represents unaccounted variance in the observed outcome. The second equation indicates that each occasion-specific state factor is a function of the prior occasion and unaccounted variance:

$$Var(\eta_t) = \beta_{t,t-1}^2 Var(\eta_{t-1}) + Var(\zeta_t) \quad (8)$$

In this study, we focus primarily on the three parameters included in Equation 7. First, we consider the variance of the stable factor, which indicates the degree of stability in academic achievement at the population and subpopulation levels. Second, we consider the variance of the state factors, which capture statistically predictable change in the rank ordering of individuals. If there is a high degree of systematic academic mobility, then state factors should account for a substantial amount of variance. Finally, we obtain an error term, which represents the idiosyncratic (or non-systematic) variance that is unexplained by either the stable factor or the previous state.

2.5.5 Measuring group differences over time

The final mobility metric gauges the extent to which inequalities between groups increase or decrease over time. If educational processes serve as an equalizer –and all additional conditions remain the same–, then one can expect that the initial differences between groups will diminish over time; if, on the contrary, educational processes aggravate group disparities, then one can expect an increasing differentiation between groups. A simple way of estimating equalization (or disequalization) over time is, then, to compare the ratio of the between-group variance to the total amount of variance. A measure of this ratio can be obtained using the familiar coefficient of determination, or R^2, which is computed as

$$R^2 = \frac{\sum(\bar{R}_r - \bar{R})^2}{\sum(R_{ir} - \bar{R})^2} \qquad (9)$$

where $\bar{R}_r$ represent the mean achievement percentile rank of individuals in race *r*; $\bar{R}$ represents the average achievement in the population; and R_{ir} represent the achievement percentile rank of individual *i* in race *r*. This ratio can be interpreted as the amount of variance that is explained exclusively by racial differences. We examine the dynamics of group differentiation by comparing this cross-sectional metric over time. Similar to achievement gap measures, this metric is more sensitive to between-group (rather than within-person) differences. However, this metric measures the overall effect of group-belonging (e.g., race), rather than differences between two particular groups.

2.6 Estimates of academic mobility across racial groups in the US

2.6.1 Overall degrees of academic mobility

We began by estimating the positional mobility for the entire population using Equation 1. Following Chetty et al., (2014, 2018), we plotted this relationship with a binned scatter plot, in which we divided the horizontal axis into 100 equal-sized bins, and then plotted the mean rank in 8th grade vs. the mean rank in kindergarten in each bin. Figure 2 shows that this relationship can be well approximated by a linear function. The results indicate that, on average, a 10-percentile difference in children's' rank-order in kindergarten is associated with a 5.4 percentile difference in the rank-order in 8th grade. Even though this estimate can be compared to other outcomes (e.g., the estimated intergeneration family income rank-rank slope in the US is 0.341; see Chetty et al., 2014), it would be more meaningful to compare it to the positional academic mobility of the same educational system at previous time points, or to other educational systems.

Subsequently, we estimated rank-mobility parameters (α_i, β_i) for each racial group using the same specification. These estimates are displayed in Table A2 in the Appendix and illustrated in Figure 2. The slope parameter is smaller for Blacks (β_b = 0.43) and Hispanics (β_h = 0.45), compared to Asians (β_a = 0.51) and Whites (β_w = 0.51). This implies that the distance between these groups gets larger as initial achievement increases. For example, while the predicted difference between Whites and Hispanics who begin in the 25th percentile is 2.3 percentile points (p = 0.144), the predicted difference when children begin in the 75th percentile is 5.3 percentile points (p = 0.005). In general, however, the differences between the rank-rank slopes are small and not statistically significant (i.e., the lines in Figure 2B are approximately parallel).

At the same time, one can perceive in Figure 2B clear differences in the estimated intercepts across racial groups. In particular, one can perceive that Blacks have a lower intercept (α_b= 11.5) compared to Whites (α_w= 26.8), Asians (α_a= 28.4) and Hispanics (α_h= 26.1). This implies that, conditional on initial achievement, Black children are more likely to move downward, and these differences get slightly larger as initial achievement increases (given that Blacks have a smaller slope).

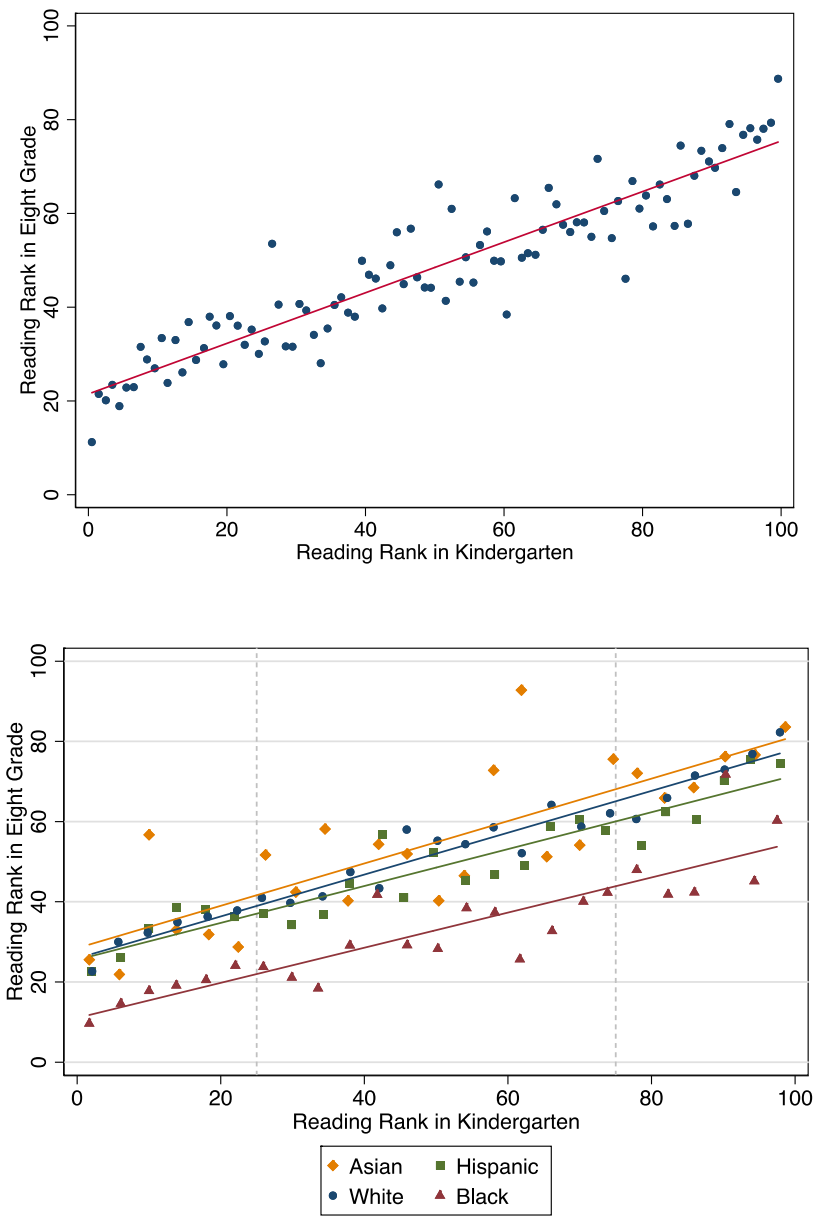


Figure 2. Scatter plots representing the relationship between students' achievement rank in eighth grade and students' achievement rank in kindergarten (Panel A) and the relationship between students' rank achievement in eighth grade and in kindergarten by race (Panel B).

***Note*. In Panel A, a 45-degree line would represent zero mobility, whereas a flat line would represent complete mobility. Panel A was constructed by dividing the horizontal axis into 100 equal-sized bins and calculating the mean percentile score in eighth grade in each of these bins. Panel B was constructed by dividing the horizontal axis into 25 equal-sized bins and calculating the mean percentile score in eighth grade for each race in each bin. The vertical lines are located in the 25th and 75th percentiles of students' reading achievement in kindergarten. An appropriate longitudinal weight was used in the calculations.**

2.6.2 Amplitude of academic mobility

We estimated the ISD and the MSSD at the population level and for each racial group using the specifications in Equations (2) and (3) respectively. As can be seen in Table A3, the ISD is similar across races, with a minimum of 13.7 percentile points for Hispanics and 14.9 percentile points for Whites. This means that, on average, White students move 14.9 percentile points around their own mean throughout the entire time-period. The differences in ISD between Whites and Hispanics (1.2 percentile points) is statistically significant ($t = 3.16$, $p < 0.01$). Overall, however, the differences in ISD across races is small, suggesting that there are no major differences in intraindividual mobility across races.

One can also perceive in Table A3 that the $MSSD_{1/2}$ (we present the square root so the estimates are comparable with the ISD) is higher than the ISD by around 2.5 points. This is expected, as the MSSD is more sensitive to idiosyncratic fluctuations and measurement error. We also computed the average positive and negative occasion-to-occasion differences across races. As Figure A2 in the Appendix shows, academic mobility is larger in the first year of kindergarten (with an average positive or negative academic mobility of around 30 percentile points), and appears to stabilize after third grade at about 5 percentile points. One can also perceive clear racial differences in the patterns of academic mobility in kindergarten: while Black students' downward movement in kindergarten is more pronounced, Hispanic students' upward movement in kindergarten is larger compared to other racial groups.

2.6.3 Transition matrices and rank reversal probabilities

In order to examine the extent to which individuals beginning in one quartile in Kindergarten remained in that same quartile in 8th grade, we examined a transition matrix for the whole population. Table A4 presents this transition matrix, showing the probability that children in each quartile in kindergarten stay or move to a different quartile in 8th grade. The largest proportion of students in each row fall in the leading diagonal (representing staying probabilities), or close to the diagonal. For example, more than 50% of the students beginning in the bottom quartile stay in that same group, and around 46% of the students beginning in the highest quartile remain at the top of the distribution. At the same time, one can see that a large number of students move to a different position. For example, around 32% of the students beginning in the 4th quartile end up in the 3rd quartile; 15% end in the 2nd quartile and 7% even end in the 1st quartile.

Given this transition matrix, it is not clear to what extent there is a large or small amount of academic mobility in the educational system, as one can observe some positional change but not complete origin independence. As with other mobility metrics, comparisons between groups might be more meaningful than overall measures. The transition probabilities of White and Black students displayed in Figure 3 indicate that while the latter are more likely to move downward the former are more likely to move upward. It can be useful to consider Figure 3 in relation to Figure A3 in the Appendix, which presents transition bar charts of systems with no mobility or complete mobility.

Differences in directional mobility are more consequential at the extremes – i.e., cases where the individuals who start in the lowest quartile move to the highest quartile or vice-versa. These measures account for initial school readiness gaps, as they compare individuals in the same original quartile. In order to perform these group comparisons, we computed the rank reversal

probabilities specified in Equations 4 and 5 by race. As Table 1 indicates (see also Figure 3 for the comparison between White and Black students), the upward rank reversal (*URR*) probabilities vary significantly by race. For example, one can see that while 11% of Asians who begin in the bottom quartile finish in the top quartile, less than 1% of Blacks achieve this. One can also see that Blacks and Hispanics have significantly different *URR* probabilities compared to Whites. This simple measure of success, which has been interpreted as the chances of achieving the "American Dream", shows substantial differences across racial groups.

Table 1 also displays the downward rank reversal (*DRR*) probabilities, indicating the chances that students who begin at the top of the distribution fall into the lowest quartile. Surprisingly, one can observe that 28% of Black students exhibit this drastic rank reversal, compared to 3% of Asians, 5% of Hispanics and 3% of Whites. The 25 percent-point difference in *DRR* between Whites and Blacks is statistically significant. However, the difference between Whites and Asians, and Whites and Hispanics, is not statistically significant. As Figure 3 shows, these findings are not limited to these extreme cases, as Black students show patterns of downward mobility (relative to White students) no matter where they begin.

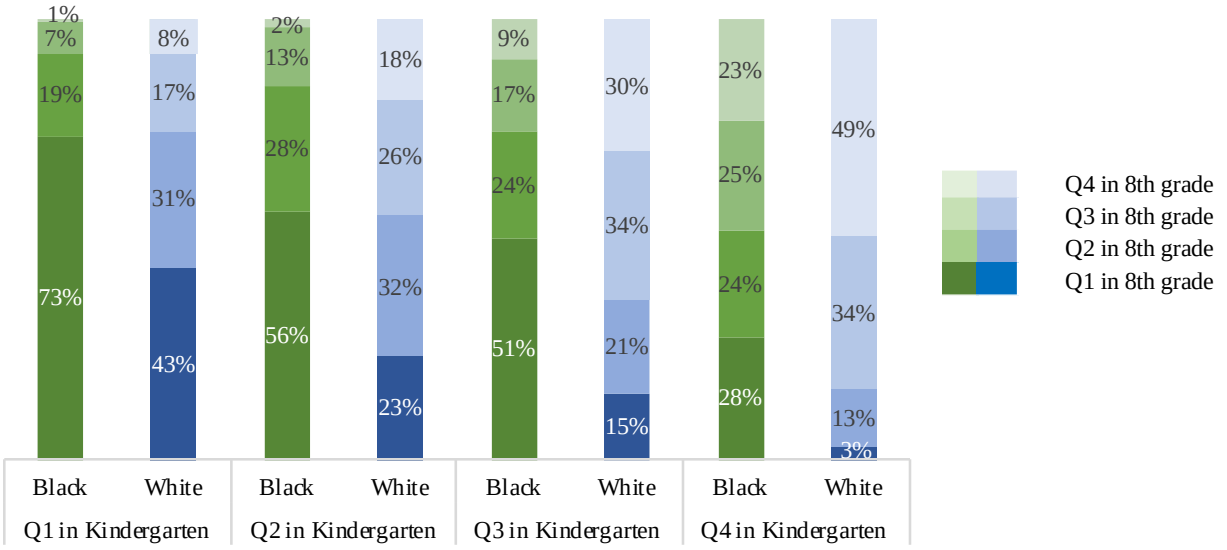


Figure 3. Transition bar chart comparing Whites' and Blacks' positional mobility.

***Note*. The first bar shows the likelihood that a Black student who begins in the 1st quartile in Kindergarten remains or moves to a different quartile in 8th grade. One can see that around 73% of students remain in the 1st quartile; 19% move to the 2nd quartile; 7% to the 3rd and only 1% to the 4th. These results suggest that there is very limited mobility for Black students who begin in the lowest achievement quartile. In contrast, the second bar shows that 43% of the white student who begin in the 1st quartile remain in that quartile; 31% move to the 2nd; 17% to the 3rd, and 8% to the 1st. Compared to Black students, White students are, then, more (upwardly) mobile at the bottom of the distribution (as just 74% remain in the bottom two quartiles, compared to 92% of Black students). These differences can be perceived by noting that the bar representing White students' mobility is closer to the second pattern in Figure 3A, while the bar associated to Black students is closer to the first pattern. Furthermore, the reverse pattern is true when White and Black students begin in the highest quartile. In that case, White students are likely to remain in the highest quartile, while Black students are likely to experience (downward) mobility. Indeed, for Black students beginning in the highest quartile, the bar looks almost exactly like the second pattern in Figure 3A, representing a system with complete mobility. In other words, being Black and starting in the highest quartile begets virtually no advantage over a system of complete mobility.**

Table 1. Rank reversal probabilities by racial group

	Estimates						Differences across groups		
	Overall	Asian	Black	Hispanic	Other	White	Asian – White	Black – White	Hispanic – White
P(Q4 in 8th grade | Q1 in Kindergarten)	0.052	0.108	0.006	0.044	0.031	0.082	0.026	-0.076***	-0.039*
	(0.007)	(0.062)	(0.003)	(0.012)	(0.019)	(0.013)	(0.064)	(0.013)	(0.018)
P(Q1 in 8th grade | Q4 in Kindergarten)	0.066	0.032	0.283	0.052	0.077	0.029	0.003	0.253**	0.023
	(0.014)	(0.023)	(0.084)	(0.025)	(0.040)	(0.007)	(0.024)	(0.084)	(0.023)

Note. The appropriate longitudinal weight was used in the estimation. Jackknife standard errors are in parentheses.

2.6.4 Stability and change

We first estimated a stable trait, autoregressive trait, and state (START) model using the entire sample. As shown in Table 2, the model had an excellent fit to the data, χ^2 = 179.11, *df* = 11, RMSEA = 0.044, CFI = 0.982, TLI = 0.975, SRMR = 0.026. The largest source of variance was related to the stable factor (49%), followed by the state variance (36%) and idiosyncratic variance (16%). This indicates that half of the variance is completely stable across time points, and 36% is related to systematic mobility over time. The estimated autoregressive coefficient was 0.82, suggesting that the changes across states is relatively gradual.

We then estimated a START model for each racial group independently. As Table 2 shows, all the models had an excellent fit to the data. One can also observe that Asian students and students belonging to other racial groups have very high academic stability, as the stable factor explains 58% and 60% of the variance respectively. These results are consistent with the rank-rank estimates. On the other hand, the stable factor only explains 38% of the variance in academic achievement for Black students, and 33% of the variance in academic achievement for Hispanic students. This means that for these students more than 60% of the variance is associated to some kind of mobility (either systematic or idiosyncratic mobility), compared, for example, to 42% for Asians and 52% for Whites.

Table 2. Estimated parameters and model fit statistics of the stable trait, autoregressive trait, and state (START) model by race

	Variance (%)			AR path coefficient		Fit statistics					
Group	Stable trait	State	Error	Estimate	SE	*df*	χ^2	RMSEA	CFI	TLI	SRMR
Overall	48.7	35.6	15.6	0.82	0.034	11	179.11	0.044	0.982	0.975	0.026
Asian	58.0	28.8	13.2	0.72	0.111	11	15.246	0.030	0.995	0.993	0.030
Black	38.0	40.7	21.3	0.79	0.163	11	21.470	0.035	0.986	0.980	0.031
Hispanic	32.8	42.8	24.4	0.85	0.077	11	32.616	0.039	0.990	0.986	0.027
Other	59.7	31.6	8.6	0.73	0.088	11	32.255	0.063	0.976	0.970	0.037
White	47.9	36.5	15.6	0.79	0.045	11	120.38	0.045	0.982	0.975	0.030

Note. The appropriate longitudinal weight was used in the estimation. Jackknife standard errors are displayed.

2.6.5 Group differences over time

In order to obtain the coefficient of determination at each time point, we regressed students' percentile rank score on race at each occasion. The omnibus *F* test indicated a significant difference in achievement across groups in every grade. As Figure 4 shows, differences among racial groups explain around 6% of the variance in achievement at the beginning of kindergarten. The coefficient of determination decreases in the spring of kindergarten (4.9%), suggesting equalizing effects in the first year of schooling. This is consistent with the high academic mobility in kindergarten depicted in Figure A2. However, as Figure 4 shows, the coefficient of determination progressively increases thereafter, and seems to stabilize at around 13.5%. The R^2 in eight grade (13.3%), is more than double the R^2 at the beginning of kindergarten. This indicates clear disequalizing effects, as the variance between groups tends to increase, while the variance within groups tend to decrease.

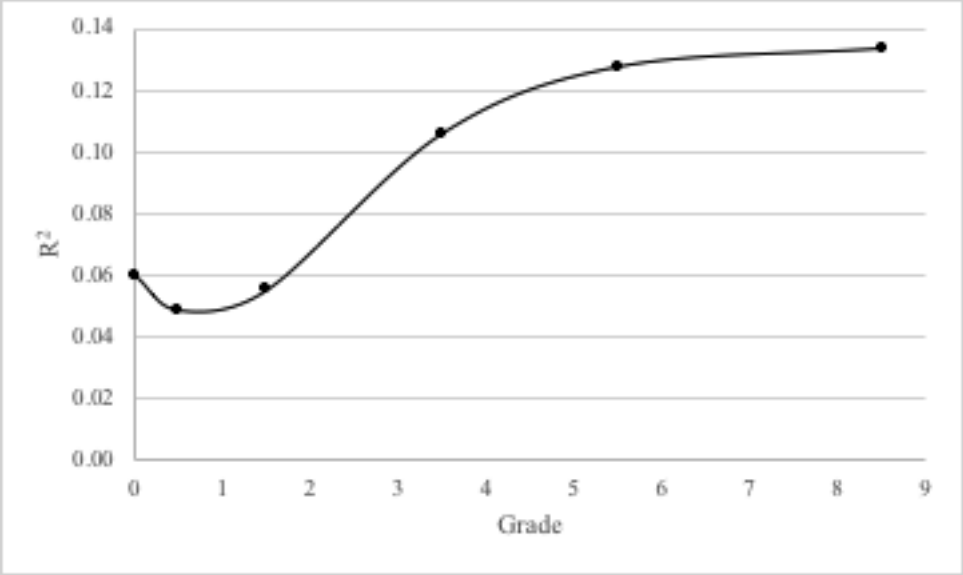


Figure 4. Coefficient of determination (R^2) of achievement predicted by race across grades.

2.7 Discussion

Descriptive analyses are a fundamental component of scientific research. By attending to particular features in the data, and summarizing these features in specific ways, descriptive analyses define the phenomena that subsequent research will try to explain. The knowledge we obtain through descriptive research is determinant, for example, for framing the research hypotheses underlying intervention and causal research (see Loeb et al., 2017). In the education field, achievement gaps have become stylized facts (i.e., simplified representations presumed to be generally true) that have shaped the way we think about educational inequality, as well as guided a wealth of descriptive and explanatory research. In this study, we reflected on the appropriateness of these comparative assessments for describing and evaluating educational inequalities. We argued that important normative dimensions are overlooked with traditional mean-based measures, and that academic mobility metrics can provide useful analytic tools to complement and go beyond achievement-gap analyses.

A naive positivist approach will assume that descriptive analysis merely consists in applying measures of central tendency and measures of variation to the data in order to describe the world in a neutral and comprehensive fashion. Value judgments and factual judgments are deeply entangled –at least in the social sciences (e.g., Anderson, 2004; Putnam, 2002; Schumpeter, 1949)–, and identifying relevant and accurate patterns in the data can be a complex endeavor. In other words, defining the most adequate comparisons for describing educational inequality is not a trivial task –both from a normative and methodological perspective. We began, then, by discussing key normative principles that should guide the adoption of particular comparative assessments. In particular, we argued that we should complement achievement-gap analyses by considering metrics that are sensitive to how individuals (rather than groups) change their relative

(rather than absolute) position in the overall distribution over time. We referred to the processes described by these metrics as academic mobility. Subsequently, we presented five metrics that are adequate for measuring different dimensions of academic mobility.

2.7.1 Complementing achievement gap analyses with academic mobility metrics

Several studies have investigated how achievement gaps change as students progress through school (e.g., Clotfelter, Ladd, & Vigdor, 2006; Fryer & Levitt, 2004; Reardon & Galindo, 2009).[14] An important shortcoming of this literature is that the use of different scale scores (e.g. standardized or unstandardized test scores) often yields different results, and most of the metrics utilized assume a measurement scale with interval properties (Reardon, 2008a). Despite these difficulties, most of the research agrees on some key points, e.g., that the Black-White achievement gaps grow during the school years, while the Hispanic-White and the Asian-White gaps decrease (Reardon et al., 2015). For instance, using the ECLS-K (which is the most commonly used dataset in this literature) Reardon et al. (2015) find that the Black-White achievement gap widens from -0.53 standard deviations in the fall of Kindergarten to -0.95 standard deviations in the Spring of 8th grade; while, in the same period, the Hispanic-White gap shrinks from -0.48 to -0.36 standard deviations, and the Asian-White gap also shrinks (albeit in the opposite direction) from 0.22 to

[14] As mentioned before, researchers can also make inferences regarding relative change -including achievement gap differences- using growth models (e.g., by combining initial status differences and differential growth rates). However, this approach can be more cumbersome, and typically assumes the interval scale of the test scores as well as parametric assumptions (in particular related to the functional form of the achievement trajectories).

0.17 standard deviations. In addition, the achievement gap literature also concludes that much of the growth of the Black-White gap occurs during elementary school (Reardon et al., 2015).

Even though these achievement gap analyses provide valuable information and have several advantages (e.g., they are easy to understand), there are several policy and socially-relevant aspects that are not captured by these measures and that can be obtained using the academic mobility metrics presented in this paper. First, these metrics allow us to obtain an estimate of the overall academic mobility in the population. Thus, in the empirical analysis we found that around 50% of the variance in the rank-ordering of individuals in reading achievement is completely stable from kindergarten to eighth grade. Even though a more meaningful interpretation of this result would require a historical or cross-national comparison of academic mobility in the entire education system, we can acknowledge that 50% represents a considerable degree of stability. This overall lack of academic mobility should increase societal concerns regarding initial disparities in achievement. A larger degree of mobility, on the other hand, should alleviate these concerns, as – in the long run– high mobility makes the disparities at a given point less consequential (as outcomes will tend to equalize over time; see Fields, 2010). In other words, the overall degree of academic mobility is a key factor that should be considered when interpreting achievement gaps at a particular time point.

Second, we obtained academic mobility estimates for each racial group. In the changes-in-gaps paradigm, one estimates to what extent the distance between two groups widens, narrows or remains stable over time. This statistic provides limited information for understanding the processes behind the development of educational inequalities. For example, the fact that the gap between White and Black students widens from one-half to a full standard deviation might be due to (1) upward mobility in some White students; (2) downward mobility in some Black students; or

(3) some combination of (1) and (2). Without a clear understanding of these possibilities, we will remain with a nebulous grasp of the processes behind this particular phenomenon.

In the empirical analysis, we found significant differences in the amount of academic mobility across racial groups. For example, while Asian students have very high academic stability (as 58% of the variance in the rank-ordering of individuals in reading achievement remains stable across time points), Black and Hispanic students have relatively low academic stability (as less than 40% of the variance in the rank-ordering remains stable). This means that more than 60% of the variance in Black and Hispanic students' reading achievement is associated with some kind of mobility. It is worth noting that, according to these results, Hispanic and Black students have a similar degree of academic mobility, which is something that we would not be able to infer from achievement gaps alone. In addition, these results suggest that the changes in gaps of Hispanics and Blacks with respect to other groups is due more to the mobility of the former than the mobility of the latter.

Third, we presented mobility metrics that allow us to estimate the direction of academic mobility at different segments of the achievement distribution. Previous studies have investigated this issue by examining differential growth rates, which assume that test scores are interval-scaled (Reardon, 2008b). In order to address this problem, we implemented (1) linear rank-rank measures, which provide a parsimonious summary of the degree and direction of academic mobility by relying on some parametric assumptions; and (2) transition probabilities, which are completely nonparametric (McDonough, 2015). The results of these two metrics suggest that, conditional on initial achievement, Black students are more likely to move in the downward direction at virtually every segment of the achievement distribution.

These findings have important consequences for the way we think about the development of educational inequalities. Notably, they contradict the "Mathew effect" hypothesis, according to which the development of cognitive abilities (e.g., reading) can be characterized as a cumulative process in which the advantages or disadvantages of individuals accrue over time (see, e.g., Stanovich, 1986; Pfost et al., 2014). The findings described above suggest that initial status is not a determining factor for subsequent academic performance, and that there are institutional and societal forces that produce systematic inequalities, overriding the potential benefits or impediments related to initial achievement. A result that clearly supports this point – and that might be considered a case of manifest disadvantage–, is that around 28% of the Black students who begin in the highest quartile end up in the lowest quartile (compared, for instance, to 3% of Whites and 5% of Hispanics).

Fourth, we implemented mobility metrics that are only sensitive to within-person variation. Overall, the results are consistent with the achievement gap literature (e.g., Reardon & Galindo, 2009; Fryer & Levitt, 2004), according to which Hispanic students tend to be more upwardly mobile in the first years of schooling, and Black students tend to be less upwardly mobile throughout the entire period (see Figure A2). In contrast to achievement gaps, however, these metrics provide novel insights into individual-level (rather than group-level) processes of change. For example, Figure A2 indicates that all students tend to be more mobile in kindergarten (with an average mobility of around 30 percentile points), and tend to stabilize after third grade at about 5 percentile points.

Finally, we used the coefficient of determination to measure the extent to which the inequalities between groups increase or decrease over time. Traditional methods do not provide comparable metrics of group differentiation and are better suited for comparing only two groups

(e.g., White versus Black students). In the empirical analysis, we found evidence of racial disequalization over time, as the differences between racial groups increase, while difference within groups decrease.

2.7.2 Limitations and conclusion

It is worth mentioning that the purpose of the present paper was not to conduct a thorough analysis of academic mobility in the United States. The goal, rather, was to argue why the concept of academic mobility provides a useful framework for describing and evaluating educational inequalities, and present various metrics that measure different aspects of academic mobility. For that reason, we presented the academic mobility metrics in their most basic form –both in their mathematical representation and empirical application. The full potential and significance of these measures will become more apparent as subsequent studies provide more detailed descriptive and explanatory evidence on students' academic mobility, as well as more comprehensive analysis of each measure's psychometric properties. Questions that can be answered with these metrics, and could not be answered with other traditional measures, include: How has academic mobility at the population and subpopulation levels changed over time? How does academic mobility in particular educational systems (e.g., districts or countries) compare to academic mobility in other educational systems? How do mobility patterns vary across the achievement distribution? To what extent do resources and experiences within school, family and neighborhood contexts independently and cumulatively explain differences in academic mobility? It is our hope that the metrics that we have presented, which have been used in disciplines outside of education, can be applied in the field of education in order to better describe and explain educational inequalities.

3.0 Introduction to Studies 2 and 3

While the goal of the first study is entirely descriptive (i.e., it intends to characterize the observed patterns and relationships in the data in an informative and parsimonious manner), the second and third studies are explanatory (i.e., they intend to shed some light into why individual differences in skill development might arise). As explained in the introduction, it is important to distinguish between these purposes, as they correspond to substantially different claims that have different standards of success, uses and supporting assumptions.

In the introduction I also noted that, even if causal inferences are seldom justified using observational data, most of the questions that are asked in psychology and education are causal in nature. This can be easily recognized by the fact that researchers in these fields (1) typically interpret the parameter estimate associated to a particular predictor; (2) justify their data, design and modelling decisions based on causal considerations (e.g., by using methods that condition on potential confounders); and (3) intend to describe the causal mechanisms to inform theory or predict the effect of interventions. In other words, even if researchers avoid the use of causal language, they are often interested in causal questions and implicitly make causal inferences (see, e.g., Grosz et al., 2020).

Given that the aims of a large portion of empirical research is causal in nature, the avoidance of causal language can generate confusion regarding the (warranted) interpretations, standards of success, limitations, and supporting assumptions of empirical research. In addition, the ambivalence regarding the purposes of research can hinder the quality of the research itself. For example, Foster (2010, p.1760) explains that "currently, developmentalists conduct complex analyses that are not useful in pursuing either aim: The analyses are too complex to produce good

description, and the complexity is not employed in a manner that facilitates causal inference." Explicitly recognizing the goals of explanatory research can help, then, improve the quality of the research, as well as articulate in a more transparent manner the limitations and assumptions of the analyses.

Based on these considerations, and given that the aims of the second and third studies are explanatory in nature, I explicitly adopt causal language to describe the purposes of these studies; explain how the methods presented can be useful from a causal inference perspective; and describe the causal assumptions and limitations of these methods. Thus, in these studies I explicitly adopt causal language, aside from the results section, where I limit the use of this terminology (except in hypothetical form, e.g., "if the results were interpreted causally, they would imply that…"). The avoidance of causal terminology in the results section is due to the fact that some of the assumptions of the methods employed (e.g., the absence of hidden confounders) are too strong, and as a consequence the results should be interpreted with caution. Yet it is also worth noting that, to a larger or lesser extent, this caution applies to all statistical modelling, as there is no technique that can conclusively prove causation, but rather that the causal inferences made are consistent with the data (Bollen, 1989).

In addition, it is worth mentioning that there are two major traditions in causal inference, the potential outcome and graphical approaches. As several authors have noted (e.g., Imbens, 2019; Morgan & Winship, 2015; Pearl, 2009; Richardson & Robins, 2013), these frameworks are complementary, as they have different strengths that make them appropriate for answering different kinds of questions. For example, while the potential outcome framework might be better suited for defining treatment effects as well as estimating treatment heterogeneity, the graphical

approach might be better suited for conveying and assessing key identification assumptions (e.g., Imbens, 2019).

More generally, one can argue that while the potential outcome framework might be more useful for estimating causal effects when the treatment is well defined and the underlying causal structure is known (e.g., in an RCT), graphical models are useful for representing causal relationships when the underling causal structure is not well understood, or when the structure represented involves complex relationships among variables. Following this logic, one can conclude that graphical models are useful at earlier stages of explanatory research (especially with observational data), when researchers want to shed some light into the underlying structure of a set of variables.

In the second study, I explain how causal search algorithms can be used to learn the causal structure or tests particular causal hypotheses using observational data. I use causal terminology in this study (e.g., "causal structure" or "direct cause"), given that, as I will explain, (1) the aim of the research is to shed some light into the causal structure of academic achievement; (2) the methods are based on causal principles (e.g., the Causal Markov Condition) and standards of success (e.g., the identification of proximal mechanisms); and (3) this terminology is consistent with the existing literature (e.g., Peters, Janzing & Schölkopf, 2017; Spirtes et al., 2000; Morgan & Winship, 2015). I will also explain the assumptions and limitations of these methods from a causal perspective (e.g., the causal sufficiency assumption), and discuss interpretations that relax these assumptions (e.g., using proximal mechanisms for predictive purposes).

In the third study, I discuss ways of capitalizing on the longitudinal structure of the data to make causal inferences. Similar to the second study, I use causal terminology in the third study given that (1) the goal of the analysis is to estimate the causal effect of executive functions on

academic achievement; (2) the methods are based on causal principles (e.g., the ability to control for stable confounders) and standards of success (e.g., obtaining unbiased estimates); and (3) this is the terminology used to describe these methods (e.g., Allison, 2009; Angrist & Pischke, 2010; Halaby, 2004).

4.0 Study 2. The structure of academic achievement: Searching for proximal mechanisms using causal discovery algorithms[15]

Causal search algorithms have been effectively applied in different fields, including biology, genetics, climate science, medicine and neuroscience. However, there have been scant applications of these methods in social and behavioral sciences. This paper provides an illustrative example of how causal search algorithms can shed light on important social and behavioral problems by using these algorithms to find the proximal mechanisms of academic achievement. Using a nationally representative dataset with a wide range of relevant contextual and psychological factors, I implement four causal search procedures that varied important dimensions in the algorithms. Consistent with previous research, the algorithms identified prior achievement, executive functions (in particular working memory, cognitive flexibility and attentional focusing) as well as motivation as direct causes of academic achievement. I discuss the advantages and limitations of causal graphs in general and causal search algorithms in particular for understanding social and behavioral problems.

[15] Manuscript published in Sociological Methods & Research; see Quintana (in press). The final version is available at https://journals.sagepub.com/home/smr).

4.1 Introduction

Social and behavioral scientists are generally interested in examining how the environment affects human behavior. Yet the complexity of both the human organism and the environments to which humans are exposed to makes this a very difficult task. The problems of identifying the causal effect of particular environmental factors on some aspect of human behavior are widely discussed in the literature (e.g., Morgan and Winship, 2015). Typically, a major concern is that the estimated associations are spurious due to unobserved confounders. Other difficulties include measurement error, selection bias and feedback loops. Faced with these complications, many researchers renounce to the possibility of drawing causal inferences from observational data and limit themselves to examining associations. Even if correlational evidence can be used for different purposes, the investigation of causal relationships in many social and behavioral domains is too important to abandon. Causal claims are at the basis of scientific understanding, and provide the support for concrete and possibly far-reaching program and policy interventions.

In the social and behavioral sciences, causal inference is generally associated with the use of experimental or "quasi-experimental" designs (e.g., Angrist and Pischke, 2008; Cook, Campbell and Shadish, 2002). Influenced by advances in econometrics, the emphasis on research design and threats to internal validity has brought a much needed "credibility revolution" in the social sciences (Angrist and Pischke, 2010). At the same time, researchers have warned about a "black-box" approach to causal inference, which focuses on establishing causal connections between two variables, without explaining how or why this causal relationship arises (e.g., Deaton and Cartwright, 2018; Hedström and Swedberg, 1998; Knight and Winship, 2013). Critics of the "black-box" approach have argued that estimating causal effects is not sufficient, and that we should focus on identifying and describing the mechanisms that link cause and effect. In fact, some

researchers (e.g., Heckman, 2005; Deaton and Cartwright, 2018) have argued that only a mechanism-based analysis can provide the appropriate depth required for scientific explanation and policy intervention.

Even though researchers have long recognized the importance of mechanism-centered explanations, there is no clarity regarding the appropriate ways of identifying and describing mechanisms. Social and behavioral researchers usually study mechanisms using mediation analysis, which often relies on strong and untestable assumptions (Imai, Keele, Tingley and Yamamoto, 2011; VanderWeele, 2015). In addition, there is no conceptual clarity regarding what counts or does not count as a "mechanistic explanation." In this paper, I suggest that causal graphs can be useful tools for organizing and investigating in a systematic fashion mechanistic findings and hypotheses. In addition, I explain how causal structure learning methods can help us identify the causal structure behind organism-environment interactions. I consider this approach in the context of the widely-discussed issue of the environmental determinants of academic achievement.

The main purpose of this paper is to provide an illustrative example of how causal search algorithms can shed light on important social and behavioral problems. Even though I will describe some of the basic assumptions and concepts behind these methods, providing a comprehensive and detailed review of causal graphs or causal search algorithms is beyond the scope of this study[16]. The remainder of the paper is organized as follows. First, I frame the challenges of causal inference in the context of educational research. Second, I explain how causal graphs provide useful

[16] For reviews on causal graphs, see Elwert (2013), Glymour and Greenland (2008), Morgan and Winship (2015), or Pearl (2009); for reviews on causal search algorithms, see Eberhardt (2017), Glymour, Zhang and Spirtes (2019), Kalisch and Bühlmann (2014), or Spirtes and Zhang (2016); for a practical guide on the implementation of these algorithms see Malinsky and Danks (2017).

conceptual and methodological tools for investigating causal structures in general and proximal mechanisms in particular. Third, I review previous research on the proximal mechanisms of academic achievement. Fourth, I describe the basic principles and assumptions of the search algorithms implemented in this study. Finally, I present the data, methods and results of the empirical application using causal search algorithms.

4.2 The challenges of causal inference: The case of academic achievement

Researchers from different disciplines have long studied the factors affecting student academic achievement. The factors investigated range from individual-level constructs such as motivation and non-cognitive skills, to a wide-range of contextual influences including peer, teacher, classroom, home, school, neighborhood, and policy-level effects. It is unquestionable that many of these studies provide valuable insights and findings. At the same time, many of these studies do not explain why particular causal relationships arise, and how the factor under consideration relates to the other environmental and psychological causes. By focusing only on the *p*-values or effect sizes attached to particular factors, we end up with a list of disconnected and structureless influences of student achievement. An example of how these results can be used and interpreted is Hattie's (2008) attempt to synthesize the literature by providing a ranking of 138 factors affecting achievement, based on effect-size calculations using more than 800 meta-analyses, which include around 52,000 individual studies. A simple and unambiguous rank of effect sizes (e.g., in Hattie's analysis "micro teaching" ranks 4, while "home environment" ranks 31) might be appealing to many. It is evident, however, that such oversimplification of complex phenomena poses severe limitations for scientific understanding, and that we need to identify and

describe the underlying mechanisms in order to achieve a deeper comprehension of the actual data-generating processess.[17]

Broadly speaking, mechanisms refer to entities or activities that generate changes in other entities or activities (Machamer, Darden and Craver, 2000). Mechanism-based explanations should be conceived, then, as particular kinds of causal explanations, and should conform to the standard methods for investigating causal claims. For instance, a simple association or regularity between two variables is an insufficient condition for identifying a mechanism. Given the counterfactual framework of causation (Holland, 1986; Rubin, 1974; Woodward, 2002), an adequate mechanism-based explanation should support specific counterfactual claims; in particular, it would specify how a particular outcome would change if the mechanism is intervened or manipulated in some specific fashion.

What differentiates mechanism-based explanations from other causal (and in particular "black-box") explanations is the explicit reference to the components of the process. Mechanisms refer to the "cogs and wheels" of specific processes, and by explicitly modelling the "intervening variables" or "steps", mechanism-based explanations elucidate how or why particular outcomes follow from a set of initial conditions. In other words, while "black-box" explanations are primarily concerned with describing the link between inputs and outputs, mechanism-based explanations are mainly concerned with describing the structure of the process (Hedström and

[17] Hattie's work has been criticized due to the inclusion of low-quality studies in his analysis (Slavin 2018). The point I am making, however, is not related to the quality of this particular investigation. The reference to this study is intended to illustrate that (1) list-like rankings are the only way of synthesizing structureless findings (e.g., effect sizes), and (2) such syntheses can be insufficient for deep scientific understanding and warrantable extrapolations (see below).

Swedberg, 1998). It is worth noting, however, that what counts as a mechanism can vary depending on the discipline and level of analysis adopted; for instance, what can be considered a mechanistic explanation in economics might be considered a "black-box" explanation in psychology; and a mechanistic explanation in psychology might be considered a "black-box" explanation in biology or neuroscience. Regardless of the granularity of the variables considered, one can generally say that, in contrast to "black-box" theories, mechanism-based explanations make explicit reference to the components of some particular process.

Many researchers have argued that mechanism-based explanations are necessary for adequate scientific understanding (e.g., Elster, 1998; Rutter, 2007). Understanding the causal structure of a phenomenon is what allows researchers to construct theories, investigate competing explanations of particular phenomena, and use scientific theory for interpreting new evidence (Heckman, 2005). In addition, some understanding of the causal structure is required for the extrapolation or forecast of particular interventions to new cases (i.e., for "external validity"). The generalization of a particular causal relationship to a different context assumes a wide range of background conditions (referred to in the literature as "supportive factors", "interactive variables" or "moderators"; see Deaton and Cartwright, 2018). Given that "black-box" causal theories are devoid of structure and that, as a consequence, no supportive factors are explicitly considered, these theories typically do not generate warrantable extrapolations.

Apart from supporting scientific understanding and warrantable generalizations, knowledge of causal structures is required for the identification of causal effects using observational data (Pearl, 2009). In particular, knowledge of causal structures is required for (1) determining the identifiability of causal effects (Pearl, 2009), and (2) identifying these effects using conditioning strategies (Hernan et al., 2002; VanderWeele and Shpitser, 2011). Statistical

adjustment can be an effective tool to eliminate confounding bias, but adjusting for the wrong set of covariates may actually increase bias (Pearl, 2011). Consequently, understanding the causal structure among the measured variables can be beneficial to assess the credibility of causal assumptions as well as for guiding confounder selection.

Mechanism-based explanations can be characterized, then, by two fundamental properties: (1) they describe causal processes, and (2) they make explicit reference to the components (or structure) of those processes. These explanations can be useful for a variety of purposes, including synthesizing, interpreting and using research findings, as well as assessing causal modelling assumptions. Now, even though we have relatively well-established frameworks for defining causal claims (e.g., the counterfactual or potential outcome theories), and for identifying causal effects (e.g., through experimental or "quasi-experimental" designs), there is less clarity regarding adequate ways of defining and identifying causal structures. In other words, even if researchers can have confidence in specific causal relationships, many researchers lack practical frameworks for organizing these relationships in a principled and systematic fashion. Absent clear principles for thinking about causal structures, researchers might end up focusing exclusively on effect sizes, which might lead them to compare incommensurable or vague constructs (as in Hattie's ranking described above), or present effects that might not replicate or generalize due to unspecified contextual differences.

4.3 Causal structures and proximal mechanisms

Causal graphs can provide a schematic representation of causal structures, and can be used –among other things– to explicitly convey causal assumptions and assess the identifiability of a

causal effect given the observed data (Morgan and Winship, 2015; Pearl, 2009; Spirtes et al., 2000). In essence, causal graphs are composed of a set of variables (*vertices*, *nodes*), and a set of arrows (*edges*) that connect the variables. In a causally interpreted graph, a single-headed arrow (called a *direct edge*), represents a causal link between two variables in the direction of the arrow; a missing arrow, on the other hand, indicates the absence of a causal effect, and is typically considered a stronger assumption. A sequence of edges between several variables is called a *path*, and a *directed path* is a path where all arrowheads point in the same direction. Direct causes are called *parents*, and both direct and indirect causes are referred to as *ancestors*. Similarly, variables that are directly affected by another variable are called their *children*, and variables that are either directly or indirectly affected by another variable are referred to as their *descendants*. Most causal graphs considered in the literature are composed of directed paths with no feedback loops, and are therefore called *directed acyclic graphs* (*DAGs*). Contrary to path models in SEM, DAGs are completely nonparametric, allowing for any functional form between the variables.

Causal graphs represent causal relationships between variables. The causal effect of some variable X on another variable Y is defined as the counterfactual changes in the distribution of Y after some intervention on X (Pearl, 2009). Pearl (2009) formalizes this definition as $\Pr\big(Y \big|\, do(X = x)\big)$, which can be interpreted as the distribution of Y that would be generated when the variable X is forced to take on some particular value x. In Pearl's terminology, the condition $do(X = x)$ refers to some "atomic" or "surgical" intervention equivalent to an ideal experiment, i.e., where only X is manipulated. This account of causation assumes, then, that the mechanisms generating the variables are independent, that is, they can be manipulated without altering the other mechanisms in the causal structure. This condition is often referred to as modularity, autonomy or locality (Pearl, 2009), and should be a primary consideration for variable selection. In particular,

this implies that one should exclude variables that are logically or conceptually related, and try to include variables that have well-defined manipulable properties (Spirtes and Scheines, 2004; Woodward, 2016).

4.3.1 Defining proximal mechanisms

Causal inference in the social and behavioral sciences faces an important challenge, namely that human behavior can be affected by many factors; so many that researchers might desist from investigating causal structures by simply stating that "everything affects everything else." A common way of overcoming this difficulty is by noting that environmental or contextual effects are hierarchically structured. This hierarchy can be used in turn to justify differences in explanatory relevance. In particular, many researchers have defended the explanatory priority of "proximal mechanisms" (or individual-level processes) over "distal mechanisms" (or social-level phenomena; e.g., Bronfenbrenner and Morris, 2006; Elster, 1998; Hedström and Ylikoski, 2010). The main idea behind this approach is that, if a contextual factor affects human behavior, then it must necessarily be mediated by proximal individual-level processes (Diez Roux 2004). In this sense, proximal processes can be considered "the primary engines of development" (Bronfenbrenner and Morris, 2006).

Even though there is a long tradition in the social and behavioral sciences supporting the explanatory priority of proximal mechanisms, what counts as "proximal" or "distal", or how to define the hierarchical or nested structure of contextual effects, is often intuitively determined and rarely empirically established. Causal graphs provide clear principles for defining and investigating proximal mechanisms, as well as the hierarchical properties of causal structures. I briefly review some of these principles below.

A fundamental concept in causal graphs is the Causal Markov Condition (CMC), which states that every variable in a DAG is independent of its non-effects conditional on its direct causes (Spirtes et al., 2000). This condition is a central component of the d-separation criterion, which is a general principle connecting the causal relations in a graphical model with probability distributions (see Pearl, 2009). In particular, d-separation can be used to infer all dependence and independence relationships implied by a DAG.

The CMC conforms to the intuition that information flows along causal paths, and that direct (or proximal) causes mediate the effect of indirect (or distal) causes. Formally, the CMC implies that, given a set of variables $\boldsymbol{V} = \{X_1, \dots, X_n\}$ and a DAG G representing the causal structure of $\boldsymbol{V}$, then the joint probability distribution can be factored as follows:

$$P(x_1, \dots, x_n) = \prod_{j=1}^{n} P(x_j \mid pa_j) \quad (1)$$

where pa_j represents the parents of x_j in G. The fact that joint probability distributions can be factored into smaller distributions involving a subset of variables PA_j is considered one of the main advantages of graphical models in general and causal graphs in particular (Pearl 2009).

Another important implication of the CMC is related to conditional independence, and the idea that direct causes "screen off" more distal causes; that is, if we know PA_j then any other ancestor that is not in PA_j will not give us any additional information regarding X_j. Formally, X_j is independent of its non-descendants (i.e., non-effects) given PA_j:

$$X_j \perp\!\!\!\perp ND_j | PA_j \quad (2)$$

where ND_J are the non-descendants of X_j. As this equation indicates, the CMC connects causal terms (e.g., "direct causes" and "non-effects) with the informational relevance of the observed variables (Janzing and Scholkopf, 2010). In particular, the CMC implies that, ignoring a variable's effects, all the relevant probabilistic information about a variable is contained in its direct causes (Scheines, 1997). As explained below, this "bridge principle" between causation and probability allows inferences from the observed (conditional) independencies to the underlying causal structure.

It is worth noting that the CMC is expected to hold only under certain assumptions. In particular, the CMC assumes a causally sufficient set of variables, that is, that there are no unmeasured direct common causes of any pair of variables under consideration (i.e., no unmeasured confounders). This assumption is unrealistic in most cases, as we seldom measure all common causes of all pairs of variables. As in most observational studies, causal inferences using the CMC in the presence of unmeasured confounders only provides provisional results under the assumption of "what if these causal assumptions were true" (Bollen and Pearl, 2012). Thus, as I explain in the discussion, the DAG that I estimate by applying causal search algorithms (which are based on the CMC) to an observational dataset should not be interpreted as the true causal structure. However, even if dropping the causal sufficiency assumption limits our inferences considerably, the estimated DAG can provide valuable insights. In particular, the probabilistic independencies among the measured variables can be used to identify proximal and distal relationships; reveal which observed variables can be potential confounders, mediators or moderators; and indicate which causal relationships do not exist (see Scheines, 1997).

Another important assumption underlying the CMC is modularity. As explained above, this assumption implies that each variable and its causal parents represents an autonomous

mechanism, i.e., a mechanism that can be manipulated without changing other mechanisms (Pearl, 2009). This assumption can be violated if the variables considered are not sufficiently distinct or if they are too coarsely grained (Hitchcock, 2018).

Causal graphs provide, then, a clear and testable definition of proximal mechanism. To illustrate this, consider the contextual effects of academic achievement. For example, consider the finding that parental divorce can have a negative effect on academic performance (e.g., Amato and Anthony 2014). Evidently, if this event has any effect on the child's academic achievement, then it must be mediated by some proximal mechanisms, e.g., by changes in the child's motivation or the ability to focus. If, for instance, one assumes that the latter are the only mediating paths, then knowing the changes in these proximal mechanisms will be sufficient for predicting the change in academic achievement (i.e., in this scenario knowing whether the parents divorced or not would not provide any additional information). In other words, parental divorce would be independent of achievement conditional on the child's motivation and ability to focus. Given that the latter statement can be empirically tested using a test of conditional independence, we can see how the CMC provides a clear framework for investigating causal assumptions regarding proximal mechanisms.

4.4 Searching for the proximal mechanisms of academic achievement

Causal search algorithms can be effectively used to search for causal structures using non-experimental data (Spirtes and Zhang, 2016). These algorithms are based on certain assumptions (e.g., the Causal Markov Condition and Faithfulness), which connect causal graphs with probability distributions (I expand on this point below). Apart from these fundamental

assumptions, the success of these algorithms is evidently constrained by the nature of the data set. Thus, one can avoid important complications by using a data set with the following characteristics: (1) it has an appropriate sample size to ensure statistical power; (2) it includes a representative sample, so one minimizes the risk of selection bias (e.g., Elwert and Winship, 2014; Spirtes, Meek and Richardson, 1995); and (3) to the greatest extent possible, it includes the most important hypothesized or proven causal factors associated to the target variable or system under investigation. This latter point is critical and most likely the most contentious; omitting critical factors might naturally lead to mistaken representations. I will begin then by justifying the choice of the data set used in the empirical analysis by briefly reviewing three constructs that, based on the literature on skill development, are generally recognized as important proximal mechanisms (or "mediators") of academic achievement, and therefore should be included in the data set: previous academic achievement, general cognitive skills (in particular executive functions) and motivation. Apart from justifying the use of the data set, this discussion also intends to put forward a body of research that can be used either as an external criterion to judge the success of the search procedures, or a series of hypotheses that can be empirically examined using a causal search framework.

4.4.1 Previous achievement

The causal effect of previous academic achievement on subsequent achievement is supported by a wide-range of theories and correlational evidence (e.g., Bailey et al., 2018; Cunha and Heckman, 2007; Stanovich, 1986). The idea that earlier states affect subsequent development has been described in different terms within different theoretical frameworks, including "cognitive efficiency" (Stanovich, 1986), "Mathew effects" (Pfost, Hattie, Dörfler and Artelt, 2014), "state

effects" (Bailey et al., 2014), and "self-productivity" (Cunha and Heckman, 2007). In essence, what this theoretical and empirical body of research suggests is that skill formation is essentially dynamic, and that skills at one point are to some extent determined by the skills developed earlier.

4.4.2 Executive functions

Executive functions refer to a set of cognitive processes involved in the control and coordination of information in the service of goal-directed behavior (Jacob, 2015). These top-down processes are considered to be necessary for higher-order cognitive abilities such as reasoning, problem solving and planning (Collins and Koechlin, 2012). Executive functions are typically classified in three main components: (1) inhibitory control, which refers to the ability to suppress one's internal or external predispositions or distractions; (2) working memory, which refers to the capacity to hold information and mentally manipulate it; and (3) cognitive flexibility, which refers to the ability of actively changing or adjusting mental processes, overcoming predisposition and inertial tendencies (Diamond, 2013). A wealth of research has been devoted to study the relationship between executive functions and academic achievement (e.g., see the meta-analyses by Follmer et al., 2018; Jacob and Parkinson, 2015, and Smithers et al., 2018). Even if the research presents mixed results and no conclusive causal associations can be established, both theoretical and empirical considerations suggest that executive functions can be an important determinant of academic achievement (Diamond, 2013).

4.4.3 Motivation

Different motivational constructs have been defined within different theoretical frameworks, e.g., attribution theory, interest theory, expectancy-value theory or self-efficacy theory (Lazowski and Hulleman, 2016; Hulleman et al., 2010). The overall purpose of these conceptual and operational definitions is to understand the connections between affective and cognitive processes in goal-directed systems (Forgas, 2008; Kruglanski et al., 2002). According to the existing experimental research, motivational interventions –broadly defined– have moderate to large effects on academic achievement (Lazowski and Hulleman, 2016). The relevance of motivational constructs is then supported by empirical evidence within different theoretical frameworks (e.g., Huang, 2011; Marsh, 2005).

The idea that previous achievement, executive functions and motivation can be considered the main proximal "mediators" of academic performance has been suggested in the literature (Watts et al., 2015). This hypothesis is strengthened by neuroscientific evidence, which associates distinct neural circuits to these constructs: executive functions are associated to the functioning of the prefrontal cortex (Koechlin and Summerfield, 2007); motivational constructs are associated to structures involving the amygdala (Phelps, 2006); and linguistic and mathematical competencies are supported by various temporal, parietal and frontal regions (Dehaene et al., 1999; McCandliss, Cohen and Dehaene, 2003). These findings are consistent with the modularity assumption and the proximal nature of these constructs. However, these findings (along with the evidence supporting the existence of causal associations between the three constructs and academic achievement) do not directly support their status as proximal mechanisms. As explained above, identifying proximal mechanisms requires an explicit reference to the causal chain connecting contextual causes with academic achievement. More specifically, we would need to show that, conditional on the

proximal mechanisms, academic achievement is independent of all the other predecessors. I test this hypothesis in the empirical section of this paper.

4.5 Using causal discovery algorithms to search for the proximal mechanisms of academic achievement

In the empirical section of this study, I apply causal discovery algorithms to search for the proximal mechanisms as well as the entire causal structure of academic achievement. I use a nationally representative data set that includes measures of the three hypothesized proximal mechanisms, as well as a range of psychological and contextual factors that –based on previous research– are related to academic achievement. The research questions of the empirical analyses are the following:

1. What are the direct causes (or parents) of academic achievement?
2. What is the causal structure of academic achievement?

The first question is the primary focus of this paper. Identifying the direct causes or proximal mechanisms provides a principled way of investigating and explaining highly complex systems. In addition, identifying the direct causes will help guide effective interventions. A direct cause represents a stable and autonomous mechanism (Pearl, 2009), and intervening on the parent of a variable will change the distribution of that variable (assuming everything else remains constant). More generally, causal graphs can be used for representing and estimating the effect of interventions (see Pearl, 2009).

Identifying the direct causes also provides a set of variables that can be used for predicting the value of academic achievement. As previously explained, the Causal Markov Condition implies that, ignoring a variable's effects, all the relevant probabilistic information about a variable is contained in its direct causes. It follows from this principle that knowing the parents of academic achievement will help us obtain accurate predictions of academic achievement. Furthermore, these predictions will remain invariant under different contextual conditions (assuming all the parents of the variable have been identified; see Peters, 2016). Given that a variable is only sensitive to its direct causes, the conditional distribution of an outcome given its parents will not change if other variables in the system are manipulated. The ability to make predictions in novel environments is at the heart of our interest in causal understanding, as it justifies generalization and extrapolation (Pearl, 2009).

The idea that identifying the parents of a target variable can be useful for constructing a predictive model reveals a natural connection between causal discovery and feature selection techniques, and researchers have explored and developed the associations between these methodological approaches (e.g., Pellet and Elisseeff, 2008). At the same time, researchers have also emphasized that we should clearly distinguish between causal and "probabilistic" (or "predictive") modelling, as they represent fundamentally different problems, with different assumptions and standards of success (see Spirtes and Zhang, 2016 for a discussion on this topic). Some important differences between causal and probabilistic modelling are the following. First, they have different goals: while causal modelling intends to predict the effect of some intervention (i.e., the value of a variable in a "manipulated" distribution), probabilistic modelling intends to predict the value of some variable in the observed or "unmanipulated" distribution (Spirtes and Zhang, 2016). In Pearl's (2009) terminology, while the causal effect $\Pr\big(Y\big|do(X = x)\big)$ describes

the distribution of Y that would be generated when the variable X is forced on some particular value x, the probabilistic prediction $\Pr(Y|(X = x))$ refers to the distribution of Y for the sub-population where X = x. Second, while causal modelling assumes that the underlying causal relations are correctly specified, probabilistic modelling can be ambiguous regarding the data-generating process. In other words, contrary to probabilistic prediction, estimating $\Pr(Y|do(X = x))$ requires knowledge of the causal graph. Finally, while the success of probabilistic modelling can be assessed using samples from the observed distribution (e.g., by examining the mean squared error of the predictions), the adequacy of a causal model cannot be estimated using non-experimental data (Spirtes and Zhang 2016).

Given that probabilistic and causal prediction represent substantially different problems, they also warrant different methodological approaches. By comparing several probabilistic and causal search algorithms, Aliferis et al. (2010) found that the latter outperformed the former in recovering the local causal structure (i.e., the direct causes and direct effects of a target variable) while achieving high prediction performance. Probabilistic feature selection methods also achieved high prediction performance, but selected variables that were scattered all over the graph (e.g., in the periphery), and therefore did not provide any useful or consistent causal results. As the authors remark, these findings suggest that good predictive performance of non-causal algorithms can be greatly misleading if it is used as a criterion for investigating causal hypotheses. As it is often noted, from the simple observation that a variable is highly predictive of another variable one cannot infer anything about whether (or how) these variables are causally related. An example that is frequently used to illustrate this point is that, even if the tar-stained fingers of the heavy smoker predict negative health outcomes (e.g., lung cancer), this relationship is not causal

(i.e., washing the yellow stain in the fingers will have no effect on the probability of lung cancer or other health outcomes; Aliferis et al., 2010).

In sum, even if having some knowledge of the causal structure (in particular regarding the parents of a target variable) can be useful for prediction, if one is exclusively interested in predictive performance then one can use non-causal feature selection or other machine learning techniques (e.g., Friedman, Hastie and Tibshirani, 2001). If, on the other hand, one is interested in understanding the underlying process that generated the data, then causal search algorithms offer a suitable approach, as they provide a principled method for uncovering how the variables in a system are causally related to each other.

I have emphasized on the importance of identifying the proximal mechanisms for understanding the processes that influence academic achievement as well as for designing effective interventions. However, causal discovery algorithms also provide important insights into the overall causal structure of academic achievement. As explained above, knowledge of the causal structure can be useful for a variety of purposes, including interpreting and using research findings as well as assessing causal modelling assumptions (e.g., classifying a variable as a "proximal" or "distal cause", a "confounder" or a "mediator" requires knowledge of the causal structure). The results of the causal search algorithms can be used to support or contradict some of these assumptions. Consequently, apart from searching for the proximal mechanisms, I also present the entire causal graph generated by the search procedures.

4.5.1 Causal learning algorithms

In this section, I briefly review the main ideas and underlying assumptions behind the learning algorithms that I apply in this study (for a more in-depth discussion of these and other

causal learning methods, see Glymour, Zhang and Spirtes, 2019; or Spirtes and Zhang, 2016). These algorithms can be used to estimate causal structures from observational or experimental data (or a combination of both). The fundamental strategy employed by these algorithms is to estimate a DAG (which encodes causal structures via conditional independencies) based on the actual independencies observed in the data.

The main principles that warrant the inference from observed dependence structures to the causal DAG that generated the data are: (1) the Causal Markov Condition (CMC), according to which every variable is independent of its non-effects conditional on its direct causes; and (2) Faithfulness, according to which the independencies in the data must also exist in the associated DAG. The first condition, previously discussed, implies that every distribution produced by a causal graph has the independence relations obtained by applying d-separation to that graph. The second condition can be interpreted as the converse, as it implies that *all* the independence relations in a distribution generated by a causal graph can be obtained by applying d-separation to the corresponding graph. This condition rules out the possibility that causal paths coincidently cancel each other out, generating independencies that are not implied by the associated graph. In other words, while the CMC implies that the distribution has all the independence relations produced by the graph, the Faithfulness condition implies that the distribution has *only* those independence relations. Together, the CMC and faithfulness warrant some inferences from the independence relations observed in the data to the underlying causal structures (Spirtes et al., 2000, Pearl, 2009).

Common causal search algorithms can be divided into (1) constraint-based algorithms, which construct a DAG based on the conditional independence relations found in the data; and (2) score-based algorithms, which optimize some score (e.g., BIC) given the observed dependence structure (Spirtes and Zhang, 2016). I will describe one algorithm belonging to the first type (PC)

and another belonging to the second type (FGES). In addition to CMC and Faithfulness, these algorithms assume causal sufficiency, i.e., the absence of hidden confounders (Spirtes et al. 2000). Existing algorithms relax this assumption (e.g., FCI), but I do not implement these algorithms in the present study. Finally, the algorithms assume acyclicity, i.e., the absence of cycles in the graph.

The PC algorithm. The Peter-Clark (PC) algorithm (Spirtes et al., 2000) constructs a DAG by conducting a series of conditional independence tests among the observed variables. The algorithm has two stages. First, it determines the undirected edges in the graph (referred to as the "skeleton identification") by starting with a complete undirected graph (i.e., a graph in which each variable is connected to each other), and then removing any edges if the variables involved are conditionally independent. Second, the edges are oriented according to some rules (see Spirtes et al. 2000 for details).

The FGES algorithm. The Fast Greedy Equivalence Search algorithm is an optimized version of the greedy equivalence search (GES) algorithm (see Chickering, 2002), which is a Bayesian algorithm that searches for the DAG with the highest score. Typically, the score used is the Bayesian Information Criterion (BIC), which measures model fit while penalizing the complexity of the DAG. The algorithm has two main stages. First, it starts from an initial estimate and adds edges until the BIC is maximized. Second, the algorithm removes edges until no single edge removal increases the score (see Ramsey et al., 2017 for details).

It is important to note that the two algorithms described above can output multiple causal models rather than a single DAG. The reason for this is that in some cases the algorithm cannot determine the direction of some edges given the available data. If an edge cannot be oriented, then the algorithm outputs an undirected edge, X – Y, which indicates that the true DAG could include $X \rightarrow Y$ or $Y \rightarrow X$. A graph containing both directed and undirected edges is called a *completed*

partially directed acyclic graphs (*CPDAG*) and represents a Markov-equivalence class (Kalisch and Bühlmann, 2014). Two DAGs are considered Markov-equivalent if they imply the same (conditional) independence structure among the observed variables (Eberhardt, 2017).

The FGES algorithm can be used with continuous, discrete or a mixture of discrete and continuous data (i.e., "mixed" data), and the score used in the algorithm as a measure of fit typically makes distribution assumptions (Chickering, 2002). Scores commonly used with continuous variables (e.g., the SEM BIC score) assume Gaussian distributions; scores for discrete variables (e.g., the BDeu score) assume multinomial distributions; and scores for mixed datasets (e.g., the Conditional Gaussian BIC Score) assume that the continuous variables are jointly Gaussian (Andrews Ramsey Cooper, 2018; Ramsey et al., 2017).

The PC algorithm, on the other hand, does not make any assumptions regarding the functional form of the causal relationships between the variables. However, the specific tests used to assess conditional independence typically depend on parametric assumptions. If the variables are continuous, multivariate Gaussian and linearly related, then conditional independence can be assessed using partial correlations, which can be tested using Fisher's z-test; if the variables are categorical, then conditional independence can be examined using G^2 or χ^2 tests; and for a combination of discrete and Gaussian data one can use the Conditional Gaussian test or the Multinomial Logistic Regression Wald Test (Ramsey and Malinsky, 2017; Spirtes et al., 2000; Spirtes et al., 2010). Non-parametric tests (e.g., the conditional correlation independence [CCI] test) have been developed, but can be quite slow and do not scale well to large samples (Ramsey, 2014).

In addition to these parametric assumptions, the statistical tests used in the PC algorithm require a decision regarding the alpha value for rejecting the null hypothesis, which is always that

two variables are (unconditionally or conditionally) independent. This decision should depend on the sample size, with bigger values (e.g. $\alpha = 0.05$) appropriate for small samples, and smaller values (e.g. $\alpha = 0.01$) appropriate for larger samples.

4.6 Data

The data comes from the Early Childhood Longitudinal Study (ECLS-K:2010), conducted by the National Center for Education Statistics (see Tourangeau et al., 2018, for more information regarding this study). The study tracks a nationally representative sample of 18,170 U.S. children who entered kindergarten in the 2010–2011 school year through fifth grade. Sampling weights are provided in the data set in order to account for differential selection at each sampling stage and to adjust for the effects of nonresponse (Tourangeau et al., 2018). In this study, the analytic sample was defined as 6,509 individuals that had a valid sampling weight that maximized the number of sources included in the analysis (which involved child, parent, teacher and school-administrator data from multiple waves).

The amount of missing data in the analytic sample was not substantial, as in most variables the complete cases amounted to around 99%. The only variables with a considerable amount of missing data were *school district poverty* (13%), *peer academic level* (8%) and *neighborhood safety* (8%). In order to include the entire sample in the analysis, single imputation methods were conducted for replacing a single value for each missing data point. Longitudinal imputation using the most recent non-missing value was conducted for *school district poverty, school neighborhood safety* and *school safety*. For the remaining variables, stochastic regression imputation using the appropriate sampling weight was conducted. In order to improve the imputation models, several

auxiliary variables were added, including the previous value of the imputed variables (if available) and demographic characteristics (Nguyen et al., 2017).

The tests used in the search algorithms typically depend on parametric assumptions (e.g., that the variables are multivariate Gaussian and linearly related). I assessed univariate normality by examining frequency distributions and quantile-quantile plots for each variable. These visual representations revealed that 21 variables had clear departures from gaussianity, even after transformation (e.g., they were highly skewed or had gaps in the data). I discretized these 21 variables into 5 categories using the Hartemink discretization algorithm, which considers the mutual information score between all pairs of variables to discretize with minimal information loss (Hartemink, 2001). I also maximized the gaussianity of the remaining 9 continuous variables by conducting a *nonparanormal transformation* (Liu, Lafferty and Wasserman 2009) in Tetrad 6.6.0 (Spirtes et al., 2000). Finally, I created a mixed data set by combining the 9 continuous paranormal variables, the 21 Hartemink discretized variables, and the two categorical variables (gender and race).

Variable selection was determined by two considerations: (1) modularity, i.e., variables should be conceptually distinct from each other and have well-defined manipulable properties; and (2) based on previous research, the variables should be considered important factors for academic achievement. In view of these considerations, thirty-two variables were included in each analysis. These variables comprised demographic, psychological (or child-level) and contextual influences, which can be broadly divided into family, school and neighborhood effects (McKown, 2013). Below, I provide a brief description of all the variables included in the analysis (for a more comprehensive description of most variables see Tourangeau et al. [2018]), as well as the literature supporting their association with academic achievement.

4.6.1 Child-level variables

Reading and math achievement scores. Children completed individualized cognitive assessments in reading and math that were developed for the ECLS (Tourangeau et al., 2018). The items on the assessments combined questions from well-validated and reliable tests, as well as newly developed items. The direct cognitive battery used adaptive testing, based on a three-parameter item response theory (IRT) model, in order to create a common scale and minimize the possibility of ceiling and floor effects (see Najarian et al., 2018, for a detailed psychometric report). I included the estimated ability score (theta) in reading and math in both kindergarten and fourth grade.

Including the previous value of academic achievement is justified by the wide-ranging consensus regarding the existence of state-effects in skill development. However, given the high degree of relative stability in academic achievement, the scores in fourth and third grade are highly correlated (r = .89 in math and r = .83 in reading). The high predictive power of the lagged-value might screen-off other causes of achievement. In order to avoid this problem, I included the lagged-values in kindergarten (rather than in third grade), which have lower correlations with the scores in fourth grade (r = .74 in math and r = .67 in reading). With these lagged scores, collinearity diagnostics were at acceptable levels.

Executive function. Children completed direct assessments of cognitive flexibility using the *Dimensional Change Card Sort* (Zelazo, 2006), and working memory using the *Numbers Reversed task of the Woodcock-Johnson III (WJ III) Tests of Cognitive Abilities* (Woodcock McGrew, and Mather, 2001). Two additional measures of executive function included teacher-

reported attentional focusing and inhibitory control, based on 6 and 7 items from the *Temperament in Middle Childhood Questionnaire* (Simonds and Rothbart 2004), respectively.[18]

Motivation. Ten items adapted from the *Self-Description Questionnaire* (Marsh et al. ,1984) measuring perceived interest and competence in math and reading were included in a child questionnaire administered in third grade (five items for reading and five for math). Perceived interest and self-competence beliefs have been considered important constructs in academic motivation and determinants of academic achievement (e.g., Susperreguy et al., 2018).

Socioemotional behaviors. Four teacher-reported social skills based on the *Social Skills Rating System* (SSRS) (Gresham and Elliott, 1990) were included: self-control (4 items), interpersonal skills (5 items), externalizing problem behaviors (6 items), and internalizing problem behaviors (4 items). In addition, four socioemotional factors were constructed from the *Child Questionnaire* administered in third grade: life satisfaction, composed of six items from the *Assessment of Neurological and Behavioral Function* (Salsman et al., 2013); perceived interest and competence in peer relationships, composed of six items from the *Self-Description Questionnaire* (Marsh et al., 1984); peer victimization, based on four items from the scale by Espelage and Holt (2001); and social anxiety and fear of negative evaluation, based on three items

[18] Even if there is a wide agreement on the importance of executive functions, researchers disagree on what are the most adequate operational definitions and measurement approaches (e.g., Jacob and Parkinson, 2015). Thus, in the present study I include both performance-based measures (related to cognitive flexibility and working memory) and teacher-reported measures (related to attentional focusing and inhibitory control). I included teacher-reported rather than parent-reported ratings, as previous research suggests that the former are more highly predictive of academic outcomes (e.g., Miranda et al., 2015).

from the *Social Anxiety Scale for Children* (Greca and Stone, 1993). Socioemotional behaviors have been considered important determinants of academic achievement (e.g., van Lier et al., 2012).

4.6.2 Family-level variables

Parent education and household income. Parents were asked to report household income in an income range comprised of eighteen categories. Parent education level was recorded in nine categories. The influence of both parent education and family income on academic achievement has been widely described in the literature (e.g., Davis-Kean, 2005).

Home literacy environment. A sum score of three items related to the child's opportunities to engage with texts outside of school was constructed. The items were: the frequency with which parents engaged in joint book reading with the child (4-point scale); the frequency with which children read (or pretended to read) books outside of school (4-point scale); and the frequency with which household members visited the library with the child (dichotomous). Previous findings suggest that home literacy environment is significantly related to academic outcomes (Aikens and Barbarin, 2008).

Parental strain and parental warmth. Parental strain was constructed as the average of four items representing the difficulty and strain of functioning as a parent, and parental warmth as the average of four items representing the closeness of the parent-child relationship (Aikens and Barbarin, 2008). Parental strain can be regarded as a proxy of parental stress, and parental warmth as a proxy of parental care and sensitivity. Previous research suggests that these parental practices and behaviors can affect achievement and executive function (Aikens and Barbarin, 2008; Bernier et al., 2010).

Parental involvement in the child's school. A factor was extracted from four items related to parents' involvement in the child's school: attending a parent–teacher conference; attending a parent–teacher association (PTA) meeting; volunteering; and attending to a school event. Parental involvement is believed to be associated to children's academic outcomes (e.g., Galindo and Sheldon, 2012).

4.6.3 School-level variables

Peer academic level. Teachers reported on the number of students below grade level in the classroom (in both reading and math). Previous research suggests that students are affected by the achievement level of their peers (e.g., Hoxby, 2000).

Student-teacher relationships. Measures of closeness and conflict between the teacher and the child were included, based on 7 and 8 items from *The Student-Teacher Relationship Scale* (STRS) (Pianta, 2001), respectively. Teacher-Child relationships have been found to predict academic achievement (e.g., Lowenstein et al., 2015).

Socioeconomic composition of schools. School administrators were asked for the percentage of children eligible for free or reduced-price lunch. In addition, a measure of the percentage of children in a school district who are in poverty was derived from the 2013 *Small Area Income & Poverty Estimates* (SAIPE). The effects of poverty on academic achievement have been widely discussed in the literature (e.g., Hair et al., 2015).

School safety. A factor was extracted from 6 items from the school administrator questionnaire, regarding several school safety issues (weapons, theft, physical conflicts, illegal drugs, vandalism and student bullying). Previous studies suggest that safe and supportive environments in schools can have an effect on academic achievement (e.g., Burdick-Will, 2013)

4.6.4 Neighborhood-level variables

Home neighborhood safety. A measure was constructed by taking the average of parent's responses to 3 items related to neighborhood safety (how safe is to play outside; burglary/robbery in the area; and problems with selling/using drugs or alcohol). Previous research suggests that disadvantaged neighborhoods can have lasting negative effects on academic outcomes (e.g., Wodtke, Harding and Elwert, 2011).

School neighborhood safety. A factor was extracted from the school administrator's answers to 6 items related to school neighborhood safety (regarding problems with violence, crime, vacant spaces, gangs, drugs and tensions based on racial, ethnic, or religious differences). Previous findings have found associations between differences in school neighborhood environments and academic outcomes (e.g., Milam et al., 2010).

4.6.5 Demographic variables

Finally, I included three demographic variables (age, gender and race) which have been consistently found to be associated with academic achievement (e.g., Aikens and Barbarin, 2008). The child's age reflected the age in months at the time of the 4th grade assessment (M = 121 , SD = 4.4). There were approximately the same number of males and females in the analytic sample. The variable for race consisted of five categories: White (58%), Black (8%), Hispanic (22%), Asian (6%) and other (6%).

Some of the data gathered by the ECLS differed across waves. In particular, the motivation measures, which have been considered important proximal mechanisms, were only included in 3rd grade. Thus, the present study focuses on 3rd and 4th grade, as for this period one can include

simultaneous measures of most relevant variables. The only variables that were not gathered in 4th or 3rd grade are: *school safety*, *school neighborhood safety* and *home literacy environment* (2nd grade); *parental strain* and *the percentage of children eligible for free or reduced-price lunch* (1st grade); and *home neighborhood safety* (kindergarten). Given that the outcome of interest is academic achievement (either reading or math), I included one achievement score in 4th grade, and another one in kindergarten (the inclusion of two achievement scores allowed to investigate the causal link with previous achievement).

Two datasets including reading-related and math-related measures were created. These datasets differed in four content-specific variables: the two achievement scores; the motivation measures; and peer academic level. The reason for creating separate dataset is that considering the two outcomes jointly might pose some threats to the modularity assumption, as reading and math achievement have been considered overlapping constructs, especially in the age range examined in this study. In particular, previous studies suggest that some cognitive processes (e.g., phonological processing abilities) are required for both reading and math skills (e.g., Harlaar et al., 2012), and this common set of processes presumably explains the high correlation that is typically found between reading and math abilities, as well as the high comorbidity of dyslexia and dyscalculia (e.g., Bailey et al., 2017; Pimperton and Nation, 2010). In the sample considered in the analysis, the correlation between reading and math scores is 0.75 in kindergarten and 0.72 in fourth grade. Given that the modularity assumption might be violated, I estimated the causal structure for the two outcomes separately. However, in order to examine the extent to which the results are sensitive to model specification, I also considered the two outcomes jointly (i.e., including math and reading scores in both kindergarten and 4th grade). For this analysis, I only present the variables directly related to academic achievement, which is the main focus of this study.

Finally, it is worth noting that most of the variables included contain some measurement error, which can affect the output of causal search algorithms (see Zhang et al., 2017 for a discussion on this topic, as well as on the identifiability conditions of causal models underlying measurement-error-free variables). Even if the algorithms that I implement do not take into account measurement error, concerns in this regard are alleviated by the fact that (1) the majority of variables are factors composed of multiple items; and (2) most variables have good reliability (e.g., Cronbach's alpha for the achievement scores, executive functions, motivation measures, social skills and student-teacher relationship measures is above 0.8; see Tourangeau et al., 2018 and the references cited therein for measurement details).

4.7 Empirical analysis

The purpose of the empirical analysis is to identify the DAG encoding the causal structure of academic achievement. This DAG will contain the parents (direct causes) and ancestors (indirect causes) of both reading and math achievement. The search procedure was divided in two steps. First, I identified the skeleton of the graph, i.e., the graph representing the adjacencies (but not the orientations) common to a class of graphs encoding the same conditional independencies (a Markov equivalence class) as the causal DAG. Second, I used a different set of algorithms to orient the edges of the undirected graph estimated in the first step. All search algorithms were implemented using the Tetrad 6.6.0 Freeware (Spirtes et al., 2000).

4.7.1 Skeleton identification

Given the scarcity of applications of causal search algorithms to real-world data, we do not have yet clear guidelines regarding the accuracy of these algorithms under different scenarios. In order to identify the skeleton, I imposed then stringent requirements. In particular, I conducted four different search procedures and considered the agreement in the results. The rationale for choosing these procedures is that they vary important aspects of the algorithms that might affect the results. The search procedures were: (1) the FGES algorithm using mixed data; (2) the PC algorithm using mixed data, the Mixed Multinomial Logistic Regression Wald (MMLRW) test (Ramsey and Malinsky 2017), and an alpha level of 0.01; (3) the PC algorithm with the same specifications as (2) but with an alpha level of 0.001; and (4) the PC algorithm using the untransformed continuous data, the Conditional Correlation Independence (CCI) test (Ramsey 2014), and an alpha level of 0.01. Due to computational demands, procedure (2) was performed with a random subsample of 2,000 individuals, and procedure (4) with a random subsample of 1,000 individuals.

Demanding an agreement across these procedures ensures that the results are consistent across important dimensions of causal search algorithms. In particular, the results will hold across search methods (as I use both constraint-based [PC] and score-based [FGES] algorithms); the alpha level of the conditional independence tests (as I use both $\alpha = 0.01$ and 0.001); parametric assumptions related to the distributional form (as I use both the transformed and untransformed data sets); as well as linearity assumptions of the statistical tests (as I use both parametric [MMLRW] and non-parametric [CCI] tests of conditional independence). The problem of demanding complete agreement across the four procedures is that the resulting graph might be too sparse. Thus, I considered the possibility of partial agreement across procedures: "high-

confidence" edges refer to the edges that were identified across all four procedures, while "low-confidence" edges refer to the edges that were identified in only three procedures.

In order to increase the robustness of the results, each search procedures was conducted multiple times on random samples of the data. More specifically, the FGES algorithm was bootstrapped (with replacement) 500 times, and the PC algorithms were bootstrapped (with replacement) 20 times (the difference in the number of resampling runs is due to different computational demands). If an edge (either directed or undirected) appeared in 50% or more of the bootstrap samples, then the edge was *retained*. If an edge was retained across four or three of the search procedures, then it was *identified*.

4.7.2 Edge orientation

In order to orient the edges, I followed a three-step process. First, I conducted two of the search procedures described above (PC using mixed data and an alpha level of 0.01, and PC using the raw data and the CCI test; each bootstrapped 20 times) but imposed constraints on the orientation of the edges based on time order. In particular, I forbid achievement scores in 4th grade from causing anything else, and the variables measured in 3rd grade from causing the variables that were already predetermined (parent education; family income and district poverty) or measured earlier (home neighborhood safety; achievement scores in kindergarten; home literacy environment; percent of students eligible for free or reduced-price lunch; and parental strain). Finally, I considered the three demographic characteristics (age, gender and race) as exogenous variables (i.e., they cannot be caused by any other variable in the dataset). I oriented the edge if the results were not contradictory among the two procedures.

In order to orient the remaining edges, I considered whether the 4 or 3 procedures that identified the edge agreed on the orientation. Finally, for all the edges that remained unoriented, I used the raw data and considered the agreement between two orienting algorithms, R3 and RSkew (with 500 bootstraps each), which appeal to the non-Gaussian information of the variables to orient the edges (see Ramsey, Sanchez-Romero and Glymour, 2014 for details on these algorithms). It is worth noting that R3 and RSkew were applied to the skeleton containing the high and low-confidence edges previously identified. On the other hand, the two other orientation methods were not based on any particular graph (the first method only imposed some constraints based on time order, and the second did not imposed any constraints).

4.8 Results

4.8.1 Reading achievement

Four causal search procedures that varied important dimensions in the algorithms were applied to the reading dataset. Using the mixed reading dataset, the FGES algorithm retained 41 edges; the PC algorithm with an alpha level of 0.01 retained 76 edges; and the PC algorithm with an alpha level of 0.001 retained 77 edges. Compared to the PC algorithm, the FGES algorithm tends to generate sparse graphs, as it penalizes complexity by optimizing the BIC. Using the continuous (raw) reading dataset, the PC algorithm with a nonparametric test (CCI) retained 81 edges. Among all edges, 29 were identified in the 4 procedures (high-confidence edges), and 27 were identified in 3 procedures (low-confidence edges). Table 3 presents the frequencies in which the edges were found in each bootstrapped sample. For example, the first row in the top panel

(*High-confidence edges)* indicates that the edge between *Working memory and Reading (4th)* was found in every bootstrapped sample across all search procedures; on the other hand, the first row in the bottom panel (*Low-confidence edges)* indicates that the edge between *Cognitive flexibility and Reading (4th)* was found in 7% of the bootstrapped samples in the FGES procedure (and, given that the value is below 50%, it was not retained); in 100% of the PC (both $\alpha = 0.01$ and 0.001) procedures; and in 55% of the bootstrapped samples in the PC CCI procedure.

In order to orient the edges, I considered the agreement between two search procedures (PC with $\alpha = 0.01$ and PC with the CCI test) with some constraints based on time order. Most edges (48 out of 56) were oriented using this method. Subsequently, I examined whether the procedures used to identify the edge agreed on the orientation, and three edges were oriented utilizing this method. Finally, I considered the agreement of two orientation algorithms (R3 and RSkew) that use the non-gaussianity of the distributions to orient the edges, and another three edges were oriented using this method. Two edges (*Attentional focus — Inhibitory control* and *Home literacy environment — FRPL*) remained unoriented.

The estimated DAG related to reading achievement is displayed in Figure 5. As expected, the contextual and psychological factors considered in the analysis are related in intricate ways. Moreover, the likely presence of unobserved nodes and other additional complexities (e.g., cycles) might complicate the true causal structure even further. As explained above, however, an important advantage of causal graphs is that it allows us to decompose the joint distribution over all of the

variables in terms of the conditional probability distribution of each variable given a (potentially) smaller group of variables.[19]

As Figure 5 illustrates, the results suggest (with high confidence) that the parents of reading achievement are working memory, previous reading achievement, attentional focus and parent education; with low-confidence, the results suggest that peer reading level, perceived interest/competence in reading, peer victimization and cognitive flexibility have also direct relationships with reading achievement. Interpreted causally, these direct links imply that manipulating these variables can lead to a change in reading achievement. Probabilistically, these links indicate that, in order to estimate the value of reading achievement, we only need to consider the value of its 8 parents rather than all 31 variables (I expand on this point below). Figure 5 also displays directed paths representing indirect causes of reading achievement. Some of these paths can involve several nodes, for example: *Race → School district poverty → School safety → Peer reading level → Reading (4th)*.

[19] It is worth noting that the data was constructed in such a way that the target variable (academic achievement in 4th grade) could not have any effects, as no endogenous variable was measured after or during 4th grade (and the cause always precedes its effect). As a consequence, the only variables that could "shield" the target variable from other variables are its parents. However, this is not the case for other variables in the dataset, which might have both direct causes and direct effects in the dataset. In this case, the set of variables that shield the target variable from other variables is called the "Markov blanket", which is composed by the variable's direct causes, direct effects, and the direct causes of the direct effects (Pellet and Elisseeff, 2008).

Table 3. High and low-confidence directed and undirected edges using the reading dataset

	Edges identified in the bootstrapped samples (%)				Orientation method		
	FGES	PC (α = .01)	PC (α = .001)	PC CCI	PC + Constraints	FGES & PCs	R3 & RSkew
High-confidence edges							
Working memory → Reading (4th)	1.00	1.00	1.00	1.00	Yes		
Attentional focus → Reading (4th)	0.95	0.90	1.00	0.90	Yes		
Parent education → Reading (4th)	1.00	1.00	1.00	1.00	Yes		
Reading (K) → Reading (4th)	1.00	1.00	1.00	1.00	Yes		
Reading (K) → Working memory	0.95	1.00	1.00	1.00	Yes		
Interpersonal skills → Teacher closeness	1.00	1.00	1.00	1.00	Yes		
Family income → School district poverty	1.00	1.00	1.00	1.00	Yes		
FRPL → School district poverty	1.00	1.00	1.00	1.00			Yes
Inhibitory control → Externalizing behaviors	1.00	1.00	1.00	1.00	Yes		
Self-control → Externalizing behaviors	1.00	0.95	1.00	1.00	Yes		
Gender → Inhibitory control	1.00	0.55	0.60	1.00	Yes		
Attentional focus − Inhibitory control	1.00	1.00	1.00	1.00			
Attentional focus → Internalizing behaviors	0.66	0.80	1.00	1.00	Yes		
Self-control → Interpersonal skills	1.00	1.00	1.00	1.00	Yes		
Life satisfaction → Int./Comp. in peer relationships	1.00	1.00	1.00	1.00	Yes		
Family income → Parent education	1.00	1.00	1.00	1.00	Yes		
Parental strain → Parental warmth	1.00	1.00	1.00	0.95	Yes		
Race → School district poverty	0.57	1.00	1.00	1.00	Yes		
Race → Home neighborhood safety	0.86	0.95	1.00	1.00	Yes		
Peer victimization → School anxiety	1.00	1.00	1.00	1.00	Yes		
FRPL → School safety	0.99	1.00	1.00	0.95	Yes		
School neighborhood safety → School safety	0.98	1.00	1.00	1.00	Yes		
Inhibitory control → Self-control	0.91	0.90	1.00	1.00	Yes		
Teacher closeness → Teacher conflict	0.80	0.65	1.00	1.00	Yes		
Externalizing behaviors → Teacher conflict	1.00	1.00	1.00	1.00	Yes		
Cognitive flexibility → Working memory	1.00	1.00	1.00	1.00	Yes		
Gender → Int./Comp. in reading	0.50	0.50	0.90	1.00	Yes		
School anxiety → Int./Comp. in peer relationships	0.99	1.00	1.00	0.55		Yes	

Race → Home literacy environment	0.57	0.95	1.00	0.55	Yes		
Low-confidence edges							
Cognitive flexibility → Reading (4th)	**0.07**	1.00	1.00	0.55	Yes		
Peer reading level → Reading (4th)	**0.00**	0.90	1.00	0.80	Yes		
Int./Comp. in reading → Reading (4th)	**0.00**	1.00	1.00	1.00	Yes		
Peer victimization → Reading (4th)	**0.00**	0.80	0.95	0.65	Yes		
School safety → Peer reading level	**0.11**	0.95	1.00	0.55			Yes
School neighborhood safety → Peer reading level	**0.13**	0.90	1.00	0.95			Yes
School district poverty → Parent involvement in school	**0.00**	1.00	1.00	0.80	Yes		
School district poverty → School safety	**0.33**	1.00	1.00	0.95	Yes		
Attentional focus → Externalizing behaviors	**0.00**	0.90	0.90	1.00	Yes		
Gender → Teacher closeness	**0.08**	0.55	0.75	0.55	Yes		
Gender → School anxiety	**0.39**	0.65	1.00	0.95	Yes		
Home literacy environment → Parental warmth	**0.01**	0.95	0.65	0.95	Yes		
FRPL − Home literacy environment	**0.04**	0.95	1.00	0.60			
Life satisfaction → Peer victimization	**0.18**	0.55	0.60	0.60	Yes		
Race → Family income	**0.38**	0.95	1.00	1.00	Yes		
Race → Parent education	**0.44**	1.00	1.00	0.90	Yes		
Race → School neighborhood safety	**0.27**	0.90	1.00	0.75	Yes		
Family income → FRPL	**0.26**	1.00	1.00	0.95	Yes		
Home neighborhood safety → School neighborhood safety	**0.00**	0.75	1.00	1.00	Yes		
Internalizing behaviors → Teacher conflict	**0.35**	0.55	0.95	1.00	Yes		
Self-control→ Teacher conflict	**0.03**	1.00	1.00	1.00	Yes		
Interpersonal skills → Teacher conflict	**0.19**	0.95	0.85	0.75	Yes		
Inhibitory control → Teacher conflict	**0.35**	0.60	0.85	0.90	Yes		
Internalizing behaviors → Int./Comp. in peer relationships	**0.02**	0.90	0.80	0.90	Yes		
FRPL → Peer reading level	0.60	**0.10**	1.00	0.90		Yes	
FRPL → School neighborhood safety	0.50	1.00	0.95	**0.00**	Yes		
Peer victimization → Externalizing behaviors	1.00	0.90	1.00	**0.40**		Yes	

Note. The high-confidence edges refer to the edges that were identified across all four search procedures, while low-confidence edges refer to the edges that were identified in only three procedures. An edge was retained if it was found in at least 50% of the bootstrapped samples. Bolded numbers indicate that the edge was found in less than 50% of the bootstrapped samples, and therefore it was not retained. For computational reasons, the FGES algorithm was bootstrapped 500 times and the PC algorithms 20 times. In addition, the PC ($\alpha = .01$) was applied to a random subsample of 2,000 individuals, and the PC CCI to a random subsample of 1,000 individuals. FRPL refers to the percentage of students eligible for free or reduced-price lunch. See text for details on the search procedures.

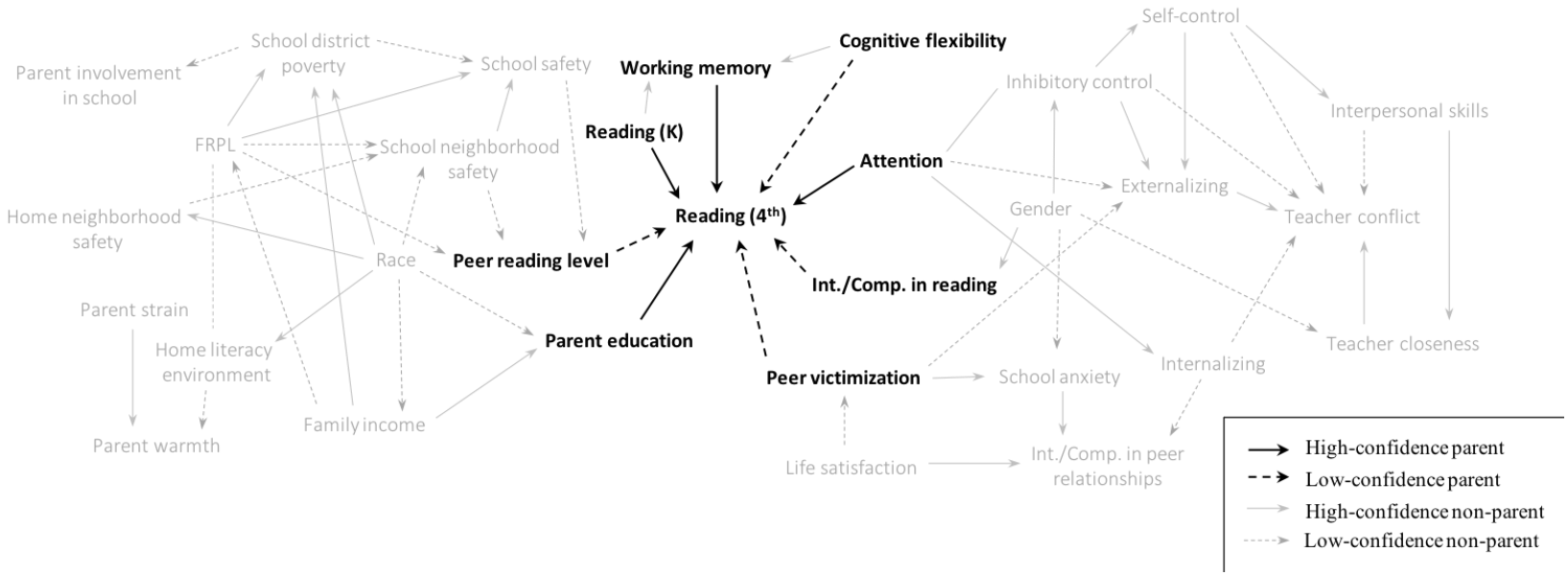


Figure 5. Estimated DAG with the parents and non-parents of reading achievemen

4.8.2 Math achievement

The four search procedures were also applied to the math dataset. As in the reading dataset, the FGES algorithm retained relatively few edges (39) and the PC algorithm with the CCI test retained the largest number of edges (81). The PC algorithm with an alpha level of 0.01 retained 73 edges and PC an alpha level of 0.001 retained 77 edges. Among all edges, 26 were identified in all 4 procedures (high-confidence edges), and 28 were identified in 3 procedures (low-confidence edges). As with the reading dataset, most edges were oriented using the two PCs plus constraints method, except for two edges that were oriented using the agreement in the original three procedures, and four edges that were oriented using the R3 and RSkew algorithms. Two edges remained unoriented. Table 4 displays the oriented high and low-confidence edges with their respective frequencies across the bootstrapped samples.

The estimated graph using the math dataset is depicted in Figure 6. Similar to the estimated causal structure using the reading dataset, one can perceive that previous achievement, working memory, attentional focusing, cognitive flexibility and peer victimization are directly related to math achievement. Contrary to the reading case, however, gender and race have a direct edge with math achievement, while peer academic level, perceived interest/competence in the skill, and parent education have no direct connection to math achievement in 4_{th} grade. One also notice that, contrary to the reading case, the child's age has a direct edge to math achievement in kindergarten.

Finally, by comparing the results in Tables 3 and 4 one can perceive that most edges were identified in both datasets. Ignoring the subject-specific edges, all high-confidence and most low-confidence edges were identified in both datasets; this consistency strengthens the confidence in the results.

Table 4. High and low-confidence directed and undirected edges using the math dataset

	Edges identified in the bootstrapped samples (%)				Orientation method		
	FGES	PC (α = .01)	PC (α = .001)	PC CCI	PC + Constraints	FGES & PCs	R3 & RSkew
High-confidence edges							
Working memory → Math (4th)	1.00	1.00	1.00	1.00	Yes		
Math (K) → Math (4th)	1.00	1.00	1.00	1.00	Yes		
Cognitive flexibility → Math (4th)	0.82	0.95	1.00	0.60	Yes		
Child age → Math (K)	0.94	1.00	1.00	1.00	Yes		
Math (K) → Working memory	0.74	0.95	1.00	1.00	Yes		
Attentional focus → Inhibitory control	1.00	1.00	1.00	1.00	Yes		
FRPL → Peer math level	0.95	1.00	1.00	1.00	Yes		
Interpersonal skills → Teacher closeness	1.00	1.00	1.00	1.00	Yes		
Cognitive flexibility → Working memory	0.98	0.95	1.00	0.85	Yes		
Family income → School district poverty	0.98	1.00	1.00	1.00	Yes		
FRPL → School district poverty	1.00	1.00	1.00	1.00			Yes
Inhibitory control → Externalizing behaviors	1.00	1.00	1.00	1.00	Yes		
Self-control → Externalizing behaviors	1.00	0.95	1.00	1.00	Yes		
Family income → Parent education	1.00	1.00	1.00	1.00	Yes		
Attentional focus → Internalizing behaviors	0.61	0.95	1.00	1.00	Yes		
Self-control → Interpersonal skills	1.00	1.00	1.00	1.00	Yes		
Life satisfaction → Int./Comp. in peer relationships	1.00	1.00	1.00	1.00	Yes		
School anxiety − Int./Comp. in peer relations	1.00	1.00	1.00	0.95			
Parental strain → Parental warm	0.99	1.00	1.00	1.00	Yes		
Race → Home neighborhood safety	0.95	1.00	1.00	1.00	Yes		
Peer victimization → School anxiety	1.00	1.00	1.00	1.00	Yes		
School neighborhood safety → School safety	0.99	1.00	1.00	1.00	Yes		
Inhibitory control → Self-control	0.99	0.95	1.00	1.00	Yes		
Teacher closeness → Teacher conflict	0.83	0.95	1.00	1.00	Yes		
Externalizing behaviors → Teacher conflict	1.00	1.00	1.00	1.00	Yes		
Race → School district poverty	0.50	1.00	1.00	1.00	Yes		
Low-confidence edges							

Race → Math (4th)	**0.14**	1.00	1.00	0.50	Yes		
Gender → Math (4th)	**0.07**	0.80	1.00	0.90	Yes		
Peer victimization → Math (4th)	**0.00**	0.95	0.80	0.60	Yes		
Attentional focus → Math (4th)	**0.11**	0.50	0.85	0.80		Yes	
Math (K) → Cognitive flexibility	**0.13**	0.90	1.00	0.70	Yes		
School safety → Peer math level	**0.12**	1.00	1.00	0.90			Yes
School neighborhood safety → Peer math level	**0.03**	0.90	1.00	0.80			Yes
Home neighborhood safety → School district poverty	**0.00**	0.95	1.00	0.85	Yes		
School district poverty → Parent involvement in school	**0.00**	1.00	1.00	0.65	Yes		
School district poverty → School safety	**0.42**	0.80	1.00	0.95	Yes		
Gender → Teacher closeness	**0.15**	0.90	0.70	0.90	Yes		
Gender → School anxiety	**0.47**	0.55	0.95	1.00	Yes		
Internalizing behaviors → Int./Comp. in peer relationships	**0.06**	0.75	0.80	0.80	Yes		
Race → Family income	**0.36**	0.85	1.00	1.00	Yes		
Race → Parent education	**0.36**	1.00	1.00	1.00	Yes		
Family income → FRPL	**0.36**	0.95	1.00	1.00	Yes		
Family income → School neighborhood safety	**0.10**	0.65	0.80	1.00	Yes		
Inhibitory control → Teacher conflict	**0.00**	0.60	0.95	0.55	Yes		
Internalizing behaviors → Teacher conflict	**0.37**	1.00	1.00	0.85	Yes		
Self-control → Teacher conflict	**0.05**	0.85	1.00	1.00	Yes		
Race → School neighborhood safety	**0.25**	0.60	1.00	0.80	Yes		
Interpersonal skills → Teacher conflict	**0.14**	0.65	0.90	0.80	Yes		
Internalizing behaviors → School anxiety	**0.00**	0.50	0.55	0.60	Yes		
Gender → Inhibitory control	0.99	**0.25**	0.85	0.85	Yes		
Externalizing behaviors → Peer victimization	1.00	**0.05**	1.00	0.85			Yes
Parent education → Peer math level	0.68	**0.45**	0.85	0.50	Yes		
FRPL → School safety	1.00	1.00	1.00	**0.25**		Yes	
Int./Comp. in math − Int./Comp. in peer relationships	0.80	0.55	1.00	**0.10**			
Math (K) → Attentional focus	0.89	0.70	0.80	**0.35**	Yes		
Race → Home literacy environment	0.69	0.60	1.00	**0.30**	Yes		

Note. The high-confidence edges refer to the edges that were identified across all four search procedures, while low-confidence edges refer to the edges that were identified in only three procedures. An edge was retained if it was found in at least 50% of the bootstrapped samples. Bolded numbers indicate that the edge was found in less than 50% of the bootstrapped samples, and therefore it was not retained. For computational reasons, the FGES algorithm was bootstrapped 500 times and the PC algorithms 20 times. In addition, the PC (α = .01) was applied to a random subsample of 2,000 individuals, and the PC CCI to a random subsample of 1,000 individuals. FRPL refers to the percentage of students eligible for free or reduced-price lunch. See text for details on the search procedures.

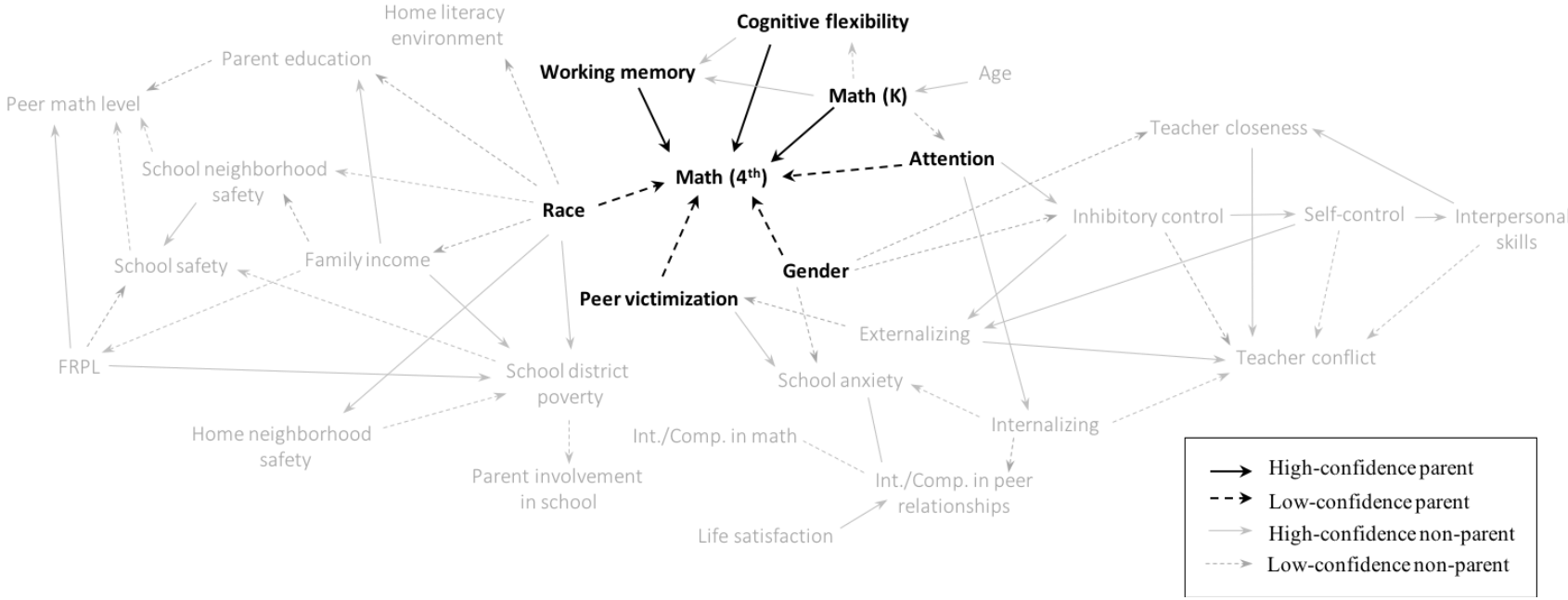


Figure 6. Estimated DAG with the parents and non-parents of math achievement.

4.8.3 Reading and math achievement combined

As previously explained, I considered reading and math achievement separately, due to concerns regarding the modularity assumption (as these constructs might overlap, particularly in the age range under consideration). However, in order to test the sensitivity of the results to model specification, I implemented the four search procedures previously described to a combined dataset with all 36 subject-specific and non-subject specific variables. Figure 7 illustrates the high and low-confidence edges directly connected to any achievement score. As the figure shows (see also Table 5), previous achievement and working memory remain directly linked to both reading and math achievement. Perceived interest/competence in reading, attentional focusing and parent education also remains directly related to reading achievement, while cognitive flexibility, gender and the child's age remains directly related to math achievement. In contrast to prior results, cognitive flexibility, peer reading level and peer victimization were not identified as parents of reading achievement, and race, attentional focusing and peer victimization were not identified as parents of math achievement.

Table 5. High and low-confidence directed and undirected edges using the combined dataset

	Edges identified in the bootstrapped samples (%)			
	FGES	PC (α = .01)	PC (α = .001)	PC CCI
High-confidence edges	1.00	1.00	1.00	1.00
Math (K) − Reading (K)	1.00	1.00	1.00	1.00
Math (K) → Math (4th)	0.71	1.00	1.00	1.00
Math (K) → Cognitive flexibility	1.00	1.00	1.00	1.00
Working memory → Math (4th)	0.78	0.95	1.00	0.85
Child age → Math (K)	0.72	1.00	1.00	0.65
Reading (K) → Working memory	1.00	1.00	1.00	1.00
Reading (K) → Reading (4th)	0.98	0.55	0.65	0.70
Reading (4th) → Math (4th)	1.00	0.95	1.00	0.85
Int./Comp. in reading → Reading (4th)	1.00	1.00	1.00	1.00
Low-confidence edges				
Math (K) → Working memory	**0.00**	0.60	1.00	0.80
Cognitive flexibility → Math (4th)	**0.24**	0.70	1.00	1.00
Gender → Math (4th)	**0.00**	1.00	1.00	0.85
Working memory → Reading (4th)	**0.04**	0.65	1.00	1.00
Parent education → Reading (4th)	**0.00**	1.00	1.00	0.65
Attentional focus → Reading (4th)	0.78	**0.40**	0.95	0.95

Note. The skeleton identification and orientation procedures were the same as above. Except for one edge that remained unoriented, all edges were oriented using the two PCs plus constraints method.

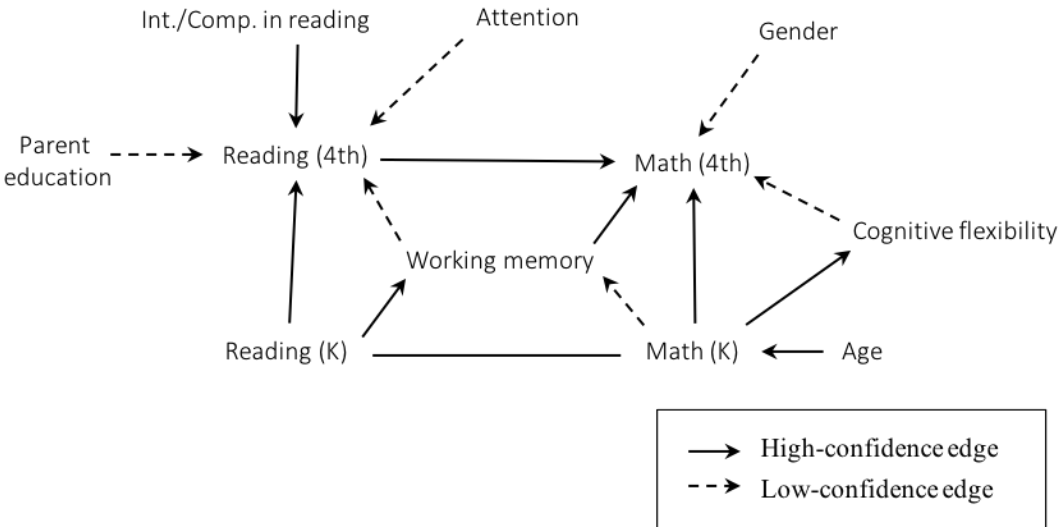


Figure 7. Estimated DAG with the edges directly related to academic achievement.

4.8.4 Using the estimated proximal mechanisms to predict the value of academic achievement

As explained above, identifying the parents can be useful for obtaining accurate predictions of academic achievement. In order to corroborate this claim, I compared the predictive performance of an (ordinary) least squares (OLS) regression model containing the entire set of variables (i.e., the thirty-one predictors in either the reading or math datasets) with an OLS model containing only the estimated high and low-confidence parents of reading or math achievement. Following the results presented in Tables 3 and 4, eight variables were considered the parents of reading achievement: cognitive flexibility, working memory, previous achievement, peer reading level, parent education, peer victimization, attentional focus and perceived interest/competence in reading; and seven variables were considered the parents of math achievement: working memory,

attentional focus, previous achievement, cognitive flexibility, race, gender and peer victimization. I focused on OLS regression, given that it is the most common method in the field (a full comparison with the predictive performance of other estimation procedures is beyond the scope of this study).

In order to assess the accuracy of the predictions, I used *k*-fold cross-validation, which estimates a model's ability to fit out-of-sample data (Friedman, Hastie and Tibshirani, 2001). In this method, the observed sample is randomly divided into *k* groups of approximately equal size, $k - 1$ groups are used as training data and the remaining group is used as validation or test data. The procedure is repeated *k* times, and as a consequence each observation is used in both the training and validation set. The predictive accuracy is assessed using the average prediction error across the *k* estimation procedures.

Following standard practice, I set $k = 10$, and I used the root mean square error (RMSE) as a measure of fit, which can be interpreted as the average distance between the observed values and the model predictions. It is calculated as follows:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

where *n* is the number of observations in the test set, y_i is the observed value, and $\hat{y}_i$ is the predicted value for the *i*th observation in the test set. I also report the test set R^2, which indicates the proportion of variance explained by the model in the test set (and which should be interpreted as a measure of correlation rather than accuracy; e.g., Kuhn and Johnson 2013).

A summary of the cross-validation results is displayed in Table 6. One can perceive that the predictive performance of the models with all thirty-one predictors is comparable to the performance of the models that only include the estimated parents of reading or math achievement. In particular, the RMSEA indicates that, conditional on its parents, the remaining variables do not considerably improve the prediction accuracy (especially in the case of reading). The R^2 is also similar across specifications, and indicates that 56% of the variance in reading and 63% of the variance in math achievement can be explained using the identified parents. These results are consistent with the idea that knowledge of the causal structure can be useful for constructing predictive models.

Table 6. Summary of cross-validation results.

	RMSEA		R^2	
Target variable	All variables	Parents only	All variables	Parents only
Reading	0.40	0.40	0.57	0.56
	(.005)	(.003)	(.010)	(.011)
Math	0.45	0.46	0.65	0.63
	(.005)	(.006)	(.008)	(.011)

Note. The numbers in parentheses refer to the standard error of the cross-validation estimate. All models were fit using ordinary least-squares regression. The models with "all variables" included 31 predictors, and the model predicting reading and math with "parents only" included 8 and 7 predictors, respectively. The untransformed dataset was used in the analysis and, except for gender and race, all variables were considered continuous.

4.9 Discussion

Causal search algorithms have been effectively applied in different fields, including biology (Sachs 2005), genetics (Verdugo et al., 2013), climate science (Ebert-Uphoff and Deng, 2012), medicine (Kalisch et al., 2010) and neuroscience (Sanchez-Romero et al., 2019). However, there have been scant applications of these methods in social and behavioral sciences. This paper argues that concepts and methods related to causal graphs in general and causal search algorithms in particular can offer valuable tools in these domains, as the latter normally deal with large numbers of variables with a (mostly) unknown causal structure. Furthermore, in these domains conducting randomized experiments to identify entire causal structures is often practically or ethically unfeasible. I also argued that a principled and effective strategy for examining complex causal structures (such as the contextual determinants of academic achievement) is to search for the proximal mechanisms (i.e., the direct causes) of the variable of interest. Identifying these proximal mechanisms would help investigate and explain the phenomena under investigation; guide effective interventions; and obtain stable and accurate predictions of the target variable.

Using a nationally representative dataset with a wide range of relevant contextual and psychological factors, I applied four causal search procedures that varied important dimensions in the search algorithms. Consistent with previous research, the algorithms identified prior achievement and executive functions (in particular working memory, cognitive flexibility and attentional focusing) as direct causes of both reading and math achievement. The motivational construct included was identified as a direct cause of reading but not of math achievement. In addition, while parent education, peer victimization and peer reading level were only directly related to reading achievement, race and gender were only directly related to math. As noted in the text, however, these results are maintained with different degrees of confidence.

Even if I included a wide range of variables measuring relevant psychological and contextual factors, the presence of unmeasured confounders remains a likely possibility. Confounding variables can generate erroneous graphs, for example by including spurious direct connections (Spirtes et al., 2000). Additional errors in the estimated graph might be generated by lack of statistical power; violations of faithfulness and modularity assumptions; feedback loops; selection bias; measurement error and inadequate time intervals. Given that –to a greater or lesser extent– these assumptions could have certainly been violated in the present analysis, the estimated graphs should not be considered representations of the true causal structures.

It is worth mentioning, then, that causal search algorithms do not "solve" the problem of causal inference, and that causal inference from observational data remains, as several authors have noted (e.g., Dawid, 2008; Greenland, 2010), a largely speculative and difficult to validate exercise. However, even if the estimated graphs should not be considered representations of the true causal structures, this approach provides valuable insights that cannot be obtained with other commonly used methods. First, the structural relationships estimated using causal search algorithms support interesting causal conjectures that can be examined using experimental designs (Dawid, 2008). For example, the results of the present study can be further investigated by examining whether manipulating working memory, cognitive flexibility or attentional focus affect academic achievement.

Second, causal graphs provide useful tools for synthesizing and organizing previous research. Compared to list-like or other structureless representations, causal graphs can be used to convey what we know or do not know about the actual data-generating process. This would allow us to focus not only on which factors are causally related to a target variable, but also on how different causes relate (or do not relate) to each other, as well as to other supporting factors.

Third, causal graphs can be useful predictive devices. For example, the results of this study suggest that, among the thirty-one variables considered, changes in eight or seven variables (the estimated parents of reading and math achievement, respectively), are sufficient to obtain accurate predictions of academic achievement. These results can be useful in a field accustomed to condition on a large number of variables in a regression framework (e.g., using a similar dataset to the one used in this study, Fryer and Levitt [2004] implement regression models with 98 covariates).

Fourth, causal graphs can be used for determining the identifiability (or non-identifiability) of causal effects from observational data (e.g., Pearl, 2009; Morgan and Winship, 2015). For example, based on Figure 3 (and assuming causal sufficiency), one can conclude that the effect of cognitive flexibility on reading achievement can be estimated by conditioning on working memory (as this would block the only back-door path, i.e., *Cognitive flexibility* ← *Working memory* → *Reading (4th)*. Fifth, and related to the prior point, the estimated graphs provide support for covariate selection. Proper statistical control requires knowledge of the causal structure, as the latter dictates which variables should be conditioned on (e.g., confounders) and which variables should not be conditioned on (e.g., colliders) in order to obtain unbiased estimates (e.g., Hernán et al., 2002). More generally, knowledge of the causal structure is required to define a variable as a "confounder", a "mediator", an "instrument", a "collider", and so forth.

Finally, the estimated causal structure can support claims regarding the explanatory relevance of particular variables. I have argued that a principled and effective strategy to analyze complex organism-environment interactions is to identify the immediate causes or proximal mechanisms of the variable of interest. Even if the results of this study do not prove that particular variables are direct causes of academic achievement (due to the strong assumptions required,) they

do indicate that some variables are *not* directly related to academic achievement. This approach makes clear, then, what variables we should focus on and what variables we can safely ignore. Causal graphs encode these assumptions in an explicit manner, which is a key requirement for transparency and testability (Pear,l 2019). In other words, the strategy adopted, but more generally the concepts and methods discussed in this paper, can provide principled methods to guide the choice of the particular factors we decide to examine.

5.0 Study 3: Thinking within-persons: Using unit fixed-effects models to describe the causal relationship between executive functions and academic achievement

Psychological research is mainly concerned with causal questions, and as a consequence can benefit from explicitly adopting a causal inference framework. In this paper, I review some basic concepts in the causal inference literature related to the definition, identification and estimation of causal effects. I also explain that the main contribution of longitudinal analysis, from a causal perspective, is the ability to control for time-invariant unobserved heterogeneity, which can be achieved by focusing exclusively on within-person variation. I review different estimation procedures to estimate within-person effects, as well as different modelling strategies that can be used to test substantive hypotheses regarding within-person asymmetrical causation, moderation, effect heterogeneity, and reciprocal causation. I provide an empirical illustration of these methods by estimating the within-person effect of executive functions on academic achievement.

5.1 Introduction

There has been a dramatic increase in the number and sophistication of longitudinal datasets and models in social and behavioral research (e.g., Castellano & Ho, 2013; Falkenström et al., 2017; Usami, Murayama & Hamaker, 2019). The wide variety of modelling techniques available can expand the toolkit of researchers using longitudinal data, but can also give rise to misconceptions, misspecifications, or misinterpretations (Shanley, 2016). In the psychological literature, there has been a recent emphasis on the importance of disaggregating within and between-person variability (e.g., Curran & Bauer, 2011; Hamaker, 2012; Hoffman & Stawski, 2009), and on the difference between so-called random and fixed-effects models (e.g., Baird & Maxwell, 2016; Hamaker & Muthén, 2019; McNeish & Kelley, 2019). Unit fixed-effects regression models have long been the preferred method for analyzing panel data in econometrics and other social sciences (e.g., Allison, 2009; Angrist and Pischke, 2008), as they allow to control for time-invariant confounders. However, as several authors have noted (e.g., Usami et al., 2019; McNeish et al., 2019; Zyphur et al., 2019) studies in psychology often fail to capitalize on the opportunities that panel data offer, in particular regarding the ability to control for stable confounders.

This paper reviews the importance of disaggregating within and between-person variation and using unit fixed-effects models in psychological research. The paper also reviews some modelling strategies that can be used to test several hypotheses regarding the data-generating process at the within-person level. In particular, it considers modelling strategies that can be used to shed light on within-person processes related to asymmetrical causation, moderation, effect heterogeneity and reciprocal causation. These processes are often investigated by psychological

researchers, but require special considerations when one intends to clearly distinguish within and between-person variation.

The paper presents an empirical application of the modelling strategies discussed using a nationally representative dataset, and explores the hypothesized causal effect of children's executive functions (EF) on their academic achievement. Numerous studies have investigated this relationship with contradictory results (see, e.g., the meta-analyses by Follmer, 2018; Jacob & Parkinson 2015, and Smithers et al. 2018). A variety of modelling strategies have been implemented to describe this relationship (e.g., multiple regression; cross-lagged panel models; latent growth curve models; latent trait-state models and unit fixed-effects models), and researchers have pointed to the difficulties of drawing causal inferences from these findings (Jacob et al., 2015; Smithers et al., 2018). The literature shows, then, a lack of clarity regarding what modelling strategies should be preferred, and the extent to which the estimated coefficients can be interpreted as "causal." This confusion emanates primarily from the difficulty of judging the credibility of the assumptions implied by different methodological strategies. The objective of this paper is to clarify some of these assumptions, while discussing the possibilities and limitations of describing causal mechanisms using observational data.

The paper is structured as follows. First, I discuss the advantages of focusing on within-person variation. In light of this discussion, I review some key concepts in the causal inference literature regarding causal effect definition and identification, as well as different estimation procedures that can be used to isolate within-person variation. Second, I review several modelling strategies that can shed light into different aspects of the data-generating process at the within-person level. Finally, I provide an empirical application of the modelling strategies presented using a nationally representative dataset.

5.2 Basic concepts in causal inference

5.2.1 Causal inference and the purposes of statistical modelling

Statistical models can be used for answering different questions and, depending on the purpose at hand, the model might be subject to different standards of success. Similar to discussions regarding the validity of test scores (Kane, 2013), one can generally say that (1) the interpretation or use of a model (rather than the model itself) can be said to be "valid" or "invalid" (i.e., plausible and implausible, or appropriate and inappropriate); (2) more ambitious interpretations typically require more supporting assumptions than less-ambitious interpretations; and (3) more ambitious interpretations tend to be more useful than less-ambitious interpretations. The idea that the plausibility or appropriateness of a statistical model depends on its intended interpretation or use implies that clarifying the purpose of a model is a necessary condition for judging the adequacy of the modelling undertaking.

In statistical analyses, a key distinction is often made between descriptive, predictive and explanatory (or causal) purposes (Shmueli, 2010). These distinctions correspond to substantially different interpretations (or claims) that have different standards of success, uses and supporting assumptions. The goal of a descriptive model is to characterize or synthesize the observed patterns and relationships in the data in a parsimonious or informative manner. An example of a descriptive claim is that the black-white reading gaps grow during the school years (e.g., Reardon & Galindo, 2009). This claim synthesizes the observed patterns in the data, without any attempt to explain why these patterns emerge or make any future predictions. The mean-based model implemented is based on some assumptions regarding the underlying test scores (e.g., that they have interval properties), and its standard of success is related to the descriptive adequacy and usefulness of the

mean of the distribution for the purposes at hand (see Quintana & Correnti [2020] for a discussion on these assumptions).

The goal of a predictive model is to maximize the predictive accuracy of a target variable using some measured variables. Predictive models are often used in natural language processing and automated scoring systems, which focus on generating the most accurate predictions of an achievement score compared to the scores provided by human raters (e.g., Quinlan, Higgins & Wolff, 2009). Apart from these applications, predictive models are infrequently used in psychology and education. Yet as Shmueli (2010) explains, predictive models can serve multiple scientific purposes, even if high predictive power can be achieved with minimal or no reliance on an underlying causal theory.

Finally, the purpose of causal or explanatory models is to support claims regarding the processes in the world that generated the data, i.e., they intend to shed some light into how the world works. Thus, a causal model succeeds when it adequately represents (or "mirrors") the system under investigation, which implies that it should be able to predict how the system would behave if it were manipulated in some specific fashion (Spirtes and Zhang, 2016). Causal claims are generally considered more "useful" than descriptive or predictive claims, as they enable us to understand at a deep level the phenomenon under investigation and predict the effect of interventions. However, the explanatory power of causal explanations comes with a price; contrary to descriptive or predictive claims, the success or adequacy of a causal claim cannot be assessed using only the observed data, as it requires assumptions regarding the data-generating process that cannot be directly observed.

It is clear that causal inference is the driving force of most empirical research in psychology. A central goal in this field is to explain the causal mechanisms underlying human

cognitive and emotional processes, often with the purpose of informing potential interventions or policy changes. In the case of student achievement, for example, researchers are typically interested in whether and to what degree some factor *X* (say executive functions, motivation or family background) "affects" or "influences" student achievement. Clearly, the goal in these investigations is not to describe the data in a novel fashion nor maximize predictive accuracy, but to uncover the causal mechanisms that operate in the world. Even if researchers frequently avoid causal terminology, it is worth stressing that any claim about how the world works (or how it would behave under a potential intervention) necessarily encodes some causal information. Acknowledging this fact can have important consequences for the transparency and testability of empirical research in psychology. As previously stated, one can only assess the plausibility or appropriateness of a modelling endeavor if the intended use and interpretation is clearly stated; ambiguity in the inferences and goals (e.g., implying causal relationships without using causal language) can only confuse and mislead what the research is really about.

In sum, empirical investigations should clearly specify what the ultimate goal of the analysis is, as it is only in light of an intended use and interpretation that one can judge the success or adequacy of a particular claim. If the goal is to describe the underlying mechanisms (as it is often the case in psychological studies), researchers can benefit from explicitly adopting a causal framework –even when dealing with observational data, where causal inferences are rarely warranted. Incorporating a causal inference approach can help define the causal effect of interest and clarify the assumptions required to identify such effect from the observed data. Stating these assumptions in a clear and transparent fashion allows researchers to discuss the credibility of these assumptions; explore different modelling strategies that might relax some of these assumptions; and design future studies that can help test these assumptions. In the following section, I review

some key ideas in the causal inference literature regarding the definition, identification and estimation of causal effects.

5.2.2 Defining causal effects

In the causal inference literature there is a clear distinction between defining and estimating causal effects. Defining the target causal effect means that we clearly specify the causal quantity or information that we want to learn. According to counterfactual theories of causation, causal claims are essentially comparative, as they describe the difference in some outcome when a particular mechanism or input is altered in a specific fashion. In other words, the goal is to know the value of the outcome when the mechanism is manipulated and when the mechanism is not manipulated. This intuition is formalized in the Neyman-Rubin model (Rubin, 1974), which defines causal effects in terms of potential outcomes in hypothetical experiments.

Assuming a dichotomous treatment, the individual-level causal effect is defined in the Neyman-Rubin model as the difference between the potential outcome of an individual *i* under treatment (y_i^1) and the potential outcome for the same individual under control (y_i^0). Given that typically we cannot observe the same individual in two different states (treatment and control), it is generally not possible to calculate individual-level causal effects. Consequently, researchers focus instead on aggregate causal effects by estimating, for example, the average treatment effect (ATE). In order to estimate these aggregate effects, however, researchers need to impose certain assumptions. A fundamental assumption is that individuals' assignment to treatment or control is independent of their potential outcomes. If this is the case, then the assignment mechanism is said to be "ignorable." The independence of the potential outcomes with the treatment indicator (T) can be formally represented as $(Y_i^0, Y_i^1) \perp\!\!\!\perp T$, and the ATE is defined as $E[\delta] = E[Y^1] - E[Y^0]$.

Randomized control trials (RCTs) are often considered the gold standard for causal inference, as the assignment mechanism is forced to be independent of the potential outcomes (assuming that the RCT is properly designed and implemented). As it is frequently noted, however, in many cases it is impractical or unethical to perform RCTs. Due to these difficulties, researchers often rely on observational data to estimate causal effects. Given the lack of control over the assignment mechanism, it is likely that individuals' potential outcomes differ across conditions, and that ignorability does not hold. The most common concern in this scenario is that the causal variable *X* and the outcome *Y* are both caused by a third variable *Z*. In this case, the total association between *X* and *Y* would be composed of (1) the causal effect of *X* on *Y*, and (2) the common dependence of *X* and *Y* on *Z* (Morgan & Winship, 2015). In order to differentiate the genuine causal effect from other spurious associations, researchers adopt an "identification strategy", which specifies how causal effects are going to be estimated given the available data. Below, I briefly review the basic concepts of graphical causal models, which provide a clear framework and simple criteria for determining the identifiability of causal effects from observational data.

5.2.3 Causal identification

After the target effect has been defined, the next step is to determine whether the effect of interest can be learned (i.e., "identified") from the data. The approach adopted to learn the target causal effect is commonly referred to, then, as the "identification strategy." In the case of an RCT, the identification strategy might simply consist in subtracting the control and treatment differences in the outcome. Given that in this scenario the treatment and control groups are comparable across all dimensions except for treatment status, the ATE represents an unbiased estimate of the treatment in the population from which the sample was drawn (Murnane & Willett, 2010).

However, in observational studies the lack of control over the assignment mechanism implies that the treatment and control groups might not be comparable at baseline. This implies that the difference in outcomes might not represent an unbiased estimate of the treatment effect, as it combines any potential causal effect with the effect of preexisting differences. In observational studies, the identification strategy consists then in specifying how the causal associations will be distinguished from the non-causal associations.

Using causal graphs, Pearl (2009) provides a clear test called the "back-door criterion" that can be used to determine the identifiability of causal effects using observational data[20]. The objective of this strategy is to isolate the causal effect of some variable *X* on an outcome *Y* by conditioning on a set of variables *Z*. The set *Z* is referred to as the "deconfounding set" and, if all the non-causal associations are effectively removed by conditioning on *Z*, then *X* would be considered conditionally ignorable given *Z* (Pearl, 2009). This strategy is at the core of common conditioning estimation procedures such as regression or matching.

The back-door criterion specifies which variables should be included in the conditioning set in order to identify the target causal effect. By doing so, this strategy also specifies the set of variables that need *not* be included in the conditioning set. Distinguishing between these two sets is important, as adjusting for the wrong variables can generate bias in the estimated coefficients (e.g., it might conceal a true effect or create a spurious effect). In order to understand why conditioning can be both a bias-removing and a bias-amplifying mechanism, it is useful to consider how different causal structures can generate associations between variables. Figure 8 illustrates

[20] Other strategies for identifying causal effects using observational data are instrumental variables and the front-door criterion (e.g., Pearl, 2009; Morgan & Winship, 2015).

three basic causal structures that can generate non-causal associations between two variables *X* and *Y*. Panel A shows that confounding bias arises when the variables *X* and *Y* have a direct common cause *Z* that is not controlled for. Panel B shows that, in order to identify the causal effect of *X* on *Y* in this scenario, one needs to condition on the common cause *Z*. Panel C shows that selection (or "collider") bias arises when one conditions on a common effect *Z*. Panel D indicates that, in order to remove this bias, one needs to *not* condition on *Z*. Panel E shows that overcontrol bias arises when the causal path between variables *X* and *Y* is blocked by conditioning on a mediator *Z*. As Panel F illustrates, the solution is again to *not* condition on Z. As this figure illustrates, then, statistical conditioning can be used to remove non-causal associations (in the case of confounding bias) but can also introduce or even amplify bias (e.g., in the case of selection or overcontrol bias; see Pearl [2011] for a discussion on other bias amplification mechanisms).

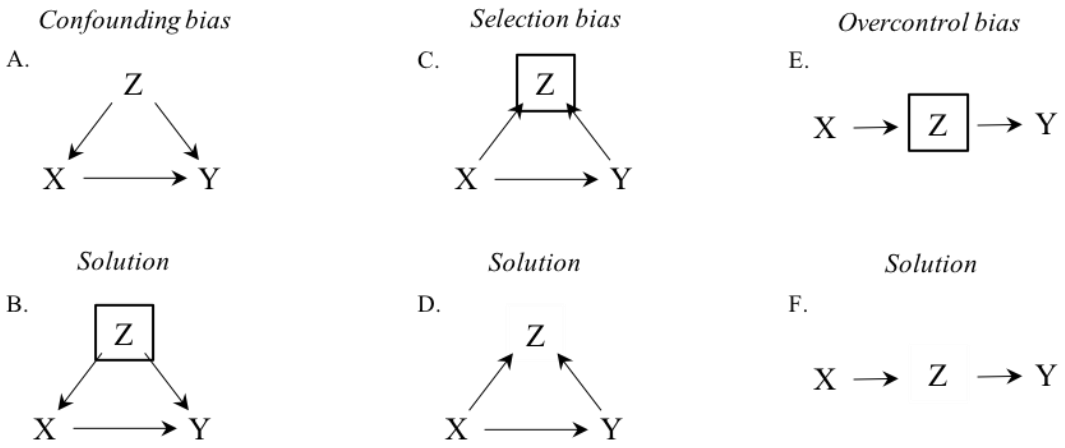


Figure 8. Three types of bias with their respective solutions. The square represents that the variable has been conditioned on.

Given a particular causal structure (which might include a complex combination of the different sub-structures depicted in Figure 8), the back-door criterion states that the effect of *X* on *Y* is identified if one adjusts for all the direct common causes of *X* and *Y*, i.e., all the confounding variables (or the variables that start with an arrow into *X* –thus the name "backdoor path"). Controlling for these variables "blocks" all non-causal paths from *X* to *Y*, and allows us to isolate the causal relationship between these variables.

The previous discussion makes two important points. First, researchers should only adjust for a variable if they believe it may be a confounder. In order to avoid the risks of collider and overcontrol bias, researchers should make special attention of not conditioning on variables that are affected by the treatment or the outcome variables (Hernan et al., 2002). Second, the back-door criterion (as well as other identification strategies such as instrumental variables) requires prior assumptions about how the variables are causally related to each other. One can only describe a variable as a "confounder", an "instrument" a "mediator" and so forth in the context of a previously specified causal structure. In other words, causal inference requires prior causal assumptions about how the data was generated. In RCTs, the researcher has control over the data-generating mechanism, and as consequence can discard the possibility of any confounding biases (Steiner et al., 2017). In observational studies, on the other hand, the researcher needs to assume an underlying causal model that adequately represents the data generating process, which implies that the causal relationships among the variables in the system (including the unobserved variables), as well as the data collection mechanism, is adequately specified. Causal graphs are effective tools for encoding the causal assumptions regarding the presumed data generating process of the observed data, and are therefore valuable tools for identification analyses (Pearl, 2009).

5.2.4 Using longitudinal data for causal identification

Identification using conditioning methods (e.g., matching or regression) relies on the unconfoundedness assumption, which requires that researchers know all the confounding variables and measure them reliably (Kim & Steiner, 2019a). In observational studies, however, one can never be sure that all confounding variables have been considered. In fact, it is almost certain that important confounders have been omitted (Berk, 2004). Consequently, the use of conditioning methods using observational data typically generate biased effect estimates. These spurious results might be misleading in regards to the presence or absence of a causal relationship, the magnitude of these relationships, or the sign of the true causal effect. The bias resulting from unblocked confounding paths has been referred to as omitted-variable bias and the independent variable in question is said to be "endogenous."

If relevant variables are not measured, then the causal effect of interest might not be identified (Pearl, 2009). For example, if variable *Z* in Panel A of Figure 8 is not measured, then the causal effect of *X* on *Y* is not identifiable with the given data. This example illustrates how the problem of causal inference is fundamentally one of unobservables; and, as Halaby (2004, p.508) remarks, "unobservables are at the heart of the contribution of panel data to solving problems of causal inference." In order to understand how longitudinal data can be used to identify causal effects, it is important to distinguish between two types of unobservables: (1) unit-specific unobservables representing stable properties of individuals (so-called "unit effects") and (2) time-varying unit-specific unobservables representing transitory and idiosyncratic alterations or disturbances (Halaby, 2004). As Halaby (2004) explains, panel data offers certain advantages for dealing with unobservables (in particular unobservables of the first kind), but only through particular statistical methods that capitalize on the longitudinal structure of the data.

It is widely-known that, if the data under consideration is longitudinal, unit fixed-effects (FE) models can alleviate the effects of confounding variables (e.g., Allison, 2009). In particular, FE models can be used to adjust for unobserved time-invariant confounders when estimating causal effects from observational data. The basic idea behind FE models is to use each individual as his or her own control, which removes confounding bias from all the observed, unobserved, and mismeasured time-invariant individual and group-level characteristics. These models require treatment variation within units over time, and eliminate the spurious associations due to omitted variables as long as these confounding paths remain stable over time (Kim & Steiner, 2019b). In other words, the key insight behind FE models is that we can identify causal effects by performing analyses within individuals, and that ignorability might hold conditional on a time-invariant unit-specific effect. This conditional ignorability assumption can be stated as follows:

$$(Y_i^0, Y_i^1) \perp\!\!\!\perp T_{it} \mid \boldsymbol{X}_{it}, U_i \qquad (1)$$

where $\boldsymbol{X}_{it}$ represents observed time-varying covariates and U_i a stable unit effect. In other words, FE models replaces the unrealistic assumption that we can block all backdoor paths by conditioning on all confounders, by the less restrictive assumption that the effects of all confounding variables are stable over time (Firebaugh, Warner & Massoglia, 2013).

It is worth stressing that FE models do not guarantee that the causal effect of interest is identified, but allows us to rely on less restrictive identifying assumptions. Importantly, given that FE models assume that the effect of unmeasured confounders is constant, these models do not remove the bias generated by unmeasured time-varying confounders. Furthermore, traditional FE

models impose additional assumptions (e.g., the absence of reciprocal causation and state effects; see Imai & Kim, 2019), some of which I discuss below.

5.2.5 Estimating FE models

FE models can be estimated using different methods (see, e.g., Falkenström et al., 2017; Hamaker & Muthén, 2019; McNeish & Kelley, 2019; Usami et al., 2019). The idea behind these estimation procedures is to consider only the within-person variance of the variables in the model and discard all the between-person variation. As equation 1 indicates, the key identifying assumption of FE models is that causal effects are identified within-individuals. By focusing exclusively on within-individual variation, one adjusts for all stable confounders. I will briefly review the basic idea behind four common methods for estimating FE models (see McNeish & Kelley [2019] and Hamaker & Muthén [2019] for more comprehensive and detailed discussion on these and other estimation procedures). However, it is worth stressing that the goal of all the methods is the same, namely to remove all between-person stable differences in the variables considered.

One method for disaggregating within- and between-person variation consists in including a dummy variable for each individual in the sample (the so-called Least Squares Dummy Variable (LSDV) regression model). The LSDV has the following form:

$$y_{it} = \beta_1 x_{it} + \sum_{i=1}^{i} \gamma_i D_i + e_{it} \qquad (2)$$

where x_{it} represents the time-varying predictor of interest and D_i represents a dummy variable for each individual *i*. These dummy variables represent the combined effect of all the variables that have a time-invariant effect on the outcome.

Another estimation procedure is referred to as "demeaning", and consists in subtracting the individual's mean value across time from all variables (also known as person-mean-centering). The estimation equation is the following:

$$y_{it} - \bar{y}_i = \beta_1(x_{it} - \bar{x}_i) + (\varepsilon_{it} - \bar{\varepsilon}_i) \quad (3)$$

where $\bar{y}_i$, $\bar{x}_i$ and $\bar{\varepsilon}_i$ refer to the individual-specific mean of the respective variable across time. This equation isolates the within-person variation by considering the individual's deviation from their own mean. By subtracting the individual's mean, one also removes all time-invariant confounders.

It is often noted that a disadvantage of traditional FE models as estimated by equations 2 and 3 is that the effects of observed time-invariant covariates (e.g., gender or race) cannot be estimated, as only within-person variation is considered. Researchers have solved this problem by implementing a random-effects model with a demeaned predictor (as in equation 3,) and the individual's mean of the time-varying predictor (see Bell & Jones, 2015). This "hybrid random effects" model allows simultaneous estimation of the within and between-person effects through these two terms, as well as estimating the effect of observed time-invariant covariates. The model can be expressed as follows:

$$y_{it} = \beta_1(x_{it} - \bar{x}_i) + \beta_2(\bar{x}_i) + \zeta_i + \varepsilon_{it} \quad (4)$$

where β_1 represents the within-person effect, $\beta_2$the between-person effect, and ζ_i represents a random intercept.

Finally, FE models can be estimated in a SEM framework by including a person-specific latent variable with the different values of the outcome variable as indicators (Allison, 2009). This model can be expressed as follows:

$$y_{it} = \beta_1 x_{it} + \lambda_t \eta_i + e_{it} \quad (5)$$

where η_i represents the unit-specific factor and λ_t the factor loadings (which are typically constrained to be 1.0 but can be allowed to vary, e.g., Usami et al., 2019). Importantly, the person-specific factor is allowed to covary with the values of the time-varying predictor x_{it} (see Bollen & Brand, 2010). In this way, the latent factor subtracts all the between-person differences, leaving only the within-person effect of *X* on *Y*.

The four estimation procedures presented effectively disaggregate within and between-person variation, and as a consequence the parameter of interest (β_1) will be identical across specifications (e.g., Hamaker & Muthén, 2019). Typically, however, the results will be different from other estimation procedures (e.g., pooled OLS or random-effect models) that do not disaggregate within and between-person variation. The latter generate results that are difficult to interpret from a causal perspective, as they blend processes that occur at different levels.

5.3 Describing causal mechanisms at the within-person level

5.3.1 The advantages of thinking "within-persons"

We have seen that one advantage of distinguishing between-person from within-person relationships is that it allows us to control for all time-invariant confounders. It is worth noting that this is a *causal* reason, as the notion of a "confounder" can be understood only by reference to a causal structure (i.e., as a direct common cause of two variables). In addition, it is worth noting that by separating these two levels of analysis one assumes that the causal mechanisms one is interested in can be described within individuals. This is not a trivial assumption, as some mechanisms might operate between-individuals (i.e., they might generate relatively stable differences among individuals). In a FE framework, between-person differences may moderate within-person effects, but the processes of change are assumed to be situated within individuals.

Researchers in the social and behavioral sciences have long recognized the importance of distinguishing between-person from within-person relationships, arguing that group-derived estimates do not necessarily represent individual-level phenomena (e.g., Molenaar, 2004). Applying the findings obtained from aggregate data to individuals constitutes an error of inference normally referred to as the "ecological fallacy" (Curran & Bauer, 2011). In psychology, this problem has attracted growing attention in recent years, as researchers have shown that the estimates derived from models that focus on the variation between subjects rarely apply to individual subjects (Molenaar, 2004; Fisher, Megdalia & Jeronimus, 2018; Hamaker, 2012). Several examples provided in the literature intuitively demonstrate the lack of group-to-individual generalizability. For instance, Curran and Bauer (2011) note that, at the group level, there is a negative correlation between exercise and heart attack –as the people who exercise more tend to

have better health, and are therefore less likely to have heart problems. However, within individuals the correlation is positive, as individuals are more likely to experience a heart attack while exercising, or when they exercise more than usual. This reversal of effects (commonly referred to as Simpson's paradox) occurs because, at the group level, the causal effect of exercise on heart attack is confounded by person-level characteristics (e.g., age, nutrition habits, etc.). FE models deal with this problem by removing all person-level confounders and focusing exclusively on within-person change.

In sum, FE models can contribute to solving problems of causal inference by removing the effects of time-invariant confounders. However, apart from estimating the causal effect of a predictor on an outcome variable, researchers in psychology are often interested in describing the underlying data-generating process at a more fine-grained level. In particular, researchers are often interested in testing substantive hypotheses regarding asymmetrical causation, moderation, effect heterogeneity, and reciprocal causation. In the following sections, I describe ways of investigating these issues in the context of within-person processes.

5.3.2 Within-person asymmetrical causation

A common and often unrecognized assumption behind the modelling strategies employed by social and psychological researchers is that causation is symmetrical; that is, that the effect of the independent variable on the dependent variable has the same magnitude regardless of whether the independent variable is increasing or decreasing (York & Light, 2017). This wide-spread assumption was put into question by Lieberson (1987), who argued that many social phenomena might involve asymmetrical causation. A similar intuition is present in the literature on risk and resilience (e.g., Gutman, Sameroff & Cole, 2003), which commonly distinguishes between factors

with predominant upward effects (promotive factors) from factors with predominant downward effects (risk factors). Furthermore, this literature also distinguishes between factors associated to predominantly positive interactive effects (protective factors) from factors associated to predominantly negative interactive effects (vulnerability factors).

A causal relationship is symmetrical if equivalent positive or negative changes in the cause are associated with equivalent positive or negative changes in the outcome, respectively[21]. Many social and psychological phenomena might involve asymmetrical forms of causation. For example, the negative consequences of stressful life events (e.g., child abuse, maternal depression, etc.) might be stronger than the positive effects associated to the reduction of these stressors (Hanson et al., 2012). This example also illustrates how the distinction between symmetrical and asymmetrical causation relates to the reversibility or irreversibility of events (Lieberson, 1987).

Even though asymmetrical forms of causation and related ideas (e.g., the notions of promotive and risk factors) have been extensively used and recognized, there have not been clear methodological tools for examining these phenomena. Based on a solution put forward by York and Light (2017), Allison (2019) provides a simple method for estimating asymmetrical effects using FE models. The main idea behind this approach is to estimate different coefficients for the positive and negative changes in the predictor. More precisely, given a panel data where individual *i* is observed at time *t*, Allison defines the positive (X_{it}^{+}) and negative (X_{it}^{-}) changes of a predictor X_{it} as

[21] It is worth noting that this distinction is different from the linearity assumption, which implies that the changes in the outcome are constant across different levels of the predictor

$$X_{it}^{+} = X_{it} - X_{it-1} \text{ if } (X_{it} - X_{it-1}) > 0, \text{ 0 otherwise,}$$

$$X_{it}^{-} = -(X_{it} - X_{it-1}) \text{ if } (X_{it} - X_{it-1}) < 0, \text{ 0 otherwise.} \quad (6)$$

The terms X_{it}^{+} and X_{it}^{-} represent, then, the successive positive and negative changes in X, respectively. Typically, the coefficients associated to these terms will have opposite sign; thus, in order to test the equality of the coefficients, one needs to multiply one of them (in this case X_{it}^{-}) by –1.

These terms can then be used to create cumulative positive and negative changes in X:

$$Z_{it}^{+} = \sum_{s=1}^{t} X_{is}^{+}$$

$$Z_{it}^{-} = \sum_{s=1}^{t} X_{is}^{-} \quad (7)$$

These expressions represent the accumulation of positive and negative changes in the predictor *X* until time *t*. These terms can then be included in a traditional fixed-effects as follows:

$$Y_{it} = \beta^{+} Z_{it}^{+} + \beta^{-} Z_{it}^{-} + \alpha_i + \varepsilon_{it} \quad (8)$$

As Allison notes, the cumulative terms imply that, all else being equal, changes in X in a particular direction persist over time. The coefficient β^{+}can be interpreted as the effect associated to a 1-unit increase in the predictor, and β^{-} as the effect associated to a 1-unit decrease in the predictor.

5.3.3 Within-person interactions

Social and psychological researchers are frequently interested in testing interaction hypotheses. An interaction refers to the phenomenon whereby an exposure, characteristic, or state alters the effect of a different exposure, characteristic or state –which is also referred to as "moderation" or "effect modification" (VanderWeele, 2015). Interactions constitute an important part of understanding mechanisms, as they can shed light into why a particular cause can have differential effects on the outcome based on other individual characteristics. Interaction analyses are also motivated by an increasing interest in "intersectionality theory", according to which demographic characteristics (e.g., gender, ethnicity, social class, etc.) can have multiplicative rather than additive effects (Dubrow, 2008). The idea behind this theory is that we should not conceive demographic attributes as having autonomous and separate effects, given that advantages or disadvantages emerge from the intersection of these different attributes. In order to test for intersectionality, one needs to consider, then, the cross-product of explanatory variables, e.g., or $gender \times ethnicity \times class$.

Despite their widespread importance and use, the implementation of interaction terms can present several complications. First, in order to precisely estimate interaction effects, one normally needs very large sample sizes (VanderWeele, 2015). Second, interaction terms can be difficult to interpret, especially if more than two terms are involved (as intersectionality theory recommends). Third, interaction terms are based on assumptions that are seldom examined, e.g., linearity and sufficient common support in the moderator (Hainmueller, Mummolo & Xu, 2019). In this section, however, I focus on a complication that arises in the context of within-person analyses. Given that the terms included in the interaction can have both time-varying and time-invariant components, the interaction term should be constructed in a way that only considers the desired component.

Below, I explain how to construct within-person interaction terms when only one term has a time-varying component, and when the two terms have a time-varying component.

Let us examine first the interaction between a time-invariant and a time-varying predictor. In particular, consider the interaction between the dummy variable *gender* (a time-invariant variable *G*) and a continuous measure of executive function (a time-varying variable *E*). In the context of student achievement, one can include this interaction in order to examine to what extent the within-person effect of executive function on achievement varies between genders. The common way of including this interaction is by treating the interaction term $G \times E$ as any other predictor (e.g., Schunck, 2013); that is, by including the term $(GE_{it} - \overline{GE}_i)$ in addition to the within-person effects of *G* and *E*. Given that $G = \bar{G}$, this term can be factor out of the equation:

$$y_{it} = \beta_1(E_{it} - \bar{E}_i) + \beta_2(\bar{E}_i) + \beta_3 G(E_{it} - \bar{E}_i) + \beta_3(G_1) + \zeta_i + \varepsilon_{it} \qquad (9)$$

In this equation, β_3 can be interpreted as the differences in the within-unit effects of *E* on *Y* across values of *G*. In our empirical example, this can be interpreted as within-person differences in executive function on achievement depending on gender.

Let us consider now an interaction between two continuous and time-varying predictors. Researchers have often constructed this interaction in the same way as in equation 9, i.e., by demeaning the interaction term. However, Giesselmann and Schmidt (2018) show that this strategy does not generate a within estimator of the interaction. That is, they show that the time-invariant component of two time-varying variables, say, P_{it} and E_{it}, is not eliminated when the interaction term is demeaned (i.e., by including $P_{it}E_{it} - \overline{P_{it}E_i}$), conflating within and between person variation. In order to solve this problem, Giesselmann and Schmidt (2018) recommend a "double-

demeaned" estimator, produced by first demeaning the predictors and then demeaning the product. That is, by including the following term:

$$(P_{it} - \overline{P_{it}})(E_{it} - \overline{E_{it}}) - \frac{\sum_{t=1}^{T_i}(P_{it} - \overline{P_{it}})(E_{it} - \overline{E_{it}})}{T_i} \quad (10)$$

As the authors explain, the unbiasedness of the proposed estimator comes at the cost of lower efficiency, which implies that it can generate large standard errors and require a large sample size.

5.3.4 Estimating between-person differences in within-person effects

Researchers are often interested in examining the extent to which causal effects differ across individuals, and consider variables that might explain that effect heterogeneity. A common way of exploring this issue in a FE framework is by treating the within-person coefficient as a random slope (Baird & Maxwell, 2016.; Falkenström et al., 2017). This specification can be implemented in a hybrid model as follows:

$$y_{it} = (\beta_1 + \mu_i)(E_{it} - \bar{E}_i) + \beta_2(\bar{E}_i) + \zeta_i + \varepsilon_{it} \quad (11)$$

where μ_i allows the within-person effect $\beta_1$vary across individuals, and can be used as a dependent variable at level-2.

5.3.5 Modelling within-person reciprocal causation

As explained above, the key reason why psychological researchers should conduct within-person studies is a causal reason, related to the possibility of controlling for all measured and unmeasured time-invariant confounders. I have also stressed that within-person models do not guarantee causal identification, but rather allow us to rely on less restrictive identifying assumptions. In particular, the traditional fixed-effects model that one estimates by using equations 2-5 generates unbiased estimates if the underlying data-generating process conforms to the causal graph represented by Figure 2. As Imai and Kim (2019) note, this graph makes several assumptions. Apart from the assumption of no time-varying confounders, the graph assumes the absence of causal relationships between (1) the outcome variable Y_{it} at different time-points; (2) the predictor X_{it} on future values of the outcome variable Y_{it}; and (3) the outcome variable Y_{it} on future values of the predictor X_{it}.

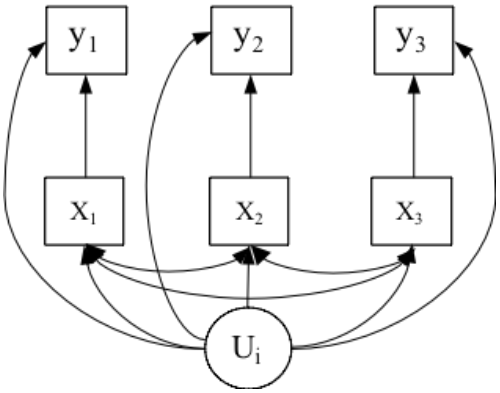


Figure 9. Causal graph representing the data-generating process implied by traditional FE models.

It is worth noting that the plausibility of these assumptions depends on the particular causal process that the graph is intended to represent. In the case of academic achievement and executive

functions, for example, assumptions 1 and 3 appear to be particularly problematic, as previous research suggests that (1) skills at one point are to some extent causally determined by the skills developed earlier (often referred to as "state effects"; e.g., Bailey et al., 2014; Stanovich, 1986); and (2) there are bidirectional associations between executive functions and academic achievement (commonly referred to as "reciprocal effects"; e.g., Fuhs et al., 2014; McKinnon & Blair, 2019; Schmitt et al., 2017). As a consequence, the absence of any causal relationships between these nodes (i.e., between $Y_{i,t} \rightarrow Y_{i,t+1}$ and $Y_{i,t} \rightarrow X_{i,t+1}$) appears to be a strong and unwarranted assumption, which might lead to biased results.

In addition to addressing potential bias in the estimated coefficients, relaxing assumptions 1 and 3 might help researchers test substantive hypotheses related to the data-generating process. That is, estimating the strength of the state and reciprocal effects between executive functions and academic achievement has theoretical importance. Finally, assumption 2 can be problematic if the lagged-effect of X (rather than the contemporaneous effect, or both the lagged and contemporaneous effects) generates the changes in Y (see Leszczensky & Wolbring, 2019; Vaisey & Miles, 2017).

Researchers have proposed different models that relax the three assumptions presented above (while separating within and between-person variation) in the framework of linear structural equation models (Allison, Williams & Moral-Benito, 2017; Hamaker & Muthén, 2019; Usami et al., 2019). These models allow us to control for (1) time-invariant heterogeneity by including a unit-specific factor; (2) state effects by including autoregressive components; and (3) reciprocal effects by allowing cross-lagged components (see Zyphur et al. [2019] and Usami et al. [2019] for a synthesizes and comparison of several of these models). I will briefly review the ML-SEM model presented by Allison et al.'s (2017) and the GCLM model presented by Zyphur et al. (2019).

Allison et al.'s (2017) ML-SEM model

The ML-SEM model presented by Allison et al.'s (2017) controls for time-invariant confounders by including a unit-fixed effect as depicted in Figure 9, as well as autoregressive effects among the predictor. Besides modelling state-dependence and unobserved heterogeneity, the ML-SEM can also account for reciprocal causation by allowing the *X* variables to be correlated with the error term of *Y* at any prior time point (Allison et al., 2017). The model estimated by ML-SEM can be represented by the following equation:

$$y_{it} = \beta_1 x_{it} + \beta_2 y_{i,t-1} + \alpha_i + \varepsilon_{it} \qquad (12)$$

where β_1 represents the contemporaneous effect of *x* on *y*, β_2 represents the autoregressive effect, and α_i represents individual fixed effects. In order to account for reciprocal causation (i.e., causal effects of *Y* on *X*), x_{it} is allowed to correlate with $\varepsilon_{i,s<t}$.

The model represented by equation 12 assumes (as all FE models presented so far) that the effect of *X* on *Y* is contemporaneous rather than lagged. However, as Vaisey and Miles (2017) indicate, the obtained estimates can be severely biased if this specification does not correspond to the true timing of causal effects. This is a major concern when dealing with observational data, where we usually lack precise theoretical knowledge regarding the temporal structure of the underlying data-generating process. In this scenario, Leszczensky and Wolbring (2019) show that including both contemporaneous and lagged effects is the best way to deal with the lag specification problem.

Zyphur et al.'s (2019) GCLM model

A similar model that accounts for time-invariant confounders as well as reciprocal causality is the "general cross-lagged panel model" (GCLM) presented by Zyphur et al. (2019). A full discussion of this model is beyond the scope of this paper; however, I will review the advantages of this model from a causal inference perspective.

The authors argue that, if one is interested in estimating the causal effect of a predictor *X* on an outcome *Y*, then one should focus on the residuals in a model (e.g., the GCLM) that explicitly accounts for all "systematic" between and within-person processes. This idea is consistent with conditional ignorability, which states that treatment assignment should be (conditionally) independent of the potential outcomes that would result if the unit is assigned to treatment or control. That is, treatment assignment should be considered (conditionally or unconditionally) random. As previously noted, however, establishing conditional ignorability with observational data is rarely, if ever, fully justified, due to the impossibility of identify and measuring all confounding variables. Yet the GCLM goes a long way in controlling for a range of potential confounding differences among units. The baseline model is specified as follows:

$$y_{it} = \alpha_t + \lambda_t \eta_i + \beta_1 y_{it-1} + \beta_2 x_{it-1} + u_{it} \qquad (12)$$

where α_t represents occasion-specific effects (which can be included in the model or accounted for by previously detrending the data); $\lambda_t \eta_i$ represents a constant or time-varying unit effect; $\beta_1 y_{it-1}$ represents an autoregressive effect; and $\beta_2 x_{it-1}$ represents a cross-lagged effect. The residual term u_{it} represents an "impulse" or "shock" that cannot be explained by any of the previous terms. An equivalent equation is specified for x_{it} (the predictor of interest) as the outcome

variable and, importantly, the residual terms of the two equations are allowed to covary. By including these covariances (referred to as "co-movements" by Zyphur et al., 2019) the model explicitly considers any unpredictable but simultaneous (and, as a consequence, "systematic") within-person changes in both *X* and *Y*.

The residual term u_{it} represents, then, idiosyncratic within-person variation that is uncorrelated with several systematic between and within-person changes. To recapitulate, variation in u_{it} is not affected by time-invariant confounders; state and reciprocal effects; general time trends; or person-specific co-movements. Given that u_{it} isolates idiosyncratic (or un-systematic) variation, it can be used to approximate "random" assignment. The authors take advantage of this by including in equation 12 previous values of the residual:

$$y_{it} = \alpha_t + \lambda_t \eta_i + \beta_1 y_{it-1} + \beta_2 x_{it-1} + \partial_1 u^y_{it-1} + \partial_2 u^x_{it-1} + u_{it} \qquad (13)$$

where $\partial_1 u^y_{it-1}$represents the effect on *Y* of a previous impulse of *Y*, and $\partial_2 u^x_{it-1}$represents the effect on *Y* of a previous impulse of *X*. A similar equation is implemented with *X* as the outcome variable.

Note that, contrary to the traditional FE models presented above, the GCLM estimates the lagged effect of the predictor on the variable. As noted, misspecification regarding the temporal structure of the underlying data-generating process can generate biased estimates. From a causal inference perspective, however, this lag specification might be preferred (compared to the contemporaneous effect), as the cause needs to precede the effect.

5.4 Data

The data comes from the Early Childhood Longitudinal Study (ECLS-K:2010), conducted by the National Center for Education Statistics (see Tourangeau et al. 2019, for more information regarding this study). The study tracks a nationally representative sample of 18,170 U.S. children who entered kindergarten in the 2010–2011 school year through fifth grade. Sampling weights are provided in the data set in order to account for differential selection at each sampling stage and to adjust for the effects of nonresponse (Tourangeau et al. 2019). In this study, the analytic sample was defined as 7,956 individuals that had a valid sampling weight that maximized the number of sources included in the analysis (which involved child and teacher data from multiple waves).

5.4.1 Measures

Reading and math achievement scores. Children completed individualized cognitive assessments in reading and math that were developed for the ECLS (Tourangeau et al. 2019). The items on the assessments combined questions from well-validated and reliable tests, as well as newly developed items. The direct cognitive battery used adaptive testing, based on a three-parameter item response theory (IRT) model, in order to create a common scale and minimize the possibility of ceiling and floor effects (see Najarian et al. 2018, for a detailed psychometric report). I included the estimated ability score (theta) in reading and math in both kindergarten and fourth grade.

Given that the IRT scores change systematically over time and that uncontrolled time-trends in the data can generate spurious correlations (e.g., Falkenström et al., 2017), prior

detrending of the achievement scores was performed by centering the variables by the weighted mean of each measurement occasion.

Executive function. Children completed direct assessments of cognitive flexibility using the *Dimensional Change Card Sort* (Zelazo 2006), and working memory using the *Numbers Reversed task of the Woodcock-Johnson III (WJ III) Tests of Cognitive Abilities* (Woodcock McGrew, and Mather, 2001). Two additional measures of executive function included teacher-reported attentional focusing and inhibitory control, based on 6 and 7 items from the *Temperament in Middle Childhood Questionnaire* (Simonds & Rothbart, 2004), respectively.

Researchers disagree on what are the most adequate operational definitions and measurement approaches of executive functions (e.g., Jacob & Parkinson, 2015). Thus, in the present study I considered both performance-based measures (related to cognitive flexibility and working memory) and teacher-reported measures (related to attentional focusing and inhibitory control). I included teacher-reported rather than parent-reported ratings, as previous research suggests that the former are more highly predictive of academic outcomes (e.g., Miranda et al., 2015).

A factor that controls for measurement error is desirable when we focus on within-person variation, as in this case the proportion of measurement error increases. Consequently, a factor representing executive function was extracted using the two performance-based measures and the two teacher-reported measures.

Measures of academic achievement and executive functions were included in 7 waves of data collection: fall and spring of kindergarten; and spring of first, second, third, fourth and fifth grade. Data from all 7 measurement occasions were included in the analysis.

The intraclass correlation for EF was 0.66, indicating that 34% of the observed variance in executive functions is within-person variance, while 66% can be considered between-person variance. Similarly, the intraclass correlation of reading and math achievement were 0.68 and 0.81, respectively. This indicates that 32% of the variance in reading achievement and 19% of the variance in math achievement is within-person variance.

5.5 Results

5.5.1 Within and between-person effects

A hybrid model as represented in equation 4 was implemented to estimate the contemporaneous within and between-person effects of executive functions on reading and math achievement. The estimated within-person effect of executive functions on reading and math achievement was 0.060 (SE = 0.004, $p < 0.001$) and 0.049 (SE = 0.003, $p < 0.001$) IRT units respectively (see Table 7). In order to interpret the relative magnitude of within-person effects, researchers recommend standardizing them using the within-person standard deviation (Schuurman, Ferrer, de Boer-Sonnenschein, & Hamaker, 2016), which was estimated as 0.28 for reading and 0.24 for math. The within-person standardized coefficients are then 0.21 *SD* for reading and 0.20 *SD* for math. These results suggest that executive functions have a positive and significant effect on academic achievement, and that that the effect on reading is comparable to the effect on math.

Table 7. Estimated within-person effects of executive functions on reading and math achievement using eight different models.

	Outcome variable	
	Reading	Math
Hybrid model		
Executive function, within	0.060	0.049
	(.004)	(.003)
Executive function, between	0.331	0.403
	(.006)	(.007)
Asymmetric model		
Executive function, within positive	0.057	0.048
	(.005)	(.004)
Executive function, within negative	-0.052	-0.044
	(.004)	(.003)
Interaction model (one time-varying)		
Executive function, within	0.057	0.047
	(.005)	(.004)
Executive function × Gender, within	0.007	0.004
	(.008)	(.006)
Interaction model (two time-varying)		
Executive function, *within*	0.061	0.048
	(.004)	(.003)
Self-control, within	-0.007	-0.001
	(.014)	(.011)
Executive function × Self-control, within	-0.080	-0.020
	(.023)	(.018)
Random slopes model		
Executive function, within fixed effect	0.057	0.046
	(.004)	(.003)
Executive function, within random effect	0.362	0.295
	(.013)	(.013)
ML-SEM model (contemporaneous effect)		
Executive function, within	0.019	0.019
	(.005)	(.004)
ML-SEM model (contemporaneous and lagged effect)		
Executive function, within contemporaneous	0.068	0.029
	(0.010)	(.007)
Executive function, within lagged	0.035	0.011
	(0.006)	(.005)
General cross-lagged panel model (lagged effect)		
Executive function, within	0.001	0.027
	(.006)	(.006)

Notes. An appropriate longitudinal weight was used in all calculations.

Table 7 also displays the between-person effects of executive functions on reading and math achievement. One can perceive that these effects are considerably larger than the within-person estimates: β = 0.331 (*SE* = 0.006, $p < 0.001$) for reading and β = 0.403 (*SE* = 0.007, $p < 0.001$) for math. The standardized coefficients using the between-person standard deviation (0.41 for reading and 0.49 for math), is 0.81 *SD* for reading and 0.82 *SD* for math. The fact that the standardized between-person effect is more than 3 times larger than the standardized within-person effect reinforces the importance of clearly distinguishing between these levels of analysis.

5.5.2 Within-person asymmetric effects

I estimated an asymmetric fixed-effects model as represented by equation 8. Interpreted causally, the results indicate that one positive-unit change in executive functions generates a 0.057 (*SE* = 0.005, $p < 0.001$) change in reading achievement; while one negative-unit change in executive functions generates a –0.052 (*SE* = 0.004, $p < 0.001$) change in reading achievement. However, the difference between these coefficients is not statistically significant, $F(1, 7953) = 1.28$. A similar result was found in the case of math achievement, where one positive-unit change in executive functions generates a 0.048 (*SE* = 0.004, $p < 0.001$) change in math achievement; and one negative-unit change in executive functions generates a –0.044 (*SE* = 0.004, $p < 0.001$) change in math achievement. Similarly, the difference between the coefficients is not statistically significant, $F(1, 7951) = 1.94$.

5.5.3 Within-person interactions

I estimated the interaction effect of executive functions and gender on reading and math achievement following equation 9. As Table 7 indicates, the interaction term for reading (β = 0.007, *SE* = 0.008, p = 0.373) is similar to the interaction term for math (β = 0.004, *SE* = 0.006, p = 0.526), and in neither case it is statistically significant. One can interpret these coefficients as implying that the within-person effect of executive functions on achievement appears to be slightly higher for females than for males, but the difference is not statistically significant.

In order to illustrate the interaction between two time-varying predictors, I constructed an interaction term between executive functions and the detrended score of students' teacher-reported self-control, following equation 10. The intraclass correlation for self-control was 0.46, indicating that 54% of the observed variance is within-person (see Tourangeau et al. 2019 for more information on the measures of self-control included). As Table 7 indicates, the interaction term is negative and statistically significant for reading (β = -0.080, *SE* = 0.023, $p < 0.001$) and also negative but not statistically significant for math (β = -0.020, *SE* = 0.018, p = 0.289). This result is consistent with the hypothesis according to which the within-person effect of executive functions on achievement is lower for individual's with higher within-person changes in self-control.

5.5.4 Within-person random slopes

I estimated a hybrid model with a random slope as specified in equation 11. The estimated standard deviation of the random slope of executive functions on achievement was 0.362 (*SE* = 0.013) for reading and 0.295 (*SE* = 0.013) for math. Assuming normality, this implies that 95% of individuals have a within-person effect of executive function on achievement between 0.77 and -

0.65 for reading and between 0.62 and -0.53 for math. Estimating these confidence intervals is important, as they indicate that the within-person effect can range widely across individuals, and can actually have different signs for different individuals. The estimated correlation between the random intercept and slope is slightly negative (-0.03 for reading and -0.01 for math), which indicates that students with higher initial performance tend to have smaller within-person effects of executive functions on achievement.

5.5.5 Within-person reciprocal causation

I estimated the ML-SEM model presented by Allison et al. (2019) using different lag specifications. The contemporaneous effect of executive functions was the same for reading and for math (β = 0.019, *SE* = 0.005, $p < 0.001$ for reading and β = 0.019, *SE* = 0.004, $p < 0.001$ for math). I also estimated the model with both contemporaneous and lagged effects. In the case of reading, the contemporaneous effect was 0.068 (*SE* = 0.010, $p < 0.001$), and the estimated lagged effect was 0.035 (*SE* = 0.006, $p < 0.001$). On the other hand, the estimated contemporaneous effect for math was 0.029 (*SE* = 0.007, $p < 0.001$), and the estimated lagged effect was 0.011 (*SE* = 0.005, $p < 0.05$). One can see, then, that in both cases the contemporaneous effect increased when the model included a lagged effect.

I also implemented a general cross-lagged panel model as described in Zyphur et al. (2019) using a 1-unit lag. As Table 8 indicates, the overall fit of the model was good (disregarding the chi square test, which is sensitive to both sample and model size, which are relatively large in the present application). Table 7 displays the estimated cross-lagged effects of the residuals as predictors. One can perceive that there was no significant effect of executive functions on reading (β = 0.001, *SE* = 0.006, $p = 0.858$). The absence of an effect of reading achievement on executive

functions is further confirmed by a slight improvement in fit (in BIC and AIC) when this effect is fixed to zero (see Table 8). The reciprocal effect of reading on executive functions was not significant either (β = 0.034, *SE* = 0.024, p = 0.153). On the other hand, the cross-lagged effects between math and executive functions were positive and significant: the estimated effect of math on executive functions was 0.027 (*SE* = 0.006, $p < .001$) and the estimated effect of executive functions on math was 0.099 (*SE* = 0.029, $p < .01$). The presence of this effect is reinforced by the fact that, as Table 8 indicates, removing this effect reduces model fit (based on BIC and AIC).

Table 8. Fit statistics of the full and constrained General Cross-Lagged Panel Model.

Reading	χ^2	*df*	RMSEA	SRMR	CFI	TLI	BIC	AIC
Full model	1,187.815	67	0.031	0.030	0.978	0.970	234,408.96	234,004.79
Constraint reading → executive functions	1,187.682	68	0.031	0.030	0.978	0.970	234,399.31	234,002.90
Math								
Full model	1,047.656	67	0.029	0.027	0.983	0.977	260,587.56	260,183.39
Constraint math → executive functions	1,077.576	68	0.029	0.028	0.982	0.976	260,659.98	260,263.58

Note. CFI = Confirmatory Fit Index; TLI = Tucker-Lewis Index; RMSEA = root mean squared error of approximation; SRMR = standardized root mean squared residual; AIC = Akaike Information Criterion; BIC = Bayes Information Criterion

5.6 Discussion

Psychological research is mainly concerned with causal questions, and can therefore benefit from explicitly adopting a causal inference framework. I reviewed some basic concepts in the causal inference literature related to the definition and identification of causal effects, particularly using observational longitudinal data. In addition, I explained that, from a causal perspective, the main advantage of longitudinal data is that it allows us to identify causal effects under less restrictive assumptions. In particular, it allows us to replace the unrealistic assumption that we can block all backdoor paths by conditioning on all confounders, by the less restrictive assumption that the effects of all confounding variables are stable over time; that is, that causal effects are identified within individuals. However, the advantages of longitudinal data can be realized only by using statistical methods that appropriately distinguish within and between person relationships. I reviewed the main within-person estimators (the so-called unit fixed effects models), as well as methodological strategies that can be used to test substantive hypotheses regarding within-person asymmetrical causation, moderation, effect heterogeneity, and reciprocal causation.

In the empirical application, I used the modelling strategies presented to estimate the effect of executive functions on academic achievement. Substantively, the results did not provide significant results regarding asymmetrical causation, nor a significant interaction with gender. However, the within-person interaction with self-control was statistically significant. In addition, the within-person random effects model suggests that there is considerable variance in the within-person effect of executive functions on achievement, with a large proportion of individuals actually

having a negative effect. Finally, the models that tested reciprocal effects generated conflicting results. For example, while the ML-SEM model suggests that the effect of reading is larger than the effect of math, the GCLPM model suggests that the effect of math is larger than the effect of reading.

Even if one cannot make definitive conclusions, the results from the GCLPM might be preferred, given that it controls for a wider range of confounders and follows the principle that the cause needs to precede the effect. In this scenario, we would conclude that the estimated effect of executive functions on reading achievement is 0.001 (*SE* = 0.006, *p*= *.858*) and on math achievement is 0.027 (*SE*= 0.006, *p*< *.001*). The latter effect is equivalent to around 0.11 within-person standard deviations. These results coincide with a general finding in the literature, according to which executive functions seem to have a higher effect on math than reading (e.g., Fuhs et al., 2014; Schmitt et al., 2017).

In conclusion, it is worth noting a simple fact: the choice of the model matters. As the results of this study indicate, one can obtain considerably different results depending on the model that one implements. Yet model choice can be a difficult endeavor for several reasons. First, it might not be possible to construct a model that tests a substantive hypothesis while controlling for all possible confounders. For example, it is difficult to implement the general cross-lagged panel model described by Zyphur et al. (2019) –which controls for several potential confounders–, while testing asymmetrical causation or within-person interactions. Second, deciding which is the best model requires knowledge about the underlying data-generation process. This is the reason why causal inference is so difficult, and impossible to validate just with the observed data. Not only we need to make assumptions about the absence of unmeasured confounders, but we also need to assume, for example, that the relationships between the observed variables is adequately specified,

or that the temporal lags in the model match the temporal processes in the real world. However, an explicit recognition of these assumptions might help researchers better assess and understand their modelling strategies.

6.0 General discussion

This dissertation discusses foundational conceptual and methodological issues in skill development research. I began by presenting a framework for describing individual differences in skill development; then, I presented an approach for identifying relevant factors for explaining those differences; and I concluded by estimating the effects of those predictors on academic achievement. Even if the specific questions and methodologies differed across the three studies, the dissertation follows a logical progression in scientific research, from descriptive to explanatory analyses, using similar datasets. Below, I briefly reprise some of the findings from each study to demonstrate how they build on one another.

In the first study I argue that the concept of academic mobility provides a useful normative and methodological framework for describing the development of educational inequalities. The mobility metrics presented provide a more comprehensive and fine-grained description of distributional change than the average-based measures typically used in educational research. Some of the findings in the empirical analysis, which cannot be easily obtained using traditional methods, include that (1) Black and Hispanic students are more mobile than Asian and White students, especially in the first years of schooling; (2) Black students are more likely to move downward at every segment of the achievement distribution; and (3) throughout K-8 schooling, differences in literacy achievement between racial groups increased, while difference within these groups decreased.

As explained in the introduction, the standards of success for descriptive models are related to their informativeness and parsimony. In accordance with these criteria, in the first study I argue that the concept of academic mobility presents a more informative and parsimonious framework

compared to traditional mean-based measures. In terms of their informativeness, I show that academic mobility metrics provide relevant information that is not captured by traditional methods, e.g., an estimate of the overall degree of mobility of a particular education system, or how mobile or persistent are the individuals at the bottom and at the top of the distribution. The ability to "shed some light in the corners" (Miller as cited in Koretz, 2017, p.31) has long been an aspiration in descriptive research in education, but has been largely restricted by the emphasis on means and variances.

In terms of parsimony, I show how the mobility metrics provide scalar summary statistics that can be easily used to make comparative assessments at the population and sub-population levels. This can be contrasted, for instance, with growth models, which typically require more involved interpretations including the estimated intercepts, slopes and variances. In addition, it is worth noting that some metrics (e.g., transition probabilities) relax important assumptions in skill development research –notably the interval scale assumption and the functional form assumption, which are important assumptions in growth modelling. In summary, I show how the concept of academic mobility provides an informative and parsimonious framework for describing individual differences in achievement.

Apart from a purely descriptive value, the concept of academic mobility has the potential to lead to new and productive explanatory research . Throughout this dissertation, I stress the distinction between descriptive and causal research, as they correspond to substantially different goals that have different standards of success, uses and supporting assumptions. At the same time, I suggest that descriptive research provides a foundation for causal research, as it helps define the phenomena, problems and hypotheses that subsequent research will try to explain. It is

unquestionable, for example, that achievement gaps have shaped in a fundamental manner the way we think and explain educational inequality.

The usefulness of the mobility framework presented for explanatory research can be justified by considering two ideas: first, the mobility metrics discussed can be used to conduct fine-grained comparative analyses; and second, causal claims are always comparative claims (e.g., Rubin, 1974). The fine-grained descriptions provided by the mobility metrics can help us, then, define or conceive more nuanced explanatory hypotheses. One way of interpreting the argument put forward in the first study is that comparing the average achievement scores of two groups (e.g., White and Black children) is too broad a comparison for both descriptive and explanatory purposes; a more productive explanatory strategy might consist in generating more detailed comparisons that can stimulate more specific research questions, for example: Why were Black students who began at the top of the achievement distribution more likely to move downward compared to students of other races who began at a similar achievement level? Why were Asian students more likely than students of other races to move from the bottom to the top of the achievement distribution? Why do students tend to be more mobile at the beginning of the schooling process? Why do some students from the same racial group move upward while others move downward, conditional on initial achievement? These questions illustrate how the mobility framework presented can be used to generate fine-grained comparisons that can be used to conceive more nuanced and –potentially– more productive research questions and explanatory hypotheses.

Apart from defining adequate comparisons (e.g., counterfactuals), causal research implies adopting a particular identification strategy (see study 3). Typically, the strategy adopted in similar studies is to estimate the change in the parameter of interest (e.g., related to the Black-White gap)

as one controls for various factors (e.g., Chetty et al., 2018; Fryer & Levitt, 2004). For example, Fryer and Levitt (2004) implement several OLS regression models with a wide range of covariates (98), and examine the extent to which controlling for particular factors reduces the unconditional Black-White achievement gap. The set of covariates that reduces this gap is said to "explain" the differences in achievement between these two groups of students.

In the second and third studies, I discuss some of the limitations of these explanatory strategies, and argue for a more principled and systematic approach. Notably, I explain that typically it is not a good idea to control for observed variables in an unprincipled manner (especially when dealing with longitudinal data), given that statistical control can be both a bias-removing as well as a bias-amplifying mechanism (see study 3). More generally, I explain that adequate statistical control requires knowledge of causal structures, as they specify which variables should be included (and which variables should not be included) in the adjustment set.

In the second study, I argue that (1) causal structural knowledge can be very helpful in explanatory research in skill development; (2) causal graphs are useful tools for representing causal structural knowledge; and (3) causal search algorithms provide a principled way to estimate causal graphs in skill development, as we normally deal with large numbers of variables with a (mostly) unknown causal structure.

Even if causal structural knowledge can be used for different purposes (e.g., synthesizing research findings and assessing causal modelling assumptions), the main purpose of the second study, in relation to the overall narrative of this dissertation, was to justify the explanatory relevance of explanatory variables. I explain that most of the explanatory research in skill development intends to identify the effect size for particular factors, and I suggest that this approach has important limitations in terms of the cumulative knowledge that it can generate for

the field; in particular, I highlight that, in this approach, the choice of explanatory variables can be unjustified (and, as a consequence, can be influenced by value judgements; see below) and that it does not elucidate how alternative factors relate to the outcome and to each other. In the long run, if we rely only on this methodological framework we will end up with a list of disconnected and structureless effect sizes that provide shallow scientific understanding and limited external validity.

In response to these limitations, I argue for the importance of mechanism-based explanations in general and of proximal mechanisms in particular. Focusing on the latter presents several theoretical, methodological, practical and policy-level advantages. Theoretically, proximal mechanisms can help us understand the phenomena under investigation, as they can be considered "the primary engines of development" (Bronfenbrenner and Morris 2006); methodologically, proximal mechanisms can help us obtain stable and accurate predictions of the target variable; practically, focusing on proximal mechanisms help us reduce the set of explanatory variables to consider; and in terms of policy-making, identifying the proximal mechanisms can help us guide effective interventions.

In the second study, I used a nationally representative dataset with a wide range of relevant contextual and psychological factors, and applied four causal search procedures that varied important dimensions in the search algorithms. A consistent finding in this study is that, in the dataset considered and using several model specifications, executive functions (in particular working memory, cognitive flexibility and attentional focusing) are directly related to academic achievement. In this study, I also present the entire causal graph generated by the search procedures, which provides useful insights into the overall causal structure of academic

achievement (e.g., it shows which variables are *not* directly related to achievement; which variables are mediated by executive functions; the causal pathways of demographic variables, etc.).

In general, however, we are not only interested in whether a set of variables is directly related to a target variable, but also on how strong or weak each relationship is. Consequently, a natural step in the dissertation was to estimate the magnitude of the coefficient describing the causal relationship between executive functions and academic achievement. Thus, in the third study I review different methodological strategies that allow us to estimate this causal relationship, as well as test substantive hypotheses regarding how these two variables are causally related (e.g., if there is reciprocal causation or effect heterogeneity). The empirical analyses show that one can obtain different results depending on the model that one implements. However, according to the preferred model and in agreement with prior research, executive functions have a stronger effect on math than on reading achievement.

6.1 Synthesis across studies

6.1.1 Theoretical contributions

This dissertation has two main theoretical contributions. First, the implementation of different academic mobility metrics to a national representative sample indicates that, throughout schooling, differences between racial groups increase, while differences within these groups decrease. In particular, I found that Black students are more likely to move downward at every segment of the distribution, which can be considered a case of manifest disadvantage. Even if around half of the variance in rank achievement remains stable from kindergarten to eight grades,

I found clear differences in mobility patterns across racial groups, conditional on initial achievement. This finding contradicts a common view in skill development, according to which initial advantages and disadvantages accrue over time (the so-called Mathew effect hypothesis). Contrary to this belief, this study shows that some students are more likely to move upward while other students are more likely to move downward, irrespective of their initial status.

The second theoretical contribution is that, among a wide range of psychological and contextual factors, prior achievement and executive functions are proximal mechanisms of math and reading achievement. In addition, while motivation, parent education, peer victimization and peer reading level are only directly related to reading achievement, race and gender are only directly related to math. I also show how 36 variables related to different child, family, peer, school and neighborhood-level effects that have been considered in the literature as important determinants of academic achievement are causally (or probabilistically) related to each other.

These are important theoretical findings that are worth investigating in future research. For example, most of the empirical research in education includes demographic variables by default, without specifying how these variables might relate to the outcome or to other covariates. The results of the second study shed some light on this issue; for instance, they indicate that while the relationship between gender and reading is mediated by motivation (and potentially inhibitory control), the relationship between gender and math is not mediated by any of the variables included in the dataset. Similarly, the results suggest that school and family characteristics mediate the effect of race on reading (but not on math) achievement. Even if these results should not be taken as conclusive, they do support interesting research hypothesis for future research, which might have important policy implications. In addition, it is worth noting that information regarding the underlying structure is not provided by traditional methods (e.g., regression analysis).

6.1.2 Methodological contributions

Methodologically, this dissertation has four main contributions. First, I present a general framework (represented by the concept of academic mobility) which operationalizes the concept of educational inequality using learning outcomes. More specifically, the first study contributes to the literature by discussing important aspects to consider in order to measure educational inequality, and by presenting five mobility metrics that can be used to measure different aspects of academic mobility.

The second methodological contribution is related to the application of causal search algorithms to social and behavioral problems. Even if these methods have been used as an exploratory tool for observational data in different fields, there have been scant implementations of these algorithms in social and behavioral sciences. In the second study, I illustrate how these methods can be used to shed light on important social and behavioral problems.

The third methodological contribution is related to the explanation and integration of a range of methodological tools that can be used to model longitudinal data. In particular, I present different modelling strategies that can be used to test substantive hypotheses regarding the data-generating process, while capitalizing on one of the main advantages that panel data offers, namely the ability to control for all time-invariant confounders.

Finally, this dissertation shows how different methodological approaches can be integrated in a principled and progressive fashion. In a highly interdisciplinary and fragmented field, it is important to distinguish between the different aims of quantitative research (e.g., description and explanation), as well as how different methodological approaches complement each other. Throughout the three studies, I showed how the methods implemented (mobility metrics, causal search algorithms and unit fixed-effects models) can be integrated in a broader program of

research, and provide important information that cannot be obtained with more traditional methods (e.g., achievement gaps, growth models, or linear regression). My goal was not to suggest that the latter methods are somehow defective or inadequate, but rather to present a variety of methods that can expand our methodological toolkit.

6.2 Supporting value judgements in educational research

In a more general way, this dissertation contributes to the field by providing conceptual and methodological tools that can be used to support or hold accountable value judgments that are often made in skill development research. One of the main reasons why education research has been so highly contested is because values play a central role (Dolle, 2008; Shavelson & Towne, 2002). Empirical research in education is often perceived as reflecting an inextricable combination of personal, disciplinary, social, moral, cultural and political values. It is of the utmost importance, then, to bring to full attention and critically scrutinize the value judgments implied in current research. In this dissertation, I present some arguments that can help us subject value judgments in both descriptive and explanatory research to rational scrutiny.

6.2.1 Value judgments in descriptive research

Value judgments provide the foundation for measurement practices as well as the ultimate justification for choosing one particular metric over another. If we disagree on the things we care about, then we cannot agree on what the appropriate measures for describing the phenomenon of interest should be. This idea is commonly recognized in the literature on educational measurement

(e.g., Betebenner, 2008, 2011; Kane, 2006; Messick, 1994). Statistical or psychometric considerations alone cannot support the use of one metric over another; ultimately, this support can only be provided by value (i.e., normative) considerations.

Even if many researchers agree with this idea, normative and methodological considerations are often blurred together. For example, studies on educational inequality are often based on intuitive ideas related to justice and equality (e.g., equality of opportunity), but these notions are rarely clearly defined. Consequently, there has been some confusion about the adequate ways of describing and evaluating education systems. The first study in this dissertation brings some clarity to these issues by relying on previous discussions regarding the normative principles of equality of opportunity. In particular, I clearly distinguish and discuss normative and methodological properties and assumptions behind metrics of educational inequality.

6.2.2 Value judgments in explanatory research

Value judgments can also play an important role in explanatory research. Modern frameworks for explanatory research in general, and causal inference in particular, focus on the effects of causes rather than the causes of effects (Holland, 1986). This implies that researchers who pursue explanatory studies typically examine the causal effect of some particular factor on the phenomenon of interest. Yet the choice of the factor itself, and the particular hypothesis to be tested, is often based on the researcher's value judgments (Longino, 1990)[22]. The potential

[22] In order to illustrate this point, one can consider the case of executive functions. As explained above, I justify the explanatory relevance of this construct based on the fact that it was consistently identified as a proximal mechanism across several causal search procedures. However, many of the reasons in the literature for justifying the

problems of this approach are apparent in explanatory research on student academic achievement, where researchers have focused on a myriad of factors based on their own background assumptions; for example, while sociologists and economists tend to examine macro-level factors and processes (e.g., institutional structures, incentive policies, neighborhood or school effects), psychologists tend to focus on micro-level processes (e.g., related to family or individual-level dynamics). Moreover, there have also been long-standing debates within disciplines regarding the importance of particular factors (e.g., school versus family effects, Coleman et al., 1966).

In sum, value judgments play an important role in establishing the causes and hypotheses that researchers decide to investigate. That is, explanatory relevance is to some extent intuitively determined rather than founded on evidence-based considerations. Yet, as cognitive scientists have shown, causal intuitions are typically affected by moral judgements, as we are "moralizing creatures through and through" (Knobe, 2010, p.328). We need then to conceive of arguments that can help us confirm or disconfirm judgments associated to the explanatory relevance of particular factors. In the second study, I discuss some conceptual and methodological tools that can help us in this task. In addition, I argue that the lack of consensus regarding ways of defining explanatory

importance of executive functions reflect –often unjustified or unrecognized– value judgements. For example, the reasons put forward by a leading scholar in the field in a highly-cited review clearly reflect moral judgments: "Self-control is about resisting temptations and not acting impulsively. The temptation resisted might be to indulge in pleasures when one should not (e.g., to indulge in a romantic fling if you are married or to eat sweets if you are trying to lose weight), to overindulge, or to stray from the straight and narrow (e.g., to cheat or steal). Or the temptation might be to impulsively react (e.g., reflexively striking back at someone who has hurt your feelings) or to do or take what you want without regard for social norms (e.g., butting in line or grabbing another child's toy)" (Diamond, 2013, p.138).

relevance (in particular in education research) has contributed to the lack of cumulative knowledge expected of scientific endeavors (Shavelson & Towne, 2002).

6.3 Limitations and future research

As noted in the first study, the academic mobility metrics presented are based on assumptions that need to be examined in future research. First, some metrics (e.g., the rank-rank slope and the START model) are sensitive to the underlying scale. Even if the use of rank scores is justified inasmuch as I was exclusively interested in relative or positional change, it is unclear to what extent these metrics will change if one considers scores with a different scale. Second, measures of the amplitude of intraindividual mobility were estimated using six waves, which – based on studies examining the reliability of these measures– is a relatively low number of measurement occasions. In order to have more precise estimates, we would need to have more waves, as well as measures that are closer in time. Finally, the transition probabilities are based on arbitrary cutoffs that affect the estimated probabilities. I used a quartile partition (rather than, for example, a quintile partition, which is common in the literature) based only on sample size considerations. Future studies can overcome this limitation by using a more principled partition (e.g., by using proficiency cut scores or future outcomes).

As noted in the first study, there are other ways in which this work can be further extended. In particular, one can use the metrics presented to answer a range of questions that cannot be easily answered with traditional methods; for example, one can investigate whether academic mobility at the population and subpopulation levels has changed over time; how does academic mobility in particular educational systems (e.g., districts or countries) compare with academic mobility in

other educational systems; or how do mobility patterns vary across the achievement distribution. In terms of more foundational research, one can investigate other mobility measures –possibly with fewer assumptions– that can be used to describe distributional change.

The main limitation of the second and third studies relates to the difficulty of making causal inferences using observational data. In the second study, the main concern is that the estimated connections are spurious due to unobserved confounders; and in the third study the main concern is that the estimated coefficients are biased due to time-varying confounders. Regarding the latter study, it is worth noting that focusing on proximal mechanisms appears to have both advantages and disadvantages from a methodological perspective. On the one hand, by focusing on proximal mechanisms the risk of overcontrol bias is reduced, as only other proximal mechanisms might mediate their effect. On the other hand, the risk of confounding bias might increase, as –compared to more distal factors– there might be more variables that have a direct effect on both the outcome of interest and the proximal mechanism. For example, even though one can conceive of different potential confounding paths between executive functions and academic achievement (e.g., related to psychological, family or school factors), it is more difficult to conceive of confounding paths between distal factors (e.g., home neighborhood safety) and academic achievement. Following this line of reasoning, focusing on distal mechanisms might be a better strategy if the goal is to minimize the risk of confounding bias.

Confounding bias might be present, then, in the estimated coefficients computed in the third study. A further limitation of this study is that all models were estimated without statistical control. Consequently, a future direction of research would be to consider a principled method for selecting the adjustment set based on proximal/distal considerations.

Regarding the second study, a possible future direction of research is to validate the results of the causal graphs presented by running similar models in other datasets, as well as implementing different algorithms (e.g., time series causal search algorithms; or algorithms that relax the causal sufficiency assumption –e.g., the FCI algorithm). One could also validate the presence or absence of a particular causal connection by considering the available experimental evidence.

There are several additional issues that were left open in the second study. First, as explained in the paper, one of the benefits of identifying causal structures is that they can help guide effective interventions. For example, interpreted causally, the path $Inhibitory\ control \rightarrow Externalizing \rightarrow Peer\ victimization \rightarrow Math\ (4th)$ implies that manipulating any variable in the chain will affect academic achievement. Given the number of proximal and distal causes identified, one might wonder which of the proximal and distal causes identified are more easily manipulated. Furthermore, one might be interested in estimating the magnitude of the coefficients of all the connections in the graph. A fully parameterized graph could help us understand the extent to which various proximal and distal contextual and psychological factors influence academic achievement.

Another issue for further investigation concerns the causal structure around executive functions. The estimated graphs indicate that there is no measured variable that has a direct effect on some of the main measures of executive functioning, namely working memory, cognitive flexibility and attention. This suggests that executive functions represent a self-contained system that is not affected by contextual factors. A further direction of research is to investigate which measured or unmeasured factors might have a direct effect on these variables, which might also help us design effective interventions. Lastly, as noted in the paper, we need a better understanding

on the extent to which the executive functions and achievement measures are distinct or overlapping constructs (i.e., whether they are conceptually related).

Finally, it is worth stressing that the fact that the explanatory (or causal) analyses in this dissertation are based on strong and –given the available dataset– untestable assumptions does not imply that these analyses cease to be explanatory (or causal). As I argue in the two papers, (1) whether a model is, e.g., "descriptive" or "causal" is not intrinsic to the model itself, but depends rather on its interpretation or use; (2) it is better for transparency and testability purposes to explicitly state the intended interpretation or use of the model (e.g., descriptive or explanatory) as well as its (testable and untestable) assumptions, rather than being ambivalent about the interpretation and use of the model (e.g., by avoiding causal terminology but deriving or suggesting causal interpretations). As Bollen and Pearl (2013, p.307) explain, causal modelling using observational data can be useful when the model is based on assumptions that 'are defensible and consistent with the current state of knowledge, and the analysis is done under the speculation of 'what if these causal assumptions were true.'' The success of an explanatory model should not be established exclusively by the conclusiveness of its assumptions, but also by the kind of insights and hypotheses it might generate.

Appendix A : Supplemental materials for study 1

Appendix Table 1. Sample size and descriptive statistics of percentile rank scores in reading achievement by race.

	Kindergarten			Eight Grade		
	N	Mean	SD	N	Mean	SD
Unweighted						
Overall	7,254	50.01	28.87	7,713	50.01	28.87
Asian	333	58.13	30.27	430	55.45	28.80
Black	762	37.93	26.98	751	29.01	24.69
Hispanic	893	39.95	28.85	1,302	37.77	27.12
Other	419	40.59	30.11	422	42.42	29.25
White	4,842	54.01	27.71	4,802	56.80	27.02
Weighted						
Overall	7,254	47.18	29.21	7,713	45.78	29.13
Asian	333	57.98	30.67	430	54.64	29.34
Black	762	37.65	27.49	751	27.91	24.36
Hispanic	893	36.94	28.20	1,302	36.33	26.69
Other	419	40.00	31.82	422	43.65	28.63
White	4,842	52.14	28.14	4,802	53.66	27.89

Note. The percentile ranks were created using only the analytic sample, i.e., individuals with a non-missing or non-zero value in the appropriate longitudinal weight. Thus, the unweighted sample displayed only includes the students with a positive value in the appropriate longitudinal weight.

Appendix Table 2. Rank-rank estimates by racial group.

	Overall	Asian	Black	Hispanic	Other	White
Rank-rank slope, β_r	0.537	0.513	0.432	0.452	0.535	0.512
	(0.018)	(0.055)	(0.047)	(0.035)	(0.055)	(0.023)
Intercept, α_r	21.874	28.926	11.548	26.068	22.678	26.847
	(1.114)	(4.168)	(1.696)	(1.476)	(3.919)	(1.465)
R_2	.29	.31	.24	0.23	0.36	0.27

Note. The appropriate longitudinal weight was used in the estimation. Jackknife standard errors are in parentheses.

Appendix Table 3. Indicators of the amplitude of intraindividual mobility by race.

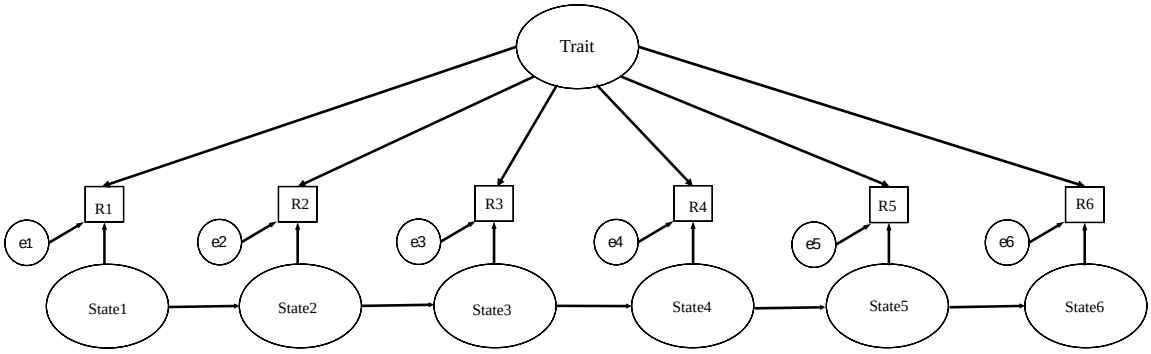
	Overall	Asian	Black	Hispanic	Other	White
ISD	14.6	14.5	14.6	13.7	14.2	14.9
	(0.153)	(0.631)	(0.516)	(0.306)	(0.699)	(0.197)
$MSSD_{1/2}$	17.1	17.2	16.9	16.4	16.2	17.5
	(0.156)	(0.716)	(0.482)	(0.374)	(0.642)	(0.221)

Note. The appropriate longitudinal weights was used in the estimation. Jackknife standard errors are in parentheses.

Appendix Table 4. National quartile transition matrix (%).

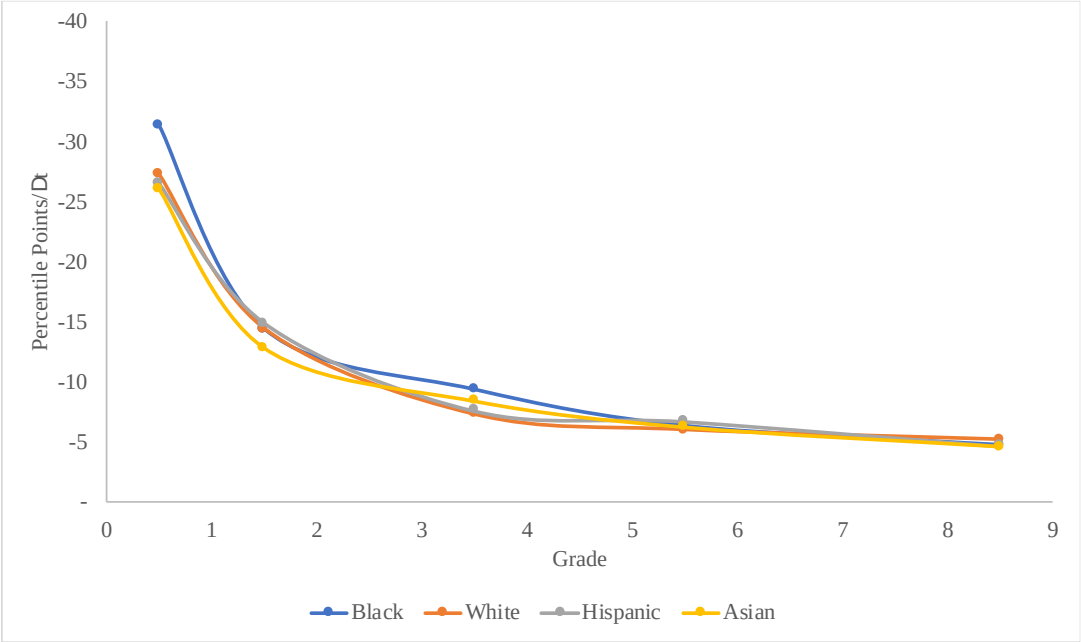
Quartile in Kindergarten	Quartile in 8th grade			
	1	2	3	4
1	51.0	29.3	14.2	5.4
2	30.8	32.0	22.5	14.7
3	19.3	23.1	31.3	26.3
4	6.6	15.0	32.4	46.0

Note. The appropriate longitudinal weight was used in the estimation.

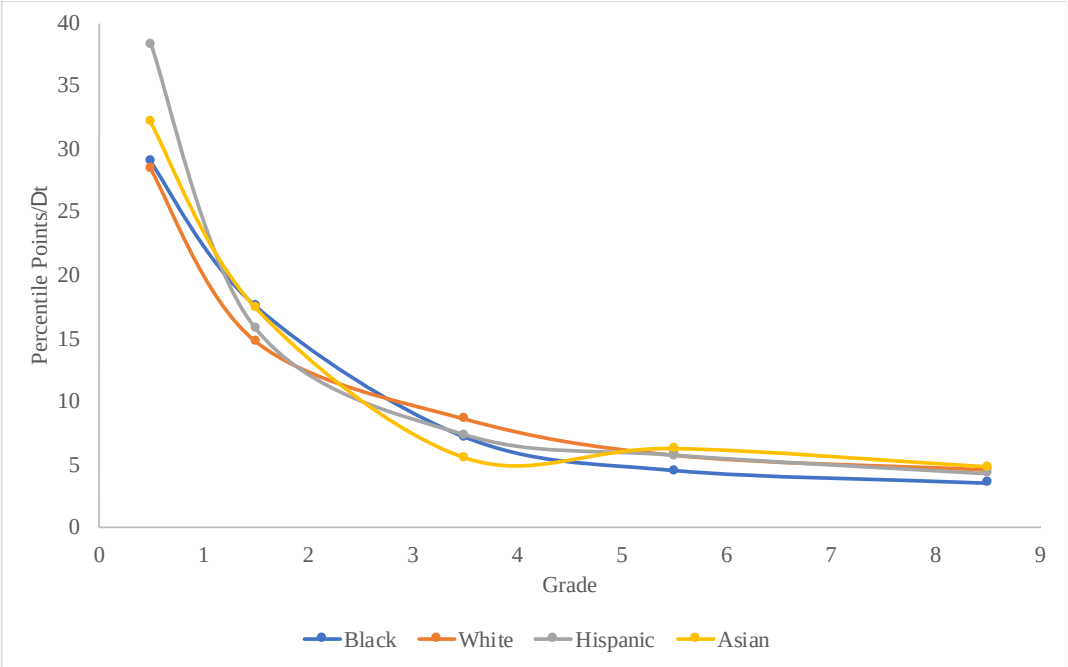


Appendix Figure 1. The stable trait, autoregressive trait, and state (START) model.

A.

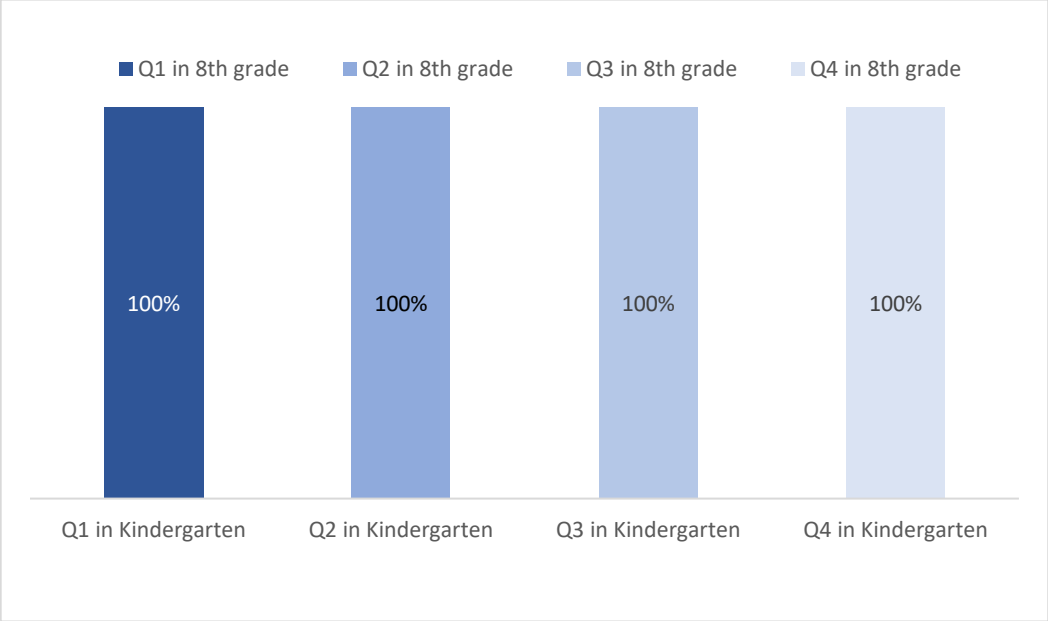


B.

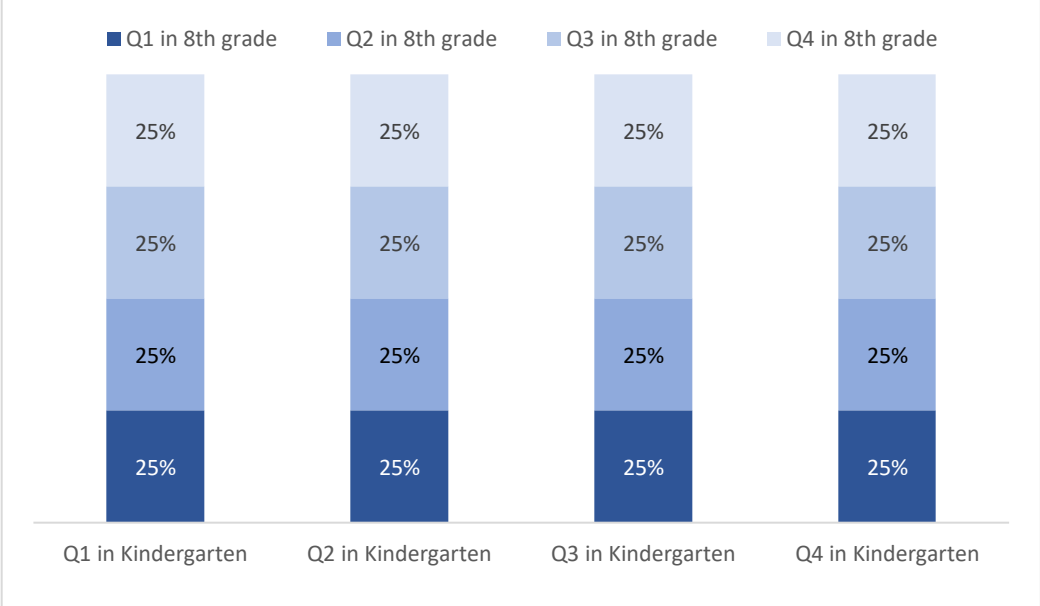


Appendix Figure 2. Mean negative (1) and positive (2) ocassion-to-occasion achievement difference by race. I divided the raw differences by the respective time interval in order to make the coefficients comparable across time (i.e., they estimate the upward or downard mob

A.



2.



Appendix Figure 3.Transition bar charts representing systems with no mobility (1) and complete mobility (2).

Bibliography

Aikens, N. L., & Barbarin, O. (2008). Socioeconomic differences in reading trajectories: The contribution of family, neighborhood, and school contexts. *Journal of educational psychology*, *100*(2), 235.

Allison, P. D. (2009). *Fixed effects regression models* (Vol.160). London, UK: Sage.

Allison, P. D. (2019). Asymmetric fixed-effects models for panel data. *Socius*, 5, 1-12

Allison, P., R. Williams, & E. Moral-Benito. (2017). Maximum Likelihood for Cross-Lagged Panel Models with Fixed Effects. *Socius* 3, 1–17.

Aliferis, C. F., Statnikov, A., Tsamardinos, I., Mani, S., & Koutsoukos, X. D. (2010). Local causal and markov blanket induction for causal discovery and feature selection for classification part i: Algorithms and empirical evaluation. *Journal of Machine Learning Research*, 11(Jan), 171-234.

Amato, P. R., & Anthony, C. J. (2014). Estimating the effects of parental divorce and death with fixed effects models. *Journal of Marriage and Family*, *76*(2), 370-386.

Anderson, E. (2004). Uses of value judgments in science: A general argument, with lessons from a case study of feminist research on divorce. *Hypatia*, 19(1), 1-24.

Anderson, E. (2007). Fair opportunity in education: A democratic equality perspective. *Ethics*, *117*(4), 595-622.

Anderson, E. (1999). What is the Point of Equality?. *Ethics*, *109*(2), 287-337.

Anderson, E. (2010). The fundamental disagreement between luck egalitarians and relational egalitarians. *Canadian journal of philosophy*, *40*(sup1), 1-23.

Andrews, B., Ramsey, J., & Cooper, G. F. (2018). Scoring Bayesian networks of mixed variables. *International journal of data science and analytics*, 6(1), 3-18.

Angrist, J. D., & Pischke, J. S. (2008). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton: Princeton university press.

Angrist, J. D., & Pischke, J. S. (2010). The credibility revolution in empirical economics: How better research design is taking the con out of econometrics. *Journal of economic perspectives*, *24*(2), 3-30.

Arneson, R. (1989). Equality and equal opportunity of welfare. *Philosophical Studies,* 56, 77-93.

Bailey, D. H., Watts, T. W., Littlefield, A. K., & Geary, D. C. (2014). State and trait effects on individual differences in children's mathematical development. *Psychological Science, 25*(11), 2017-2026.

Bailey, M.J., & Dynarski, S.M. (2011). *Gains and gaps: Changing inequality in U.S. college entry and completion* (National Bureau of Economic Research Working Paper No. 17633). Cambridge, MA: National Bureau of Economic Research

Baird, R., & Maxwell, S. E. (2016). Performance of time-varying predictors in multilevel models under an assumption of fixed or random effects. *Psychological methods*, *21*(2), 175-188.

Bell, A., & Jones, K. (2015). Explaining fixed effects: Random effects modeling of time-series cross-sectional and panel data. *Political Science Research and Methods*, 3(1), 133-153.

Bernier, A., Carlson, S. M., & Whipple, N. (2010). From external regulation to self-regulation: Early parenting precursors of young children's executive functioning. *Child development*, *81*(1), 326-339.

Berk, R. A. (2004). *Regression analysis: A constructive critique*. Thousand Oaks, CA: Sage.

Betebenner, D. W. (2009). Norm-and criterion-referenced student growth. Educational *Measurement: Issues and Practice*, 28, 42–51.

Bodovski, K., & Farkas, G. (2008). "Concerted cultivation" and unequal achievement in elementary school. *Social Science Research*, *37*(3), 903-919.

Bollen, K. A. (1989). *Structural equations with latent variables*. Wiley. New York.

Bollen, K. A., & Pearl, J. (2013). Eight myths about causality and structural equation models. In *Handbook of causal analysis for social research* (pp. 301-328). Springer, Dordrecht.

Bollen, K. A., & Brand, J. E. (2010). A general panel model with random and fixed effects: A structural equations approach. *Social Forces*, 89, 1–34.

Bond, T. N., & Lang, K. (2013). The evolution of the Black-White test score gap in Grades K–3: The fragility of results. *Review of Economics and Statistics*, *95*(5), 1468-1479.

Breen, R., & Jonsson, J. O. (2000). Analyzing educational careers: A multinomial transition model. *American sociological review*, 65(5), 754-772.

Breen, R., & Jonsson, J. O. (2005). Inequality of opportunity in comparative perspective: Recent research on educational attainment and social mobility. *Annu. Rev. Sociol*., 31, 223-243.

Briggs, D., & Betebenner, D. (2009, April 14). *Is growth in student achievement scale dependent?* Paper presented at the invited symposium Measuring and Evaluating Changes in Student Achievement: A Conversation about Technical and Conceptual Issues at the annual meeting of the National Council for Measurement in Education, San Diego, CA.

Brighouse, H., & Swift, A. (2006). Equality, priority, and positional goods. *Ethics*, *116*(3), 471-497.

Bronfenbrenner, U., & Morris, P. A. (2006). The bioecological model of human development. In R. M. Lerner (Ed.) *Handbook of child development: Vol. 1. Theoretical models of human development* (6th ed., pp. 793-828). Hoboken, NJ: Wiley

Burdick-Will, J. (2013). School violent crime and academic achievement in Chicago. *Sociology of education*, *86*(4), 343-361.

Cameron, C. E., Grimm, K. J., Steele, J. S., Castro-Schilo, L., & Grissmer, D. W. (2015). Nonlinear Gompertz curve models of achievement gaps in mathematics and reading. *Journal of Educational Psychology*, *107*(3), 789-804.

Castellano, K. E., & Ho, A. D. (2013). Contrasting OLS and quantile regression approaches to student "growth" percentiles. *Journal of Educational and Behavioral Statistics*, 38(2), 190-215.

Castellano, K. E., & Ho, A. D. (2013). A practitioner's guide to growth models. Washington, DC: Council of Chief State School Officers

Chetty, R., Hendren, N., Kline, P., & Saez, E. (2014). Where is the land of opportunity? The geography of intergenerational mobility in the United States. *The Quarterly Journal of Economics*, *129*(4), 1553-1623.

Chetty, R., Hendren, N., Jones, M. R., & Porter, S. R. (2018). *Race and economic opportunity in the United States: An intergenerational perspective* (No. w24441). Cambridge, MA: National Bureau of Economic Research.

Chickering, D. M. (2002). Optimal structure identification with greedy search. *Journal of machine learning research*, *3*(Nov), 507-554.

Clotfelter, C. T., Ladd, H. F., & Vigdor, J. L. (2006). *The academic achievement gap in grades three to eight* (Working Paper No. 12207). Cambridge, MA: National Bureau of Economic Research.

Cohen, G. A. (1989). On the currency of egalitarian justice. *Ethics*, *99*(4), 906-944.

Coleman, J. (1968). The concept of equality of educational opportunity. *Harvard educational review*, *38*(1), 7-22.

Collins, A., & Koechlin, E. (2012). Reasoning, learning, and creativity: frontal lobe function and human decision-making. *PLoS biology*, 10(3), e1001293.

Cook, T. D., Campbell, D. T., & Shadish, W. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston, MA: Houghton Mifflin.

Cunha, F., & Heckman, J. (2007). *The technology of skill formation* (No. w12840). National Bureau of Economic Research.

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.

Curran, P. J., & Bauer, D. J. (2011). The disaggregation of within-person and between-person effects in longitudinal models of change. *Annual review of psychology*, 62, 583-619.

Cutler, D.M., and Lleras-Muney, A. (2010). Education and health: Evaluating theories and evidence. In R. Schoeni, J.S. House, G.A. Kaplan, and H. Pollack (Eds.), *Making Americans healthier* (pp. 29-60). New York: Russell Sage Foundation.

Davis-Kean, P. E. (2005). The influence of parental education and family income on child achievement: The direct role of parental expectations and the home environment. *Journal of Family Psychology*, 19, 294–304.

Davis-Kean, P. E., & Jager, J. (2014). Trajectories of achievement within race/ethnicity: "Catching up" in achievement across time. *The Journal of Educational Research*, *107*(3), 197-208.

Dawid, A. P. (2008). Beware of the DAG!. *Journal of Machine Learning Research: Workshop and Conference Proceeding* 6, 59–86.

Deaton, A., & Cartwright, N. (2018). Understanding and misunderstanding randomized controlled trials. *Social Science & Medicine*, *210*, 2-21.

Dehaene, S., Spelke, E., Pinel, P., Stanescu, R., & Tsivkin, S. (1999). Sources of mathematical thinking: Behavioral and brain-imaging evidence. *Science*, 284(5416), 970-974.

Diamond, A. (2013). Executive functions. *Annual review of psychology*, *64*, 135-168.

Diez Roux, A. V. (2004). Estimating the neighborhood health effects: The challenges of casual inference in a complex world. *Epidemiologic Reviews*, 26, 104–111

Domina, T., Penner, A., & Penner, E. (2017). Categorical inequality: Schools as sorting machines. *Annual review of sociology*, *43*, 311-330.

Downey, D. B., & Condron, D. J. (2016). Fifty years since the Coleman Report: Rethinking the relationship between schools and inequality. *Sociology of Education*, 89(3), 207-220.

Dubrow, J. K. (2008). How can we account for intersectionality in quantitative analysis of survey data? Empirical illustration for Central and Eastern Europe. *ASK. Research & Methods*, (17), 85-100.

Duncan, G. J., Dowsett, C. J., Claessens, A., Magnuson, K., Huston, et al. (2007). School readiness and later achievement. *Developmental Psychology*, 43, 1428-1446.

Eberhardt, F. (2017). Introduction to the foundations of causal discovery. *International Journal of Data Science and Analytics*, 3(2), 81-91.

Ebert-Uphoff, I., & Deng, Y. (2012). Causal discovery for climate research using graphical models. *Journal of Climate*, *25*(17), 5648-5665.

Elster, J. (1998). A plea for mechanisms. In P. Hedstrom & R. Swedberg (Eds.), *Social mechanisms: An analytical approach to social theory* (pp. 45-73). Cambridge , UK: Cambridge University Press.

Elwert, F. (2013). Graphical Causal Models. In S. L. Morgan (Ed.), *Handbook of causal analysis for social research* (pp. 245–273). New York: Springer.

Elwert, F., & Winship, C. (2014). Endogenous selection bias: The problem of conditioning on a collider variable. *Annual review of sociology*, *40*, 31-53.

Espelage, D. L. and Holt, M. (2001). Bullying and victimization during early adolescence: Peer influences and psychosocial correlates. *Journal of Emotional Abuse*, 2: 123–142.

Estabrook, R., Grimm, K. J., & Bowles, R. P. (2012). A Monte Carlo simulation study of the reliability of intraindividual variability. *Psychology and Aging*, 27(3), 560-576.

Falkenström, F., Finkel, S., Sandell, R., Rubel, J. A., & Holmqvist, R. (2017). Dynamic models of individual change in psychotherapy process research. *Journal of Consulting and Clinical Psychology*, 85(6), 537.

Feinstein, L. (2003). Inequality in the early cognitive development of British children in the 1970 cohort. *Economica*, *70*(277), 73-97.

Ferreira, F. H. and, J. (2011). The measurement of inequality of opportunity: Theory and an application to Latin America. *Review of Income and Wealth*, 57(4), 622–657.

Feuer, M. (2015). Evidence and Advocacy. In Feuer, M. J., Berman, A. I., & Atkinson, R. C. (Eds.). *Past as Prologue: The National Academy of Education at 50. Members Reflect* (pp.95-101). Washington, DC: National Academy of Education.

Fields, G. S. (2006). The many facets of economic mobility. In McGillivray, M. (Ed.), *Inequality, Poverty, and Well-Being* (pp. 123–142). Hampshire: Palgrave Macmillan.

Fields, G. S. (2010). Does income mobility equalize longer-term incomes? New measures of an old concept. *The Journal of Economic Inequality*, *8*(4), 409-427.

Firebaugh, G., Warner, C., & Massoglia, M. (2013). Fixed effects, random effects, and hybrid models for causal analysis. In S. L. Morgan (Ed.), *Handbook of causal analysis for social research* (pp. 113–132). Dordrecht, the Netherlands: Springer.

Fisher, A. J., Megdalia, J., & Jeronimus, B. F. (2018). A lack of group-to-individual generalizability is a threat to human subjects research. *Proceedings of the National Academy of Sciences of the United States of America* (PNAS), 115, E6106–E6115.

Follmer, D. J. (2018). Executive function and reading comprehension: A meta-analytic review. *Educational Psychologist*, *53*(1), 42-60.

Forgas, J. P. (2008). Affect and cognition. *Perspectives on psychological science*, *3*(2), 94-101.

Foster, E. M. (2010). The u-shaped relationship between complexity and usefulness: A commentary. *Developmental Psychology, 46*, 1760-1766.

Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The elements of statistical learning* (Vol. 1, No. 10). New York: Springer series in statistics.

Fryer Jr, R. G., & Levitt, S. D. (2004). Understanding the black-white test score gap in the first two years of school. *Review of Economics and Statistics*, *86*(2), 447-464.

Fuhs, M. W., Nesbitt, K. T., Farran, D. C., & Dong, N. (2014). Longitudinal associations between executive functioning and academic skills across content areas. *Developmental Psychology*, 50(6), 1698–1709.

Galindo, C., & Sheldon, S. B. (2012). School and home connections and children's kindergarten achievement gains: The mediating role of family involvement. *Early Childhood Research Quarterly*, 27(1), 90-103.

Giesselmann, M., & Schmidt, A. W. (2018). *Interactions in Fixed Effects Regression Models*. DIW discussion paper No. 1748. Berlin: German Institute for Economic Research (DIW).

Glymour, M. M., & Greenland, S. (2008). Causal diagrams. In K. J. Rothman, S. Greenland, & T. Lash (Eds.), *Modern epidemiology* (3rd ed., pp. 183–209). Philadelphia: Lippincott.

Glymour, C., Zhang, K., & Spirtes, P. (2019). Review of Causal Discovery Methods Based on Graphical Models. *Frontiers in Genetics*, 10, 1-15.

Greca, A. M. and Stone, W. L. (1993). Social anxiety scale for children—revised: Factor structure and concurrent validity. *Journal of Clinical Child Psychology*, 22(1): 17–27.

Greenland, S. (2010). Overthrowing the tyranny of null hypotheses hidden in causal diagrams. In R. Dechter, H. Geffner, & J. Y. Halpern (Eds.), *Heuristics, probability and causality: A tribute to Judea Pearl* (pp. 365–382). London: College Publications.

Gresham, F. M., & Elliott, S. N. (1990). *The social skills rating system*. Circle Pines, MN: American Guidance Service.

Grosz, M. P., Rohrer, J. M., & Thoemmes, F. (in press). The Taboo Against Explicit Causal Inference in Nonexperimental Psychology. *Perspectives on Psychological Science*.

Gutman, L. M., Sameroff, A. J., & Cole, R. (2003). Academic growth curve trajectories from 1st grade to 12th grade: Effects of multiple social risk factors and preschool child factors. *Developmental psychology*, *39*(4), 777.

Halaby, C. N. (2004). Panel models in sociological research: Theory into practice. *Annu. Rev. Sociol.*, *30*, 507-544.

Hainmueller, J., Mummolo, J., & Xu, Y. (2019). How much should we trust estimates from multiplicative interaction models? Simple tools to improve empirical practice. *Political Analysis*, *27*(2), 163-192.

Hair, N. L., Hanson, J. L., Wolfe, B. L., & Pollak, S. D. (2015). Association of child poverty, brain development, and academic achievement. *JAMA pediatrics*, *169*(9), 822-829.

Hamaker, E. L. (2012). Why researchers should think "within-person" a paradigmatic rationale. In M. R. Mehl & T. S. Conner (Eds.), *Handbook of research methods for studying daily life* (p. 43-61). New York, NY: Guilford Publications.

Hamaker, E. L., & Muthén, B. (2019). The fixed versus random effects debate and how it relates to centering in multilevel modeling. *Psychological methods*.

Hambleton, R .K., & Pitoniak, M. J. (2006). Setting performance standards. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 433–470). Westport, CT: American Council on Education and Praeger.

Hamilton, L.S. (2003). Assessment as a policy tool. *Review of Research in Education*, *27*, 25-68.

Hanson, J. L., Chung, M. K., Avants, B. B., Rudolph, K. D., Shirtcliff, E. A., Gee, J. C., ... & Pollak, S. D. (2012). Structural variations in prefrontal cortex mediate the relationship between early childhood stress and spatial working memory. *Journal of Neuroscience*, *32*(23), 7917-7925.

Hartemink, A. J. (2001), "Principled Computational Methods for the Validation of and Discovery of Genetic Regulatory Networks," unpublished doctoral thesis, Massachusetts Institute of Technology

Hattie, J. (2009). *Visible Learning: A Synthesis of Over 800 Meta-Analyses Relating to Achievement*. London: Routledge.

Heckman, J. J. (2005). The scientific model of causality. *Sociological methodology*, *35*(1), 1-97.

Hedström, P., & Swedberg, R. (1998). Social mechanisms: An introductory essay. In P. Hedström & R. Swedberg (Eds.), *Social Mechanisms: An Analytical Approach to Social Theory* (pp. 1-31). Cambridge, UK: Cambridge University Press.

Hedström, P., & Ylikoski, P. (2010). Causal mechanisms in the social sciences. *Annual review of sociology*, *36*, 49-67.

Heinze-Deml, C., Maathuis, M. H., & Meinshausen, N. (2018). Causal structure learning. *Annual Review of Statistics and Its Application*, *5*, 371-391.

Hernán, M. A., Hernández-Díaz, S., Werler, M. M., & Mitchell, A. A. (2002). Causal knowledge as a prerequisite for confounding evaluation: an application to birth defects epidemiology. *American journal of epidemiology*, 155(2), 176-184.

Hernán, M. A. (2018). The C-word: scientific euphemisms do not improve causal inference from observational data. *American journal of public health*, *108*(5), 616-619.

Hill, C. J., Bloom, H. S., Black, A. R., & Lipsey, M. W. (2008). Empirical benchmarks for interpreting effect sizes in research. *Child Development Perspectives*, *2*(3), 172-177.

Hitchcock, C. (2018). Probabilistic Causation. In *The Stanford Encyclopedia of Philosophy*, edited by E. Zalta. Retrieved October 25, 2019 (https://plato.stanford.edu/entries/causation-probabilistic/)

Ho, A. D., & Haertel, E. H. (2006). Metric-Free Measures of Test Score Trends and Gaps with Policy-Relevant Examples. CSE Report 665. *National Center for Research on Evaluation, Standards, and Student Testing (CRESST)*. Retrieved August 7, 2019, from https://cresst.org/wp-content/uploads/R665.pdf

Hoffman, L. (2015). *Longitudinal analysis: Modeling within person fluctuation and change*. New York, NY: Routledge.

Hoffman, L., & Stawski, R. S. (2009). Persons as contexts: Evaluating between-person and within-person effects in longitudinal analysis. *Research in Human Development*, *6*(2-3), 97-120.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American statistical Association*, *81*(396), 945-960.

Hoxby, C. (2000). *Peer effects in the classroom: Learning from gender and race variation*. NBER Working Paper No. 7867. Cambridge, MA: National Bureau of Economic Research.

Huang, C. (2011). Self-concept and academic achievement: A meta-analysis of longitudinal relations. *Journal of School Psychology*, 49, 505–528.

Hulleman, C. S., Schrager, S. M., Bodmann, S. M., & Harackiewicz, J. M. (2010). A meta-analytic review of achievement goal measures: Different labels for the same constructs or different constructs with similar labels?. *Psychological bulletin*, *136*(3), 422.

Huque, M. H., Carlin, J. B., Simpson, J. A., & Lee, K. J. (2018). A comparison of multiple imputation methods for missing data in longitudinal studies. *BMC medical research methodology*, *18*(1), 168.

Imai, K., Keele, L., Tingley, D., & Yamamoto, T. (2011). Unpacking the black box of causality: Learning about causal mechanisms from experimental and observational studies. *American Political Science Review*, 105, 765–789.

Imai, K., & Kim, I. S. (2019). When should we use unit fixed effects regression models for for causal inference with longitudinal data? *American Journal of Political Science*. Advance online publication. doi:10.1111/ajps.12417.

Imbens, G. (2019). *Potential outcome and directed acyclic graph approaches to causality: Relevance for empirical practice in economics* (No. w26104). National Bureau of Economic Research.

Jacob, R., & Parkinson, J. (2015). The potential for school-based interventions that target executive function to improve academic achievement: A review. *Review of educational research*, *85*(4), 512-552.

Jahng, S., Wood, P. K., & Trull, T. J. (2008). Analysis of affective instability in ecological momentary assessment: Indices using successive difference and group comparison via multilevel modeling. *Psychological methods*, *13*(4), 354-375.

Jäntti,M & Jenkins, S. (2013). *Income Mobility*. IZA Discussion Paper, No. 7730.

Janzing, D., & Scholkopf, B. (2010). Causal inference using the algorithmic Markov condition. *IEEE Transactions on Information Theory*, *56*(10), 5168-5194.

Peters, J., Janzing, D., & Schölkopf, B. (2017). *Elements of causal inference: foundations and learning algorithms*. MIT press.

Jencks, C. (1988). Whom must we treat equally for educational opportunity to be equal?. *Ethics*, 98, 518–533.

Kalisch, M., Fellinghauer, B. A., Grill, E., Maathuis, M. H., Mansmann, U., Bühlmann, P., & Stucki, G. (2010). Understanding human functioning using graphical models. *BMC Medical Research Methodology*, *10*(1), 14.

Kalisch, M., & Bühlmann, P. (2014). Causal structure learning and inference: a selective review. *Quality Technology & Quantitative Management*, *11*(1), 3-21.

Kanbur, R. & Wagstaff, A. (2016). How useful is inequality of opportunity as a policy construct?. In K. Basu and J.E. Stiglitz (Eds.), *Inequality and Growth: Patterns and Policy, Vol. I: Concepts and Analysis* (pp.131-148). London: Palgrave Macmillan.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*(1), 1-73.

Kenny, D. A., & Zautra, A. (1995). The trait-state-error model for multiwave data. *Journal of consulting and clinical psychology*, *63*(1), 52-59.

Kenny, D. A. & Zautra, A. (2001). Trait-state models for longitudinal data. In L.M. Collins & A.G. Sayer (Eds.). *New methods for the analysis of change: Decade of behavior*. (pp. 243-263). Washington, DC: American Psychological Association.

Kim, Y., & Steiner, P. M. (2019a). Gain scores revisited: A graphical models perspective. *Sociological Methods & Research*. Advance online publication. doi:10.1177/0049124119826155

Kim, Y., & Steiner, P. M. (2019b). Causal Graphical Views of Fixed Effects and Random Effects Models. *Sociological Methods & Research*. Advance online publication. doi: 10.31234/osf.io/cxd2n.

Kerckhoff, A. C. (1993). *Diverging pathways: Social structure and career deflections*. New York: Cambridge University Press.

Knight, C. R., & Winship, C. (2013). The causal implications of mechanistic thinking: Identifi cation using directed acyclic graphs. In S. L. Morgan (Ed.), *Handbook of causal analysis for social research* (pp. 275–300). New York: Springer.

Koechlin, E., & Summerfield, C. (2007). An information theoretical approach to prefrontal executive function. *Trends in cognitive sciences*, *11*(6), 229-235.

Koretz, D. (2017). *The testing charade: Pretending to make schools better*. University of Chicago Press.

Kruglanski, A. W., Shah, J. Y., Fishbach, A., Friedman, R., Chun, W. Y., & Sleeth-Keppler, D. (2002). A theory of goal-systems. *Advances in Experimental Social Psychology*, 34, 311–378.

Kuhn, M., & Johnson, K. (2013). *Applied Predictive Modeling*. New York: Springer.

Lareau, A. (2003). *Unequal Childhoods: Class Race and Family Life*. Berkeley: University of California Press.

Lazowski, R. A., & Hulleman, C. S. (2016). Motivation interventions in education: A meta-analytic review. *Review of Educational research*, *86*(2), 602-640.

Leszczensky, L., & Wolbring, T. (2019). How to deal with reverse causality using panel data? recommendations for researchers based on a simulation study. *Sociological Methods & Research*. doi: 10.17605/OSF.IO/8XB4Z

Levy, F., and Murnane, R.J. (2004). *The new division of labor: How computers are creating the next job marke*t. Princeton, NJ: Princeton University Press.

Lieberson, S. (1987). *Making It Count: The Improvement of Social Research and Theory*. Berkeley: University of California Press.

Liu, H., Lafferty, J., & Wasserman, L. (2009). The nonparanormal: Semiparametric estimation of high dimensional undirected graphs. *Journal of Machine Learning Research*, 10(Oct), 2295-2328.

Loeb, S., Dynarski, S., McFarland, D., Morris, P., & Reardon, S. (2017). *Descriptive analysis in education: A guide for researchers* (NCEE 2017–4023). Washington, DC: U.S.

Lowenstein, A.E., Friedman-Krauss, A.H., Raver, C.C., Jones, S.M., & Pess, R.A. (2016). School climate, teacher-child closeness, and low-income children's academic skills in kindergarten. *Journal of Educational and Developmental Psychology*, 5(2), 89-108.

Lucas, S. R. (2001). Effectively maintained inequality: Education transitions, track mobility, and social background effects. *American journal of sociology*, 106(6), 1642-1690.

Lucas, S. R., & Berends, M. (2002). Sociodemographic diversity, correlated achievement, and de facto tracking. *Sociology of Education*, 75(4), 328-348.

Machamer, P., Darden, L., & Craver, C. F. (2000). Thinking about mechanisms. *Philosophy of science*, 67(1), 1-25.

Malinsky, D., & Danks, D. (2018). Causal discovery algorithms: A practical guide. *Philosophy Compass*, *13*(1), e12470.

Marsh, H. W., Barnes, J., Cairns, L., & Tidman, M. (1984). Self-description questionnaire: Age and sex effects in the structure and level of self-concept for preadolescent children. *Journal of Educational Psychology*, 83, 377–392.

Marsh, H. W., Trautwein, U., Lüdtke, O., Köller, O., & Baumert, J. (2005). Academic self-concept, interest, grades, and standardized test scores: Reciprocal effects models of causal ordering. *Child development*, *76*(2), 397-416.

May, K., & Nicewander, W. A. (1994). Reliability and information functions for percentile ranks. *Journal of Educational Measurement*, 31(4), 313-325.

Mazumder, B. (2005). Fortunate Sons: New Estimates of Intergenerational Mobility in the United States Using Social Security Earnings Data. *Review of Economics and Statistics*, 87 (2), 235–55.

McCandliss, B. D., Cohen, L., & Dehaene, S. (2003). The visual word form area: expertise for reading in the fusiform gyrus. *Trends in cognitive sciences*, *7*(7), 293-299.

McDonough, I. K. (2015). Dynamics of the black–white gap in academic achievement. *Economics of Education Review*, 47, 17-33.

McKinnon, R. D., & Blair, C. (2019). Bidirectional relations among executive function, teacher–child relationships, and early reading and math achievement: A cross-lagged panel analysis. *Early Childhood Research Quarterly*, *46*, 152-165.

McKown, C. (2013). Social equity theory and racial-ethnic achievement gaps. *Child Development*, *84*(4), 1120-1136.

McNeish, D., & Kelley, K. (2019). Fixed effects models versus mixed effects models for clustered data: Reviewing the approaches, disentangling the differences, and making recommendations. *Psychological Methods*, *24*(1), 20-35.

Milam, A. J., Furr-Holden, C. D. M., & Leaf, P. J. (2010). Perceived school and neighborhood safety, neighborhood violence and academic achievement in urban school children. *The Urban Review*, *42*(5), 458-467.

Miranda, A., Colomer, C., Mercader, J., Fernández, M. I., & Presentación, M. J. (2015). Performance-based tests versus behavioral ratings in the assessment of executive functioning in preschoolers: associations with ADHD symptoms and reading achievement. *Frontiers in psychology*, 6, 545.

Molenaar, P. C. (2004). A manifesto on psychology as idiographic science: Bringing the person back into scientific psychology, this time forever. *Measurement*, *2*(4), 201-218.

Morgan, S. L., & Winship, C. (2015). *Counterfactuals and causal inference: Methods and principles for social research*. Cambridge, UK: Cambridge University Press.

Murnane, R. J., & Willett, J. B. (2010). *Methods matter: Improving causal inference in educational and social science research*. New York, NY: Oxford University Press.

National Research Council. (2012). *Education for life and work: Developing transferable knowledge and skills for the 21st century*. Washington, DC: The National Academies Press.

Nesselroade, J. R., & Molenaar, P. C. M. (2010). Emphasizing Intraindividual Variability in the Study of Development Over the Life Span: Concepts and Issues. In W. F. Overton (Ed.), & R. M. Lerner (Editor- in-Chief). *Handbook of life-span development*, Vol. 1: Cognition, biology, and methods across the lifespan (pp. 30–54). Hoboken, NJ: Wiley.

Newsom, J. T. (2015). *Longitudinal structural equation modeling: A comprehensive introduction*. New York: Routledge.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational researcher*, *23*(2), 13-23.

Miranda, A., Colomer, C., Mercader, J., Fernández, M. I., & Presentación, M. J. (2015). Performance-based tests versus behavioral ratings in the assessment of executive functioning in preschoolers: associations with ADHD symptoms and reading achievement. *Frontiers in psychology* 6, 545.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). Focus article: On the structure of educational assessments. *Measurement: Interdisciplinary research and perspectives*, *1*(1), 3-62.

Molenaar, P. C. (2004). A manifesto on psychology as idiographic science: Bringing the person back into scientific psychology, this time forever. *Measurement*, 2(4), 201-218.

Muthén, B. (2004). Latent variable analysis: Growth mixture modeling and related techniques for longitudinal data. In D. Kaplan (Ed.), Handbook of Quantitative Methodology for the Social Sciences (pp. 345–368). Newbury Park, CA: Sage Publications.

Nagin, D. S. (1999). Analyzing developmental trajectories: a semiparametric, group-based approach. *Psychological methods*, *4*(2), 139-152

Nagin, D. S. (2005). *Group-based modeling of development*. Cambridge, MA: Harvard University Press.

Najarian, M., Tourangeau, K., Nord, C., and Wallner-Allen, K. (2018). *Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011), First- and Second-Grade Psychometric Report (NCES 2018-183)*. National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC.

Nguyen, C. D., Carlin, J. B., & Lee, K. J. (2017). Model checking in multiple imputation: an overview and case study. *Emerging themes in epidemiology*, *14*(1), 8.

Nikolaev, B. (2016). Does other people's education make us less happy?. *Economics of Education Review*, *52*, 176-191.

OECD (2019), *OECD Skills Strategy 2019: Skills to Shape a Better Future*, OECD Publishing, Paris, https://doi.org/10.1787/9789264313835-en.

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

Pallas, A. M. (2000). The effects of schooling on individual lives. In M. T. Hallinan (Ed.), *Handbook of the sociology of education* (pp. 499–525). New York: Kluwer Academic/ Plenum Publishers.

Pearl J. (2009). *Causality: Models, Reasoning, and Inference.* New York: Cambridge Univ. Press. 2nd ed.

Pearl, J. (2011). Invited commentary: understanding bias amplification. *American journal of epidemiology*, *174*(11), 1223-1227.

Pearl, J. (2019). The seven tools of causal inference, with reflections on machine learning. *Communications of the ACM*, *62*(3), 54-60.

Pellet, J.P. & Elisseeff, A. (2008). Using Markov blankets for causal structure learning. *Journal of Machine Learning Research*, 9(Jul), 1295-1342.

Peters, J., Bühlmann, P., & Meinshausen, N. (2016). Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 78*(5), 947-1012.

Pfost, M., Hattie, J., Dörfler, T., & Artelt, C. (2014). Individual differences in reading development: A review of 25 years of empirical research on Matthew effects in reading. *Review of Educational Research, 84*(2), 203-244.

Phelps, E. A. (2006). Emotion and cognition: insights from studies of the human amygdala. *Annu. Rev. Psychol*., 57, 27-53.

Pianta, R. C. (2001).*Student-teacher relationship scale.* Odessa, FL: Psychological Assessment Resources.

Putnam, H. (2002). *The collapse of the fact/value dichotomy and other essays*. Cambridge, MA: Harvard University Press.

Quinlan, T., Higgins, D., & Wolff, S. (2009). Evaluating the construct-coverage of the e-rater® scoring engine. *ETS Research Report Series*, 2009(1), i-35.

Quintana, R. (in press). The Structure of Academic Achievement: Searching for Proximal Mechanisms Using Causal Discovery Algorithms. *Sociological Methods & Research.*

Quintana, R. and Correnti, R. (2020). The Concept of Academic Mobility: Normative and Methodological Considerations. *American Educational Research Journal*, 20(10), 1-40.

Ramsey, J. D. (2014). A scalable conditional independence test for nonlinear, non-Gaussian data. arXiv preprint arXiv:1401.5031.

Ramsey, J. D., Sanchez-Romero, R., & Glymour, C. (2014). Non-Gaussian methods and high-pass filters in the estimation of effective connections. *Neuroimage, 84*, 986-1006.

Ramsey, J., Glymour, M., Sanchez-Romero, R., & Glymour, C. (2017). A million variables and more: the Fast Greedy Equivalence Search algorithm for learning high-dimensional graphical causal models, with an application to functional magnetic resonance images. *International journal of data science and analytics*, *3*(2), 121-129.

Reardon, S. F. (2008a). *Thirteen ways of looking at the Black-White test score gap* (Working Paper No. 2008-08). Stanford, CA: Institute for Research on Educational Policy and Practice, Stanford University.

Reardon, S. F. (2008b). *Differential growth in the Black-White achievement gap during elementary school among initially high- and low-scoring students*. Working Paper Series. Stanford, CA: Institute for Research on Educational Policy and Practice, Stanford University.

Reardon, S. F., & Galindo, C. (2009). The Hispanic-White achievement gap in math and reading in the elementary grades. *American Educational Research Journal*, *46*(3), 853-891.

Reardon, S. F., Robinson-Cimpian, J. P., & Weathers, E. S. (2015). Patterns and trends in racial/ethnic and socioeconomic academic achievement gaps. In H. Ladd & M. Goertz (Eds.), *Handbook of research in education finance and policy* (2nd ed., pp. 491–509). Mahwah, NJ: Erlbaum.

Reeves, R. (2015). The Measure of a Nation. *The Annals of the American Academy of Political and Social Science*, 657(1), 22-26.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, *66*(5), 688.

Rutter, M. (2007). Proceeding from observed correlation to causal inference: The use of natural experiments. *Perspectives on Psychological Science*, *2*(4), 377-395.

Sachs, K., Perez, O., Pe'er, D., Lauffenburger, D. A., & Nolan, G. P. (2005). Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, *308*(5721), 523-529.

Salsman, J. M., Butt, Z., Pilkonis, P. A., Cyranowski, J. M., Zill, N., Hendrie, H. C., ... & Lai, J. S. (2013). Emotion assessment using the NIH Toolbox. *Neurology*, *80*(11 Supplement 3), S76-S86.

Sanchez-Romero, R., Ramsey, J. D., Zhang, K., Glymour, M. R., Huang, B., & Glymour, C. (2019). Estimating feedforward and feedback effective connections from fMRI time series: Assessments of statistical methods. *Network Neuroscience*, *3*(2), 274-306.

Scheines, R. (1997). An introduction to causal inference. Pp. 185-199 in *Causality in crisis? Statistical methods and the search for causal knowledge in the social sciences*, edited by V.R. McKim and S.P. Turner. Notre Dame, IN: Notre Dame University Press.

Schmitt, S. A., Geldhof, G. J., Purpura, D. J., Duncan, R., & Mcclelland, M. M. (2017). Examining the relations between executive function, math, and literacy during the transition to kindergarten: A multi-analytic approach. *Journal of Educational Psychology*, 109, 1120.

Schumpeter, J. (1949). Science and ideology. *American Economic Review*, 39, 345-359.

Schunck, R. (2013). Within and between estimates in random-effects models: Advantages and drawbacks of correlated random effects and hybrid models. *The Stata Journal*, *13*(1), 65-76.

Shanley, L. (2016). Evaluating Longitudinal Mathematics Achievement Growth: Modeling and Measurement Considerations for Assessing Academic Progress. *Educational Researcher* , 20 (10), 1-11.

Shepard, L. A. (1997). The centrality of test use and consequences for test validity. *Educational Measurement: Issues and Practice*, *16*(2), 5-24.

Shepard, L., Hannaway, J., & Baker, E. Editors (2009). *Standards, assessments, and accountability*. Washington, DC: National Academy of Education.

Shmueli, G. (2010). To explain or to predict?. *Statistical science*, 25(3), 289-310.

Simonds, J. & Rothbart, M. K. (2004). *The Temperament in Middle Childhood Questionnaire (TMCQ): A computerized self-report measure of temperament for ages 7–10*. Poster session presented at the occasional temperament conference, Athens, GA.

Slavin, R. (2018). John Hattie is Wrong [Blog post]. Retrieved from https://robertslavinsblog.wordpress.com/2018/06/21/john-hattie-is-wrong/

Smithers, L. G., Sawyer, A. C., Chittleborough, C. R., Davies, N. M., Smith, G. D., & Lynch, J. W. (2018). A systematic review and meta-analysis of effects of early life non-cognitive skills on academic, psychosocial, cognitive and health outcomes. *Nature human behaviour*, *2*(11), 867.

Sohn, K. (2012). The dynamics of the evolution of the Black–White test score gap. *Education Economics*, 20(2), 175-188.

Solnick, S. J., & Hemenway, D. (1998). Is more always better?: A survey on positional concerns. *Journal of Economic Behavior & Organization*, *37*(3), 373-383.

Spirtes, P. (2010). Introduction to causal inference. *Journal of Machine Learning Research*, *11*(May), 1643-1662.

Spirtes P., Meek C. & Richardson, TS. (1995) Causal inference in the presence of latent variables and selection bias. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence* (pp. 296–305). San Francisco: Morgan Kaufmann.

Spirtes, P., Glymour, C. N., Scheines, R., Heckerman, D., Meek, C., Cooper, G., & Richardson, T. (2000). *Causation, Prediction, and Search*. MIT press.

Spirtes, P., & Scheines, R. (2004). Causal inference of ambiguous manipulations. *Philosophy of Science*, *71*(5), 833-845.

Spirtes, P., & Zhang, K. (2016). Causal discovery and inference: Concepts and recent methodological advances. *Applied Informatics*, 3, 1–28.

Stanovich, K. E. (1986). Matthew effects in reading: Some consequences of individual differences in the acquisition of literacy. *Reading research quarterly*, 360-407.

Steiner, P. M., Kim, Y., Hall, C. E., & Su, D. (2017). Graphical models for quasi-experimental designs. *Sociological methods & research*, *46*(2), 155-188.

Sun, Y., & Li, Y. (2009). Parental divorce, sibship size, family resources, and children's academic performance. *Social Science Research*, *38*(3), 622-634.

Susperreguy, M. I., Davis-Kean, P. E., Duckworth, K., & Chen, M. (2018). Self-Concept Predicts Academic Achievement Across Levels of the Achievement Distribution: Domain Specificity for Math and Reading. *Child development*, *89*(6), 2196-2214.

Raudenbush, S. W. (2001). Comparing personal trajectories and drawing causal inferences from longitudinal data. *Annual review of psychology*, *52*(1), 501-525.

Rawls, J. (1971) *A Theory of Justice*. Cambridge, MA: Harvard University Press.

Roemer, J. (1998). *Equality of Opportunity.* Cambridge, MA: Harvard University Press.

Roemer, J. E., & Trannoy, A. (2015). Equality of opportunity. In A.B. Atkinson and F. Bourguignon (eds) *Handbook of income distribution* (Vol. 2, pp. 217-300). Amsterdam: North Holland.

Sen, A. (2009). *The Idea of Justice*. London: Allen Lane.

Stanovich, K. E. (1986). Matthew effects in reading: Some consequences of individual differences in the acquisition of literacy. *Reading research quarterly*, 360-407.

Thomson, J. J. (2008). *Normativity*. Chicago: Open Court.

Tourangeau, K., Nord, C., Leˆ, T., Sorongon, A. G., Najarian, M., & Hausken, E. G. (2009). *Early childhood longitudinal study, kindergarten class of 1998-99 (ECLS-K) combined user's manual for the ECLS-K Eighth-Grade and K-8 full sample data files and electronic codebook*. Washington, DC: U.S. Department of Education.

Usami, S., Murayama, K., & Hamaker, E. L. (2019). A unified framework of longitudinal models to examine reciprocal relations. *Psychological Methods*.

Vanderweele, T. J. (2015). *Explanation in causal inference: Methods for mediation and interaction.* New York, NY: Oxford University Press.

VanderWeele, T. J., & Shpitser, I. (2011). A new criterion for confounder selection. *Biometrics*, 67(4), 1406-1413.

van Lier, P. A., Vitaro, F., Barker, E. D., Brendgen, M., Tremblay, R. E., & Boivin, M. (2012). Peer victimization, poor academic achievement, and the link between childhood externalizing and internalizing problems. *Child development*, *83*(5), 1775-1788.

Van de Werfhorst, H. G., & Mijs, J. J. (2010). Achievement inequality and the institutional structure of educational systems: A comparative perspective. *Annual review of sociology*, 36, 407-428.

Wang, L., & Grimm, K. J. (2012). Investigating reliabilities of intraindividual variability indicators. *Multivariate Behavioral Research*, 47(5), 771-802.

Wang, L. P., Hamaker, E., & Bergeman, C. S. (2012). Investigating inter-individual differences in short-term intra-individual variability. *Psychological Methods*, *17*(4), 567-58.

Watts, T. W., Duncan, G. J., Chen, M., Claessens, A., Davis-Kean, P. E., Duckworth, K., ... & Susperreguy, M. I. (2015). The role of mediators in the development of longitudinal mathematics achievement associations. *Child development*, *86*(6), 1892-1907.

Wodtke, G. T., Harding, D. J., & Elwert, F. (2011). Neighborhood effects in temporal perspective: The impact of long-term exposure to concentrated disadvantage on high school graduation. *American Sociological Review*, *76*(5), 713-736.

Woodcock, R., McGrew, K., & Mather, N. (2001). *Woodcock-Johnson III*. Itasca, IL: Riverside.

Woodward, J. (2002). What is a mechanism? A counterfactual account. *Philosophy of Science*, *69*(S3), S366-S377.

York, R., & Light, R. (2017). Directional Asymmetry in Sociological Analyses. *Socius***,** 3, 1–13.

Zelazo, P. D. (2006). The Dimensional Change Card Sort (DCCS): A method of assessing executive function in children. *Nature protocols*, *1*(1), 297.

Zyphur, M. J., Allison, P. D., Tay, L., Voelkle, M. C., Preacher, K. J., Zhang, Z., Hamaker, E., Shamsollahi, A., Pierides, D, Koval, P., & Diener, E. (2019). From data to causes I: Building a general crosslagged panel model (GCLM). *Organizational Research Methods*, 20(10), 1-37