Essays on the Economics of Education

Policies for Academic Transitions

by

Andreas Østbø Fidjeland

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

The PhD Program in Social Sciences
UiS Business School
2022

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

© 2022 Andreas Østbø Fidjeland

ISSN: 1890-1387
ISBN: 978-82-8439-053-6
Dr. avh nr 621

To Benjamin and Tobias

Acknowledgements

It takes a village to write a PhD. As such, the realization of this thesis is equally the result of effort on the part of numerous people other than myself. To all of them I owe a great deal of gratitude. First and foremost, I would never have made it to this point without the support, encouragement, and mentorship provided by Mari Rege and Ingeborg F. Solli. It is not every doctoral student's privilege to be given the opportunity to learn from such excellent researchers and people. Their genuine engagement and passion gave my projects a spark from the very beginning, and instilled in me the belief that I have ideas and talents worth pursuing. I do not take for granted how there has always been time to answer my questions — big and small — and provide detailed, constructive feedback on my work. Their example have not only taught me how to become a good researcher, but also how to be a good academic citizen. I will do my best to follow that lead in all my academic endeavours.

I am thankful to my third supervisor Eric Bettinger, not only for providing mentoring and co-authorship, but also for graciously hosting me at Stanford during my research stay. The semester in California was the experience of a lifetime, and I am truly grateful for the opportunity to learn hands-on from a world-class scholar.

As part of my doctoral research, I have been privileged to be affiliated with the research projects *Agderprosjektet* and *Lekbasert Læring*. Many people have worked hard over many years in order to execute these field experiments, not the least the project leaders Mari Rege and Ingunn Størksen. Their efforts not only provided me with superb experimental data, which serves as the basis for Chapter 3 of this thesis, but also pro-

vided me with many learning opportunities and allowed me to take part in a high-quality research project. Thank you to all the team members, researchers, administrators, children, and childcare center staff who made the projects possible. Thank you also in this regard to Åse Lea, for providing excellent co-ordination and administrative support for the project, as well as providing me with help and solutions across a wide range of practical issues.

A great deal of gratitude is owed to all my colleagues at the UiS Business School, and in particular the PhD community, from which I've not only learned immensely, but also made many friends. Thank you in particular to all the participants at the 9 am Coffee and the PhD Brown Bag. A special thanks to Max, with whom I've walked in tandem throughout this PhD process. I appreciate all our trips, conversations, coffees and beers, all of which have greatly improved the PhD experience. Thank you also to Nur for all the support from day 1, our many great conversations, and for sharing with me many great moments, both at, and outside of, work. Thank you to all the old PhDs who were there to welcome and include me into their group as a fresh doctoral student. From this group I am particularly thankful to May Linn for showing me the ropes, and allowing me to peak at/copy her work from time to time. Being able to follow her example greatly reduced the uncertainty and frictions of navigating the PhD experience.

Many people have generously used of their time to comment on my work. Edwin Leuven and Hans H. Sievertsen provided excellent opposition at my 50% and 90% seminars, with thorough assessments of my entire body of work. Their efforts exceeded what could be expected from the role, to which I am thankful. I am grateful to Ingunn Størksen for taking great care in explaining the theoretical underpinnings of *Agderprosjektet*, providing me with relevant literature, and carefully reviewing our manuscript for Essay II. Thank you to Tom Dee for many thoughtful suggestions for my first essay, and for without prompting finding and providing me with relevant literature. Thank you also to the many seminar and conference participants whom have engaged with my papers over the years, asking tough and constructive questions that have greatly improved

the quality of my work.

I am forever grateful to my family, Gerd, Tor Olav, and Sigrid, for providing me with the values, mindset and tools necessary to achieve my academic ambitious. They instilled in me the belief that doing well in school is cool, and that learning is fun. Thank you for supporting me in whatever choices I have made, even in my refusal to get a real job.

Both my family and my in-laws, Mai and Steinar, also deserve a great deal of thanks for helping me balance my home- and worklife. Having two children in the span of two years is a daunting task in its own right, not the least in combination with a PhD program. Thank you all for everything you do for our family, so that we can solve the logistics of everyday life.

Of course, the greatest thanks is owed to Ann Karin. At our wedding I claimed that I wouldn't have achieved any of the accomplishments I'm most proud of if it wasn't for her. For none is it more true than this thesis. Her never-ending love, support and encouragement is an invaluable force in my life, both at home and at work. We make the best team.

In conclusion, I also want to thank Benjamin and Tobias for reminding me why this research matters in the first place, and for giving me all the best reasons to stop working and go home.

Andreas Fidjeland
Ullandhaug
October, 2021

My doctoral research was funded by the Norwegian Research Council through the project Lekbasert Læring, grant number. 270703. The Research Council also funded my research stay at Stanford with a Personal Overseas Grant, grant number 290675. Both are gratefully acknowledged.

Contents

Chapter 1 – Introduction

1 Overview

This doctoral thesis builds on a rich literature investigating how education policy affects students' learning, motivation, investment, and decision-making — all of which are determinants of the productivity of education systems. Over the past decades, the education field has yielded one of the most prolific strands of literature within applied economics research (Machin, 2014). In part this reflects a growing demand for an evidence-based design of education policy. Rigorous and thoughtful economic research can often produce such evidence, which may guide policymakers in the policy-design process (Hanushek et al., 2016).

Policy questions are ubiquitous in the education domain. In particular, many dimensions of a child's environment in school are determined by policymakers, ranging from the small and specific (such as the number of students in each class or the books used) to the large and general (such as the length of compulsory education, financing, and tracking). Another prevalent structural feature of the schooling process determined by policymakers are the transitions from one educational stage to the next. These milestone moments not only involve the replacement of one set of education policies by another, but have evolved into rites of passage in children's lives, signifying the end of one stage of development and the beginning of the next (Bharara, 2020; Evans et al., 2018).

Like more traditional rites of passage, these academic transitions are often costly. Because of the institutional discontinuities they represent, they are disruptive and challenging for many students (Anderson et al., 2000; Curson et al., 2019; Rice et al., 2015; Rice, 2001; Symonds and Galton, 2014), forcing them to navigate a new educational context that often involves a new school, new peers, and new teachers. Further, at each new stage, students not only face new and challenging academic demands but also heightened expectations of their independence and ability to assume responsibility for their own schooling. Not surprisingly, these transitions represent a period of particular vulnerability for many young people. An extensive research literature has consistently found associations with neg-

ative outcomes such as a decline in academic engagement and motivation, a decline in grades, and an increased risk of dropout (see, e.g., Bharara, 2020; Eccles et al., 1993; Evans et al., 2018; Galton et al., 1999 or Mizelle and Irvin, 2000).[1] Because the number, timing, and structure of transitions are all the result of policy, and imposed on students by policymakers, there is a need for a solid base of evidence—particularly causal—on how students navigate and prepare for them that can inform policy design so as to minimize the negative outcomes associated with those transitions (Rice, 2001; van Rens et al., 2018).

My aim for the thesis is to contribute to that evidence base. Empirical studies, such as those in the following chapters, can provide insights for policy on how best to prepare students for transitions, and how best to support them in making well-informed choices. For example, ensuring that students are adequately prepared for subsequent stages of schooling is an important step in making the education system more efficient and productive. Understanding how children and adolescents make investments and choices in their schooling can help policymakers identify areas where interventions might reduce inequalities in (opportunities for) human capital accumulation. Indeed, support and preparedness have been identified in the education literature as key elements for effective transitions (Anderson et al., 2000; Bharara, 2020).

I start, in Essay I, by investigating how students may respond to the implicit incentives associated with stage transitions in cases where the transition involves a transfer to a new school, and where enrollment in specific schools is based on merit. In fact, having adequate academic abilities is vital for successfully transitioning to more advanced stages of schooling (Anderson et al., 2000). However, students often report faltering motivation and engagement in school as they enter adolescence (Eccles et al., 1993; Harter et al., 1992). A fundamental tenet of economic theory is that we respond to incentives (Fehr and Falk, 2002). Policymakers concerned with poor effort and motivation among students might therefore consider rewarding those who perform well, so as to stimulate a more optimal level of investment in schoolwork. There is indeed ample evidence

[1] I will discuss these transitions in greater detail in Section 3.

for this type of response for older students, but we know very little about whether young students respond similarly to such incentives (Bach and Fischer, 2020). Not only are the benefits of schooling are less tangible for teenagers because of the long-term nature of the pay-offs, the skills necessary for implementing their preferred decisions, such as attention and impulse control, may not be sufficiently developed in adolescence (List et al., 2021). I test the validity of the hypothesis that young teenagers in Norwegian middle school[2] will respond to incentives by exploiting reforms that caused high-school enrollment schemes to change from being strictly based on neighborhood catchment areas to being based on merit in the form of middle school grades. I find that teenage students do increase their performance on high-stakes exams in response to such incentives. Also, ability assessments suggest that the increase in performance reflects actual learning and so is relevant for human capital accumulation. Hence, my study contributes causal evidence that policymakers are indeed able to influence the level of young students' investment in school by providing them with proper incentives.

In the second essay I take a step back to early childhood to investigate gender differences in pre-academic skills among children on the cusp of formal schooling. Building on an established literature on the importance of *school readiness*, my co-authors Mari Rege, Ingeborg Solli, Ingunn Størksen and I demonstrate that girls score substantially better than boys on measures of early learning. This implies that boys enter school at a significant skill disadvantage to girls.

Policymakers routinely express particular concern for boys in the transition from childcare to formal schooling (Husain and Millimet, 2009). Generally, this concern centers on boys being perceived as relatively less "mature", and less ready for the demands of school. In addition, they are perceived as having less-developed academic and socioemotional skills than girls at similar ages (DiPrete and Jennings, 2012; Lenes et al., 2020; Stipek, 2012). In the essay I report on results from an intervention in a sample of Norwegian preschools where we introduced more structured

[2]By "middle school" I refer to grades 8–10 of Norwegian compulsory school, which roughly equates to lower secondary school in many countries.

learning activities to be carried out with the children by trained teachers. While the goal of the project as a whole was to test the efficacy of this curriculum in improving school readiness, this particular study focuses on differential benefits across gender. Although many countries are now pushing toward universal provision of early childhood education, we know very little about whether existing universal programs have a heterogeneous impact across child subgroups when it comes to preparing them for later learning (Duncan and Magnuson, 2013; Phillips et al., 2017). Since expanding equal opportunities to succeed in the transition to school is often stated as an explicit policy objective underpinning such universal provision (Havnes and Mogstad, 2015; Heckman, 2006), we need better evidence of how curricular design interacts with child characteristics. In our study, we find that the introduction of more structured activities targeting important school-readiness skills was particularly beneficial for boys. Hence, our intervention was successful in reducing the substantial skill gap between boys and girls, which remained stable in the control group over the sample period. This suggests that careful, evidence-based curricular design and pedagogical practice can contribute to ensuring that children transitioning from childcare to formal schooling will do so on a more level playing field.

In the final essay, my co-author Eric Bettinger and I move to the other end of the education system to consider the transition into higher education in the United States. A college degree can be a major driver of social mobility, with a far-reaching impact on the life trajectory of young adults. However, despite the importance of the decision as to whether and where to enroll in college, prospective students have very poor information on both the costs and the benefits of going to college (Avery and Kane, 2004; Horn et al., 2003; Jensen, 2010). This is particularly the case for high-achieving students in low-income and rural areas, who often do not to apply to college at all, or apply to less selective colleges than students from more affluent backgrounds with similar profiles (Dillon and Smith, 2017; Hoxby and Avery, 2013; Hoxby and Turner, 2015). Providing students with accurate and objective information about colleges with regard to typical graduate outcomes, such as unemployment rates and av-

erage income levels, has therefore become an important objective for US policymakers (Mabel et al., 2020).

However, numerous government-backed efforts and research-led interventions have yielded only a limited impact on enrollment rates, application patterns, or completion rates (Barone et al., 2017; Bergman et al., 2019; Bird et al., 2021; Carrell and Sacerdote, 2017; Cunha et al., 2018; Gurantz et al., 2021; Hyman, 2020; McGuigan et al., 2016). In our study, Bettinger and I use novel data from a large-scale survey of US college graduates to argue that a plausible reason for this might be that students rather seek advice from their parents (Oymak, 2018) and that their parents, when giving such advice, tend to look back on their own time at college and reflect on their subjective experiences and satisfaction. To substantiate this argument we show that alumni satisfaction and willingness to recommend one's alma mater to others are weakly correlated with labor market outcomes. In fact, even those with very poor labor market returns report a high level of satisfaction. The importance of parental advice for student decision-making, combined with the salience of subjective experiences in former college students' evaluations of the benefits of a college education suggests that informational campaigns might have more impact if they address not only prospective college students but also their parents. Further, incorporating satisfaction-based measures in existing college-quality evaluations could also improve the information set provided to students more generally.

The remainder of this chapter will proceed as follows: In Section 2, I will expand on the conceptual framework underpinning the thesis. This I will follow with a brief discussion about the nature of academic transitions and their relevance for my essays in Section 3. Next, I will describe the methodological approach used throughout the thesis in Section 4, with a particular emphasis on causal inference, before Section 5 will conclude the chapter with a summary of the essays and their findings.

2 Conceptual Framework

The topics discussed in this thesis all fall within the human capital tradition of economic research, spawned by the seminal contributions of Becker (1962, 1964), Schultz (1961), Mincer (1958), Ben-Porath (1967) and others. Human capital theory posits that education is an investment in future productivity through the development of skills valuable to the labor market—what Becker (1962, p. 9) referred to as the "imbedding of resources in people." The decision whether to partake in schooling represents an investment problem where a rational agent chooses to do so only if the expected return (in the form of expected future earnings) exceeds the costs of obtaining the schooling.

Within this general framework for human capital, there is a strand of research focusing on the production of skills and other educational outputs. This strand, often referred to as the "economics of education," is characterized by Hanushek and Welch (2006) as having a dichotomous objective: first, to use education-production functions to understand how various inputs map to observable outcomes; and, second, to understand the influence of structural and contextual factors, often resulting from public policies, on educational investments and decision-making as well as on heterogeneity in educational attainment.

To see how my three essays relate to these objectives, consider a simple yet typical production function for human capital, expressed in Equation (1):[3]

$$M_{it} = f(C_{it}, P_{it}, S_{it} \mid \Omega_{it}) \tag{1}$$

Let our output of interest be a skill, and let M_{it} be our measure of that skill—say, a test score—for student i at time t. Assume, for simplicity, that M accurately measures all abilities, cognitive and noncognitive alike, of relevance to the labor market and so is identical with i's human capital. The production of skills might be modeled as a function of inputs (each the focus in one of my essays) from the child (C), the parents (P), and the (pre)schools (S), conditional on the current state of the skill formation

[3]This setup follows List et al. (2021) in notation and style.

process, Ω, which captures the history of these inputs, the skill level in $t-1$, and individual characteristics that do not vary over time. We generally assume that the inputs in Equation (1) are complementary (so that low investment in C will also reduce the productivity of investments in S and P), that M is increasing and concave in the inputs, and that previous skills and investments influence both the skill level of the current period and the productivity of investments made in that period. One implication arising from these assumptions is that investing more in an input will produce more educational output; our ability to do so is constrained by our budget and by the concavity of $f(\cdot)$.

Production functions of this type are ubiquitous in education-economics research, in part because they can be used to analyze a wide range of policy-relevant issues (Machin, 2014). For example, even though there are significant pay-offs to be earned in the labor market from investing in M, many students will fail to maximize Equation (1). Indeed, one of the major puzzles in education economics is why so many students invest so little into their schooling, when the potential benefits are so large (Levitt et al., 2016). In the simple framework outlined above, we can characterize this as a failure to invest in the input C, for example by not putting enough effort into one's schoolwork, thereby reducing the output of schooling. Because underinvestment in C, and subsequent suboptimal production of skills, will affect not only the individual but also the aggregate (i.e., society), there is a role for policymakers to try to stimulate investments (Levitt et al., 2016). However, it is not obvious how policy can influence private investments such as effort. Essay I provides evidence about one channel through which policymakers could stimulate investment in C indirectly, through incentives, using merit-based enrolment to schools.

In contrast, policymakers have more direct influence over S, which might capture — among other things — schooling-related expenditure incurred by the government, such as investments in school finances, facilities, teachers' salaries, or tuition subsidies. Starting with the landmark report authored by Coleman et al. (1966), decades of economic research on education production centered on the relationship between school resources and student achievement (Hanushek, 2020). In recent years, however,

many economists have shifted their focus from the *quantity* of inputs to their *quality*, as illustrated, for example, by the blossoming literature on teacher quality (Hanushek, 2020). In Essay II my co-authors and I study an intervention aimed at improving the process quality of early childhood education through curricular design and pedagogical practice. Hence, our intervention does not represent an *increase* in S, but a change in *type* of S. In other words, if the intervention proves successful, the productivity of S have improved resulting in increased educational output without (or with very small) increases in expenditure by enhancing the quality of instruction. Moreover, under the assumption that the production of new skills is influenced by the stock of skills from previous periods, raising the productivity of S in period t will also make subsequent inputs of S in period $t + 1$ more productive, underscoring the importance of investing in skills early in order to be able to capitalize better on schooling at the next stage (Cunha and Heckman, 2007).

On a broader understanding, List et al. (2021) argue that models for human capital formation, such as Equation (1), can also be used to understand the formation of economic preferences, noting that human capital formation is fundamentally a social activity and that "choices are malleable through investments by children, schools, and parents" (List et al., 2021, p. 17). In Essay III my co-author and I explore how choices regarding educational investments by students might be influenced by parental preferences. For example, let M denote a child's risk aversion. The child's parent might affect M through the input P by transmitting their own risk aversion to the child over the course of his or her childhood. This may in turn cause the child to invest differently in education (changing the input C) than he or she otherwise would have, for example by choosing not to apply for college or by applying only to colleges close to home. Such a channel — from parental inputs, through preference formation, to economic decisions — is one plausible mechanism behind the "hidden supply" of high-achieving low-income students who do not attend selective colleges despite the potentially great economic returns of doing so (Hoxby and Avery, 2013). This is also the channel underpinning our proposed mechanism in the essay, where we argue that parental preferences, which

may not tend to maximize human capital or lifetime earnings, are important for understanding college choices made by students. We note that most policy interventions aimed at increasing rates of college application and enrollment have primarily targeted students and schools – that is, aimed to change the inputs C and S — but have largely left out the students' parents (input P). If, again, we assume that the inputs are complementary to one another, the lack of investment in P might explain why these interventions in C and S have failed to move the outcomes of interest, suggesting that future interventions should target a broader range of inputs.

3 Transitions Between Educational Stages

The common theme overarching the essays in the thesis is that they all examine aspects of educational success at a key transition: from compulsory school to high school (Essay I), from childcare to formal schooling (Essay II), and from high school school to higher education (Essay III). While there are several valid reasons for organizing schooling in distinct stages (e.g., capitalizing on economies of scale to departmentalize and offer more varied schooling options for older students), transitions are disruptive in that they introduce institutional discontinuities (Rice, 2001). Typically involving a cluster of changes, transitions expose students to abrupt changes in both the educational environment and the social context (in terms of the model described in Section 2, this can be seen as an abrupt change or discontinuity in the input S). For example, the transition from preschool to primary school will entail a shift in pedagogical content from a play-based to a more formal curriculum, with schedules and learning goals (Jindal-Snape (Ed.), 2010), particularly in certain countries such as Norway, where the second study was conducted. Children also face new demands on their ability to regulate behaviors, such as paying attention and following instructions (DiPrete and Jennings, 2012). In the transition to high school (or, upper secondary school), the organizing principle of in-

struction will typically change from single-teacher classrooms to subject specialists (Symonds and Galton, 2014). Students will have to manage relationships with many teachers and often with many peer groups, and they must learn how to find their way to many different classrooms on a larger campus (Bharara, 2020; Galton et al., 1999). As they grow older, students will also be expected to assume more responsibility for their own schooling, and the decision to continue their studies will ultimately be placed in their hands. All of these changes——and many others——contribute to turning educational transitions into periods of "psychological disequilibrium," where the crucial prerequisites for further learning include successfully adapting to new policies and rules, to heightened academic standards, and to increasing levels of individual responsibility (Felner et al., 1981).

The challenging nature of transitions, and the negative outcomes often associated with them, are well documented in the educational sciences (Anderson et al., 2000; Bharara, 2020; Eccles et al., 1993; Evans et al., 2018; Galton et al., 1999; Mizelle and Irvin, 2000; Rice et al., 2015; Rice, 2001), particularly when it comes to achievement, mental health, and well-being (van Rens et al., 2018). In response, substantial research efforts have been undertaken to investigate measures intended to mitigate the disruptiveness of transitions in order to minimize the risk of students falling behind or dropping out (Bharara, 2020; Curson et al., 2019). My thesis adds to this literature with regard to two key elements of educational transitions: academic preparedness and choice of institution. Below I will explain how these elements relate to points of transition and how they are conceptualized in economic research, and I will outline some of the main policy questions related to them.

3.1 Academic preparedness

A key predictor of whether the transition to a new education level will be difficult for a student is his or her preparedness. That is, "students must possess the knowledge and skills they need to succeed at the next level" (Anderson et al., 2000, p. 331). In the Norwegian context, this is evident in the fact that higher academic achievement is associated with a reduced

likelihood of dropping out after the transition to high school (Falch et al., 2010), and the Norwegian Ministry of Education highlights insufficient academic abilities as a primary predictor of high-school dropout (NOU 2019:2). Similarly, Scott et al. (1995) estimate that the dropout rate in the bottom quartile for academic ability, as measured using a standardized achievement test, is 26 times that in the top quartile. Anderson et al. (2000) describe a process in which students who are unprepared for the transition fail to adapt to new standards and expectations. This makes them feel marginalized and rejected, and their sense of failure initiates a process of gradual disengagement from school, often leading to conflict and antagonizing behavior. For older students, this process can ultimately lead to dropout: a sense of failing or not being able to keep up with one's schoolwork is one of the reasons most frequently given by students for dropping out of school (Scott et al., 1995).

However, dropout is not the only cost associated with having academically unprepared students. Within the human capital framework discussed in Section 2, the need for academic preparedness reflects the notion that "skills acquired in one period persist into future periods [and are] self-reinforcing" — the *self-productivity principle* argued by Cunha and Heckman (2007, p. 35). In other words, there is a complementarity between the skills accumulated by a student up to the point of transition, and their ability to successfully navigate it. This relationship between academic preparedness and subsequent educational productivity also reflects the notion that skills acquired in one period will bolster investments in new (other) skills in subsequent periods (List et al., 2021). If so is the case, then one logical implication is that students without sufficient skills will not be in a position to capitalize very well on the investments made in them after transitioning to higher stages of schooling, meaning that the productivity of their inputs in producing human capital will be reduced. For this reason, ensuring that students acquire sufficient academic abilities at earlier stages of schooling is an important step toward increasing productivity and enhancing human capital development in later schooling.

I discuss how policymakers can stimulate academic preparedness in Essay I. As academic standards increase at more advanced levels of edu-

cation, students will often experience a greater emphasis on measures of ability as well as higher levels of competition. For example, while primary school often centers on the mastery of core skills, which might be measured with M, there is later a gradual shift toward a greater focus on M *per se* as an observable metrics of achievement, typically in the form of grades and test scores. In some studies, this shift to the more impersonal, evaluative, formal, and comparative environment of secondary school has been linked to a decline in intrinsic motivation and in the commitment to learn (Harter et al., 1992). Middle-school students themselves report instead being more motivated by extrinsic factors, in particular by grades (Anderman and Midgley, 1997; Eccles et al., 1993; Harter, 1981; Midgley et al., 1995; Symonds, 2015).

However, the more rigorous grading practices might not compensate fully for the decline in intrinsic learning motivation among adolescents. Indeed, there is an abundant literature suggesting that motivation and effort correlate with how much is at stake in a given assessment (Napoli and Raymond, 2004; Wise and DeMars, 2005; Wolf and Smith, 1995). This manifests itself, for example, in cross-country ability assessments (such as PISA and TIMSS), where high-income countries often do worse than they would be expected to, considering their overwhelming advantage in educational expenditure. Gneezy et al. (2019) show that this paradoxical result is in fact largely explained by differences across cultures in effort expended when stakes are low: students in Western cultures are likely to put in the effort required to perform well only when a test really "matters." One policy conclusion to be drawn from this is that policymakers should ensure that students face proper incentives that reward effort.

In Essay II I study academic preparedness at the point of entry into formal schooling — a transition that is increasingly emphasized by policymakers and researchers alike. A growing literature demonstrates that effective early childhood programs can have substantial effects on early-life skill development (Berlinski et al., 2008; Cornelissen et al., 2018; Felfe and Lalive, 2018; Felfe et al., 2015; Heckman et al., 2010; Melhuish, 2011). In turn, cognitive and socioemotional skills, such as numeracy, literacy, and executive functioning, have been linked to success at the start of

formal schooling and to longer-run academic achievement and social adjustment (Bennett and Tayler, 2006). Further, skill gaps appearing in early childhood often persist into adulthood, with consequences for educational attainment and labor market participation (Cunha et al., 2006). On the hypothesis that skill beget skills, interventions aimed at closing such gaps should be targeted toward underachieving children and carried out in early childhood, so as to build a foundation of skills on which later learning can take place (Cunha and Heckman, 2007).

Many countries are concerned with easing the transition from childcare to school by mitigating the institutional discontinuities, but the pedagogical approach taken to achieve a smoother transition varies. Whereas countries such as the United States and the United Kingdom promote school readiness by investing systematically in key skills, childcare centers in Norway and other Scandinavian countries typically have a more limited curricular focus (Engel et al., 2015). Scandinavian preschool teachers tend to emphasize the value of free play rather than formal training of key skills, aiming to facilitate learning through spontaneous engagement and interaction between adults and children (Synodi, 2010). In fact, such "unstructured" curricula are becoming increasingly popular in other countries aiming to provide universal childcare. However, one major concern with this approach is that it gives preschool centers considerable freedom with respect to pedagogical content, which may lead to large differences in learning across centers (Engel et al., 2015; Rege et al., 2018). In particular, this heterogeneity in centers' effectiveness in preparing children for the transition to school could contribute to early-life skill gaps across child subgroups. In Essay II, we investigate to what extent systematic investment in key school-readiness skills has differential effects across gender, and we discuss the implications that this might have for the design of curricula for the year closest to the transition from childcare to formal schooling.

3.2 Choice of institution

A second crucial dimension of the transition from one educational stage to the next is "deciding" whether, and if so where, to go to school. The first and last essays of the thesis broadly relate to school choice—in the sense of choosing *where*, rather than *whether*, to enroll in high school and college, respectively.

I use quotation marks to indicate that this decision-making process is usually not solely a matter of preference. First, these choices are restricted in many contexts. For example, many countries including both Norway and the United States use district catchment areas based on residency to decide enrollment into primary schools. Second, school choice typically involves some sort of qualification process. Economic scholars have long argued that the competitive force of the marketplace is a channel through which we could increase the productivity of schools (Hoxby, 2003). In an influential contribution, Friedman (1962) argued that allowing parents and students to choose freely between schools would force the schools with dwindling enrollment to make efforts to improve their educational output or risk being closed down.

In the wake of Friedman's theoretical work, a number of Western countries have adopted variants of school-choice systems. Particularly in the United States, a flurry of research has studied their impact on the students who gain access to selective schools (e.g., Bütikofer et al., 2020; Cullen et al., 2006; Gibbons et al., 2008; Hsieh and Urqiuola, 2006; Lavy, 2010), on schools that face competition (e.g., Epple et al., 2002; Figlio and Hart, 2014; Hoxby, 2003; Lindbom, 2010; Robert, 2010), and on parental decision-making (e.g., Abdulkadiroğlu et al., 2018; Abdulkadiroğlu et al., 2020; Burgess et al., 2015; Hanushek et al., 2007). However, the extant literature has primarily focused on the effects of school choice *after* the right to choose has been exercised. In addition, we know much less about the extent to which school-choice systems affect younger cohorts, particularly in contexts where the choice is tied to merit (Bach and Fischer, 2020). For this reason, previous work will typically not be able to separate effects attributable to changes in student behavior from effects of changes

in peer-group composition or in the incentives facing schools, teachers, and administrators. This weakness of the literature also clouds our view when it comes to learning how students prepare academically for more advanced stages of schooling, and how that preparation might change in response to changing incentives. Essay I aims to bridge that gap in the literature.

The second strand of economic theory that relates to school choice involves inquiring into what makes a school *good*. In economic theory, this will often be operationalized as the productiveness, or value-added, of a school. Within the human capital framework outlined above, school quality plays an integral role in the investment problem facing prospective students. One of the primary predictions of the Becker model is that people choose to invest in more education if the net benefits outweigh the costs. In that regard, school quality can be thought of as an input in the profit function of schooling. More specifically to the choice context, school quality matters for the investment decision of where to enroll — conditional upon the individual having chosen to undertake more schooling in the first place. In a stylized model where agents have perfect information, we would hypothesize that prospective students would enroll in the most effective school that would accept them, conditional on their budget constraint. However, there is abundant evidence that prospective students actually have little, poor, and even wrong information about the costs and benefits associated with pursuing college degrees, and about the relative merits of potential institutions. Trusted adults such as parents play a crucial role as advisors and sources of information for students who are considering making the transition to higher education. Indeed, parents generally provide a critical support function for students at points of transition (Anderson et al., 2000), and their active participation can contribute to smooth transitions between stages of schooling (van Rens et al., 2018). However, parents may also have far from perfect information and may rely mainly on personal, subjective knowledge. In Essay III, we explore what might inform parents' advice to prospective students as well as the policy implications of how parents think about their own university experiences.

4 Methodological Approach

In methodological terms, all three studies included in this thesis can be characterized as representing empirical, or applied, microeconomics. This reflects the fact that my primary unit of analysis is the individual, in most cases a student. Microeconomics studies the behavior and decision-making of individual economic units, as well as their interaction with other agents or institutions. My research is applied in the sense that I make use of microeconomic principles and hypotheses to study real-life contexts and events. It is empirical in the sense that I employ data to investigate relationships between economic parameters of interest. In the following section, I will summarize the methods used across the three essays, reflect on why they are appropriate to answer the questions I ask, and detail some of the strengths and weaknesses of each method. I start with descriptive analysis, which I employ in Essay III, before I briefly review under what conditions and assumptions the associations uncovered in a descriptive analysis might have a causal interpretation, which is the goal of the analysis in Essays I and II.

4.1 Descriptive Analyses

Quantitative descriptive analysis uses data to answer questions of *what*, *who*, *where*, *when*, and *to what extent* (Loeb et al., 2017). Rigorous descriptive analysis also aims to answer questions relevant for policy, research, or both. For example, when discovering a previously unknown phenomenon, description is a vital first step of scientific progress to generate hypotheses and to identify potential causal mechanisms worthy of future investigation, or potential interventions that might solve problems.

Where causal research methods can uncover *whether* interventions work, and *which* ones do, careful descriptive analysis might, for example, provide insights into for *whom* it worked, and *when*: in what contexts and under what conditions. For policymakers considering changes to education policy, evidence based on causal studies devoid of descriptions — that is, lacking information about the characteristics of the population, the fea-

tures of the implementation, the nature of the setting, and so on — will be left with only half the pieces of the jigsaw puzzle. Descriptive analysis is important to understand what types of interventions might be useful or necessary in the first place. In this connection, Loeb et al. (2017, p.1) characterize descriptive analysis as a way to provide an "understanding [of] the landscape of needs and opportunities".

In Essay III, Eric Bettinger and I study a novel data set containing information on college graduates' subjective evaluation of the education they received. We use these data to construct a measure of alumni satisfaction for over 4,000 higher-education institutions. To the best of our knowledge, this is the first effort of this sort in a US context, at least at this scale. In order to provide some insights into how a measure of alumni satisfaction might be relevant for research and policy, we conduct a descriptive analysis to answer *what* satisfaction might be and what it is not, *who* the satisfied alumni are, *where* they attended college, and *to what extent* their level of satisfaction correlates with existing measures of college quality, or with individual outcomes that graduates might care about. By conducting this analysis, we also uncover a plausible hypothesis for why informational interventions targeting prospective college students seem to have limited effects on enrollment patterns. We believe that by doing so, we provide some insights of relevance to future intervention design. In other words, we contribute to the "understanding of the landscape of needs and opportunities" by suggesting a different path through that landscape where opportunities might be more plentiful.

While a descriptive analysis of this sort is thus arguably useful, it also has its limitations. We cannot, for example, answer the question of what *causes* satisfaction. All we can do is describe the patterns we observe in the satisfaction measure. While these patterns may well hint at the causal mechanisms at play, we cannot identify them with any certainty. For example, we find that alumni satisfaction is weakly correlated with labor market outcomes. However, we cannot conclude on the basis of this finding that individuals who report high satisfaction with their education despite poor returns in the labor market are irrational. As we do not manipulate college choices, we are unable to assess what their satisfaction

levels would have been in a counterfactual scenario, and therefore to judge to what extent high satisfaction reflects avoiding even worse outcomes. Such questions of causality are therefore left for future research.

4.2 Causal Inference

While description is an important first step in intervention design, *causal* evidence often has greater policy implications than descriptive evidence. For example, if it is demonstrated that student achievement fell *because* of the introduction of a new school policy, this provides policymakers with more information than if it is simply observed that a drop in student achievement *coincided* with the introduction of that policy. In the first and second essays, the goal of the analysis is to estimate causal effects of a *treatment*. In the second essay, the treatment is a new preschool curriculum, administered by ways of an experiment, where units were randomly assigned to either a treatment group, which implemented the curriculum, or a control group, which did not. In the first essay, the treatment is exposure to high school enrolment reform, with treatment assignment characterized by naturally occurring events in a manner that is often referred to as a "natural" experiment. Common to the empirical strategy in both studies is that the main goal is to estimate effects on relevant outcomes that are directly attributable to the treatment received. Below I will briefly summarize under what conditions and assumptions such estimates have a causal interpretation, and the methodological strategies used to enable such an interpretation.

A typical framework for causal inference in the social sciences rests on the consideration and characterization of the *potential* outcomes for a unit. Using the notation of the Rubin (1974, 1977) framework, let the outcome of interest be some Y. Assume that we have a treatment T and a control C, and that the unit i have an equal probability of being assigned to either. Then consider the unit prior to assignment to treatment. At this point in time, there are two possible states in which we could observe Y after the treatment has been administered: $Y_i(T)$ and $Y_i(C)$. These states are the unit's potential outcomes. The quantity of interest that we

are trying to estimate — the causal *estimand* — then involves comparing the potential outcomes for the unit with different treatment assignments. Intuitively, the causal effect we are interested in can be understood as follows: Given the treatment received by the unit and the corresponding value observed for Y, what value of Y would have been observed if the unit had been given the other treatment? Hence, the individual-level causal estimand is given by $Y_i(T) - Y_i(C)$.

The fundamental problem of causal inference, however, is that we cannot observe values for Y for a given unit i under both treatments (Holland, 1986). As Rubin (2005, p. 323) succinctly states, "[e]ach potential outcome is observable, but we can never observe all of them." In order to quantify the causal effect, we must rely on assumptions about what would have happened to i based on what happened to *other* units exposed to different treatments. A crucial component of causal inference is therefore that we observe multiple units. Assume, then, that we have two units, $i \in \{1, 2\}$. Let unit $i = 1$ be the one randomly assigned to T and $i = 2$ the one assigned to the control. In the simple two-unit case, the best we can do might simply be to calculate the difference $Y_1(T) - Y_2(C)$. Does the difference in Y between T and C have a causal interpretation? That depends on how reasonable it is for us to assume that $Y_2(C)$ is the same value that *would be* observed for unit $i = 1$ if that unit had received C instead of T. We might find this assumption reasonable if the units are fairly similar on observable characteristics prior to the treatment and there is little reason to fear that additional, unobserved "treatments" have affected the units concurrently.

However, in small samples, like the two-unit case, an abundance of differences between $i = 1$ and $i = 2$ will often lead to skepticism as to whether $Y_1(T) - Y_2(C)$ is a "sensible" estimate of the "typical" causal effect of T relative to C — in the terminology of Rubin (1974). To gain confidence in our estimate we have to replicate it and see that a similar treatment yields similar results under similar conditions. Within the context of a single study, this translates into a need for (many) more than two observations. As the sample grows larger, random assignment reduces the likelihood that all units assigned to the treatment condition

will share some characteristic thought to be relevant for Y. For large samples, comparison of the average Y for those exposed to the treatment with the average Y for those assigned to the control group will therefore often yield a reasonable estimate of the typical causal effect, when assignment to treatment is random.

In Essay II, our research design rests on these insights about the power of random assignment. In our field experiment, we tested the efficacy of a new curriculum by randomly deciding which preschools would be given access to it, and which would continue with business as usual. Randomization ensures, in terms of expected values, that there are no confounding treatments of relevance to the outcomes measured that may contaminate the estimates. One classic example of such contamination is a study of labor market outcomes after a job-training program where those individuals who are most likely to benefit from the program are also most likely to sign up for it. In such a case, it is not possible to determine whether any differences in outcomes observed after the program are *causal* effects of the program or simply manifestations of differences in baseline characteristics between those individuals who signed up for the treatment and those who did not. In our case, the causal interpretation we make of our findings is predicated upon the assumption that our randomization procedure shields us from such contamination.

A second assumption underpinning our causal interpretation is stability in treatment assignment, which is often referred to as the Stable Unit Treatment Value Assumption (SUTVA). In order to be able to infer causal effects from the observation of multiple units, we must assume that one unit's treatment status does not affect (the treatment status of) other units, and also that there are no different variants, in terms of features such as treatment intensity or dosage, at any treatment level (Imbens and Rubin, 2015). To illustrate the importance of this assumption, consider our field experiment from Essay II, for which it is relevant in at least two ways. First, SUTVA requires that the intervention received in the treatment group does not affect the potential outcomes of the control group. Since our curriculum was implemented in preschools by practitioners, keeping all participants blind to their treatment status

was infeasible. It is therefore conceivable that teachers in the control group were discouraged by not receiving the resources that went along with the treatment, and that in response they altered their pedagogical practice. It is also possible that a teacher exposed to the treatment would share the content of the intervention with colleagues working at centers assigned to the control group. To mitigate concerns for such violations of the SUTVA assumption, teachers were given strict instructions to refrain from sharing resources with other preschool teachers, and they committed to complying with this request. To minimize discouragement effects, we provided explicit information at the start of the project that all control centers would receive the intervention materials, after our posttreatment assessments were conducted. Hence the only difference between treatment and control centers related to *when* they would be able to make use of the curriculum.

The second way in which the SUTVA is relevant for our experiment regards implementation. The second element of the SUTVA requires that the efficacy of a treatment must not vary within the sample. In a medical trial, this would mean that the drug tested has the same potency for all treated participants. In our case, teachers were given ample discretion to adapt the curriculum to suit their pedagogical approach and to best serve the needs of their child group. This was done to ensure that the teachers would be comfortable with the materials and experience a sense of ownership over them, which should lead to a higher average level of implementation quality but could also lead to heterogeneity in implementation. Because we could not control what happened in preschools directly, we are forced to trust that the teachers did not approach the intervention too differently. In order to ensure high implementation quality and teacher fidelity, all participants were given comprehensive training prior to the start of the project. We also required teachers to fill out weekly questionnaires detailing what they had done and why, and to inform us of any issues, challenges, or changes. Further, members of our team regularly contacted every teacher to discuss their progress during the project period. All of these measures were taken to mitigate concerns about discouragement, spillover, lack of fidelity in implementation, and treatment

heterogeneity. While we cannot guarantee that the SUTVA holds in our study, the causal nature of our estimates rests on the assumption that it does hold.

Random assignment to treatment will have preferable properties in most settings, but for many research questions it is infeasible, either because it would be impractical or prohibitively expensive or because it would be ethically intolerable. In such cases, researchers may instead resort to using observational data based on nonrandom assignment mechanisms but still aim to estimate relationships between parameters that have causal interpretations. Typically, such studies rely on exogenous variation in some explanatory variable of interest, again referred to as the *treatment*, and measure how outcomes differ between units exposed to different types of treatments. Various research designs exploit this exogenous variation to approximate the ideal experimental design.

In the first essay, I employ the difference-in-differences (DID) design, one of the most common quasi-experimental methods for causal inference (Goodman-Bacon, 2021), in conjunction with exogenous variation in enrollment rules stemming from policy reforms. The introduction of these reforms can plausibly be deemed to be exogenous if the reforms are uncorrelated with the outcomes we are interested in measuring. This assumption would be violated if, for example, units (counties, in this case) experiencing a downward trend in student performance were more likely to adopt reforms. In DID designs, the validity of this assumption is assessed by inspecting trends in outcomes between adopting and nonadopting units in the periods prior to adoption. If these trends are found to be parallel, the causal interpretation of the DID estimates rests on the assumption that the trend in outcomes observed for the nonadopting units postreform are similar to the trend that *would* have been observed in the adopting units in the absence of the reforms. In other words, we argue that the nonadopting units reflect a reasonable approximation of the adopting units' potential outcomes.

While a causal interpretation hinges on stronger assumptions for a DID estimate than for results from a randomized controlled trial (RCT), there are several advantages to the DID design. The widespread use of

the DID approach in applied economics is due not only to the simplicity and elegance of the design, but also to "its potential to circumvent many of the endogeneity problems that typically arise when making comparisons between heterogeneous individuals" (Bertrand et al., 2004, p.250). Researchers will find this approach particularly useful in policy-relevant settings where randomization is infeasible but where endogenous variation in outcomes (due, e.g., to selection into treatment or omitted variables) is still a concern (Meyer, 1995). What is more, collecting field-experiment data is both costly and logistically challenging, meaning that it often yields small or convenience-based samples. While this does not necessarily threaten the internal validity of an experiment, it does limit our ability to generalize results to other populations. In contrast, researchers can leverage DID designs to study naturally occurring settings involving large samples of individuals, often at relatively low costs. Particularly in recent decades, comprehensive registries and data records have allowed researchers to analyze samples that ostensibly cover entire populations of interest (Hanushek, 2020; Machin, 2014). Not only can these analyses arguably provide insights that more easily generalize to other contexts, but they can also be better suited for exploring heterogeneous impacts across smaller subgroups, which might be harder to do with precision in an RCT with limited sample size.

5 Summary of Essays

Essay I

Using High-Stakes Grades to Incentivize Learning

Effort by students is critical for the production of human capital, and policymakers are often concerned that students are not motivated enough to capitalize on the learning opportunities they are given. One policy measure to consider in that regard is to provide students with incentives that encourage effort and motivation. It has been shown in the psychology literature that motivation correlates with test stakes, and the experimental economics literature has provided further causal evidence that raising the stakes of tests using financial incentives increases both motivation and effort, with some evidence that it might also increase performance. However, paying students for performing on tests is not a viable policy at scale. How, then, could policymakers use these insights at the policy level?

The first essay in the thesis builds on the above-mentioned literature investigating how raising the stakes on tests affects student performance. One way to boost student performance could be to tie school enrollment to past academic performance. If enrollment in specific schools is something students care about, such a tie should provide them with an incentive to put in the effort required to achieve the grades necessary. In fact, this line of argument partly explains why merit-based school-choice systems have become increasingly common in Norway.

In the essay, I exploit six instances of school-choice reform to investigate how students respond in terms of performance on the exit exam they take at the end of compulsory school (grades 1–10). Even though all students sit for the same test at the same time, the relevant changes in high.school enrollment rules in Norway caused the final exit exam to differ in importance across space and cohorts. My empirical strategy consists in using a staggered triple-difference model to estimate the effects on exam performance of being exposed to such a reform. The third difference leveraged is the supply of schools that a student might find to constitute

reasonable options, based on travel distance. I argue that the incentive given by merit-based enrollment should have little effect on students in rural areas, who for geographical reasons might have only a single high school that they might realistically attend.

I find that middle-school students respond to the incentive given by merit-based enrollment in a manner that economic theory would predict. Tying the compulsory school exit exam to salient outcomes improves the grades attained by 5–6 percent of a standard deviation — an effect size that is moderate, but nonetheless economically meaningful. My findings also indicate that, as expected, the introduction of school choice as such, without a sufficient supply of reasonable choices, has little effect on students. A further interesting finding is that analysis of low-stakes ability assessments suggests that actual learning — and thus not only test-taking behavior — is important for explaining the effect of the reforms. This finding adds causal evidence to an as yet limited literature investigating the extent to which young students' investments in schooling are sensitive to the structural incentives facing them. For policymakers this points to a channel, easily applied at scale, through which student learning can be stimulated.

Essay II

Reducing the Gender Gap in Early Learning: Evidence From a Field Experiment in Norwegian Preschools

with Mari Rege, Ingeborg Solli and Ingunn Størksen

Although an extensive literature documents a persistent gender gap in academic achievement, we do not fully understand its origin. Recent evidence suggests that there are substantial differences across gender in important academic skills even before children start formal schooling. Such gender differences in early learning have implications for the provision of early childhood education and care (ECEC). While existing ECEC programs have been shown to have promising effects in terms of child development and outcomes later in life, the variety of contexts and program features makes the literature far from unified with respect to the conditions and inputs that might support these beneficial effects. Even less is known about

the potential distribution of effects, and about whether the conditions that must be met are similar for all children.

Several studies report results indicating that girls might benefit more than boys from enrolling in ECEC programs in terms of making them ready for school, but so far few hypotheses or possible mechanisms for why this might be the case have been discussed. One potential explanation is that girls and boys seemingly spend their time in childcare very differently, with girls much more likely to engage in activities that promote school readiness and skills development. This suggests that boys may not be exposed to many of the stimulating learning activities that girls seem inclined to engage in of their own accord.

In this study, we use experimental data collected through an RCT in a sample of Norwegian childcare centers to investigate whether providing teachers with a curriculum of structured, yet playful, learning activities yields differential effects across gender. We hypothesized that a more structured curriculum with activities initiated by adults and including all children would be particularly beneficial for boys, who might need more support and scaffolding from teachers to engage in stimulating activities.

In line with that hypothesis, we find that the positive average effects of the intervention on children's school readiness is almost entirely driven by the effect on boys. In contrast, we find little evidence that the curriculum had any effect on girls compared with business as usual. Moreover, we also find suggestive evidence that the boys who were at the bottom of the skill distribution at baseline are the ones who improve the most. With many countries experiencing a push toward universal provision of preschool programs, our results underscore the importance of curriculum design and pedagogical practices as well as the need to consider their effects across child subgroups. Implementing curricula such as that featured in our intervention could potentially reduce gender gaps in early learning by positively impacting the development of boys in particular, thus improving their long-term academic achievement.

Essay III

Alumni Satisfaction, Rankings, and College Recommendations

with Eric Bettinger

There is concern among policymakers that prospective students do not have the information necessary to make informed decisions about their college options. Information asymmetries routinely lead to poor matches between students and colleges. For example, the vast majority of high-achieving students in low-income families fail to apply to selective and prestigious colleges (a population of students Hoxby and Avery (2013) refer to as "the missing one-offs"), even when such colleges provide generous financial aid. The policy response has been to invest heavily in college-information interventions, such as the College Scorecard. By providing high-school students with comprehensive data on college characteristics and graduate outcomes, or college rankings based on various measures of institutional quality, policymakers have hoped to improve matching and to increase both enrollment and completion rates.

However, most of these efforts have proved to have little of the desired effects. Survey evidence indicates that students, rather than relying on these more or less objective data sources, tend to use parents and other close family members as their primary advisors on college options and decisions, by quite a wide margin. Trusted adults thus seem to have a crucial role in the college-application process, but we know very little about how those adults might think about colleges or how they reflect on their own educational background. In other words, when asked to provide advice by prospective college students, what elements of college do such adults base their advice on?

In this study, we report on a novel data set containing detailed information about how former college attendees perceive their own education and about what factors might predict their willingness to recommend others, such as their children, to follow a similar path. We begin by constructing a measure of alumni satisfaction, ostensibly capturing the respondents' subjective evaluation of the value of the schooling they received. We then go on to show that this subjective satisfaction measure is poorly predicted by traditional measures of college quality. Neither in

terms of school characteristics (such as selectivity, structural quality, or ranking) nor in terms of labor market outcomes (employment status and income) do we find much evidence that satisfaction correlates with the return on investment from attending college. Instead, we show that most alumni are very satisfied with their educational path, even those whose labor market outcomes are in fact very poor. In the second part of the analysis, we explore possible predictors of the willingness to recommend one's alma mater to others. Again we find that traditional measures of college quality have little predictive power. Instead, our results indicate that alumni primarily emphasize their own subjective satisfaction, rather than more objective and tangible information. Moreover, results from survey items about self-reported reasons for choosing one's alma mater suggest that, conditional upon actually choosing to go for a college education, people primarily choose what specific college to attend for reasons other than prestige, reputation, or labor market prospects.

Our satisfaction measures can provide policymakers with a fuller picture of why individuals choose to enroll in college and of what inputs they use in deciding between college options. These insights might also be relevant for college administrators looking to improve the recruitment of future students, increase the retention of current students, and boost donations from former students. For researchers, our results provide a plausible reason for why information interventions often fail and also suggest a path for refining such interventions in the future.

References

Abdulkadiroğlu, A., Pathak, P., & Walters, C. (2018). Free to Choose: Can School Choice Reduce Student Achievement? *American Economic Journal: Applied Economics*, *10*(1), 175–206.

Abdulkadiroğlu, A., Pathak, P. A., Schellenberg, J., & Walters, C. R. (2020). Do Parents Value School Effectiveness? *American Economic Review*, *110*(5), 1502–1539.

Anderman, E. M., & Midgley, C. (1997). Changes in Achievement Goal Orientations, Perceived Academic Competence, and Grades Across the Transition to Middle-Level Schools. *Contemporary Educational Psychology*, *22*(3), 269–298.

Anderson, L. W., Jacobs, J., Schramm, S., & Splittgerber, F. (2000). School Transitions: Beginning of the End or a New Beginning? *International Journal of Educational Research*, *33*(4), 325–339.

Avery, C., & Kane, T. J. (2004). Student Perceptions of College Opportunities. The Boston COACH program. In C. M. Hoxby (Ed.), *College Choices: The Economics of Where to Go, When to Go, and How to Pay For It* (Chap. 8). University of Chicago Press.

Bach, M., & Fischer, M. (2020). Understanding the Response to High Stakes Incentives in Primary Education. *IZA Discussion Paper Series*, (13845).

Barone, C., Schizzerotto, A., Abbiati, G., & Argentin, G. (2017). Information Barriers, Social Inequality, and Plans for Higher Education: Evidence From a Field Experiment. *European Sociological Review*, *33*(1), 84–96.

Becker, G. S. (1962). Investment in Human Capital: A Theoretical Analysis. *Journal of Political Economy*, *70*(5, Part 2), 9–49.

Becker, G. S. (1964). *Human Capital*. New York, NY: Columbia University Press.

Ben-Porath, Y. (1967). The Production of Human Capital and the Life Cycle of Earnings. *Journal of Political Economy*, *75*(4, Part 1), 352–365.

Bennett, J., & Tayler, C. (2006). *Starting Strong II: Early Childhood Education and Care*. Paris, France: OECD Publishing.

Bergman, P., Denning, J. T., & Manoli, D. (2019). Is Information Enough? The Effect of Information about Education Tax Benefits on Student Outcomes. *Journal of Policy Analysis and Management*, *38*(3), 706–731.

Berlinski, S., Galiani, S., & Manacorda, M. (2008). Giving Children a Better Start: Preschool Attendance and School-Age Profiles. *Journal of Public Economics*, *92*(5), 1416 –1440.

Bertrand, M., Duflo, E., & Mullainathan, S. (2004). How Much Should We Trust Differences-in-Differences Estimates? *The Quarterly Journal of Economics*, *119*(1), 249–275.

Bharara, G. (2020). Factors Facilitating a Positive Transition to Secondary School: A Systematic Literature Review. *International Journal of School & Educational Psychology*, *8*(sup1), 104–123.

Bird, K. A., Castleman, B. L., Denning, J. T., Goodman, J., Lamberton, C., & Rosinger, K. O. (2021). Nudging at Scale: Experimental Evidence from FAFSA Completion Campaigns. *Journal of Economic Behavior & Organization*, *183*, 105–128.

Burgess, S., Greaves, E., Vignoles, A., & Wilson, D. (2015). What Parents Want: School Preferences and School Choice. *The Economic Journal*, *125*(587), 1262–1289.

Bütikofer, A., Ginja, R., Landaud, F., & Løken, K. V. (2020). School Selectivity, Peers, and Mental Health. *Working Paper*.

Carrell, S., & Sacerdote, B. (2017). Why Do College-Going Interventions Work? *American Economic Journal: Applied Economics*, *9*(3), 124–151.

Coleman, J., Campbell, E., Hobson, C., McPartland, J., Mood, A., Weinfeld, F., & York, R. (1966). *Equality of Educational Opportunity*. Washington, DC: National Center For Educational Statistics.

Cornelissen, T., Dustmann, C., Raute, A., & Schönberg, U. (2018). Who Benefits From Universal Child Care? Estimating Marginal Returns to Early Child Care Attendance. *Journal of Political Economy*, *126*(6), 2356–2409.

Cullen, J. B., Jacob, B. A., & Levitt, S. (2006). The Effect of School Choice on Participants: Evidence From Randomized Lotteries. *Econometrica*, *74*(5), 1191–1230.

Cunha, F., Heckman, J. J., Lochner, L., & Masterov, D. V. (2006). Interpreting the Evidence on Life Cycle Skill Formation. In E. Hanushek & F. Welch (Eds.), *Handbook of the Economics of Education* (Vol. 1, pp. 697–812).

Cunha, F., & Heckman, J. (2007). The Technology of Skill Formation. *American Economic Review*, *97*(2), 31–47.

Cunha, J. M., Miller, T., & Weisburst, E. (2018). Information and College Decisions: Evidence From the Texas GO Center Project. *Educational Evaluation and Policy Analysis*, *40*(1), 151–170.

Curson, S., Wilson-Smith, K., & Holliman, A. (2019). Exploring the Experience of Students Making the Transition From Primary School to Secondary School: An Interpretative Phenomenological Analysis of the Role of Friendship and Family Support. *Psychology Teaching Review*, *25*(1), 30–41.

DiPrete, T. A., & Jennings, J. L. (2012). Social and Behavioral Skills and the Gender Gap in Early Educational Achievement. *Social Science Research*, *41*(1), 1 –15.

Dillon, E. W., & Smith, J. A. (2017). Determinants of the Match Between Student Ability and College Quality. *Journal of Labor Economics*, *35*(1), 45–66.

Duncan, G. J., & Magnuson, K. (2013). Investing in Preschool Programs. *Journal of Economic Perspectives*, *27*(2), 109–132.

Eccles, J. S., Wigfield, A., Midgley, C., Reuman, D., Iver, D. M., & Feldlaufer, H. (1993). Negative Effects of Traditional Middle Schools on Students' Motivation. *The Elementary School Journal*, *93*(5), 553–574.

Engel, A., Barnett, W. S., Anders, Y., & Taguma, M. (2015). *Early Childhood Education and Care Policy Review*. Paris, France: OECD Publishing.

Epple, D., Newlon, E., & Romano, R. E. (2002). Ability Tracking, School Competition, and the Distribution of Educational Benefits. *Journal of Public Economics*, *83*(1), 1–48.

Evans, D., Borriello, G. A., & Field, A. P. (2018). A Review of the Academic and Psychological Impact of the Transition to Secondary Education. *Frontiers in Psychology*, *9*, 1482.

Falch, T., Borge, L.-E., Lujala, P., Nyhus, O. H., & Strøm, B. (2010). Årsaker til og konsekvenser av manglende fullføring av videregående opplæring. *SØF Rapport*, (6200 [03/10]).

Fehr, E., & Falk, A. (2002). Psychological Foundations of Incentives. *European Economic Review*, *46*, 687–724.

Felfe, C., & Lalive, R. (2018). Does Early Child Care Affect Children's Development? *Journal of Public Economics*, *159*, 33–53.

Felfe, C., Nollenberger, N., & Rodríguez-Planas, N. (2015). Can't Buy Mommy's Love? Universal Childcare and Children's Long-Term Cognitive Development. *Journal of Population Economics*, *28*(2), 393–422.

Felner, R. D., Primaver, J., & Cauce, A. M. (1981). The Impact of School Transitions: A Focus for Preventive Efforts. *American Jouranl of Community Psychology*, *9*(4), 449–459.

Figlio, D., & Hart, C. M. D. (2014). Competitive Effects of Means-Tested School Vouchers. *American Economic Journal: Applied Economics*, *6*(1), 133–156.

Friedman, M. (1962). *Capitalism and Freedom*. Chicago, IL: University of Chicago Press.

Galton, M. J., Gray, J., & Ruddock, J. (1999). *The Impact of School Transitions and Transfers on Pupil Progress and Attainment*. London, UK: UK Department for Education and Employment.

Gibbons, S., Machin, S., & Silva, O. (2008). Choice, Competition and Pupil Achievment. *Journal of the European Economic Association*, *6*(4), 912–947.

Gneezy, U., List, J. A., Livingston, J. A., Qin, X., Sadoff, S., & Xu, Y. (2019). Measuring Success in Education: The Role of Effort on the Test Itself. *American Economic Review: Insights*, *1*(3), 291–308.

Goodman-Bacon, A. (2021). Difference-in-Differences With Variation in Treatment Timing. *Journal of Econometrics*, *forthcoming*.

Gurantz, O., Howell, J., Hurwitz, M., Larson, C., Pender, M., & White, B. (2021). A National-Level Informational Experiment to Promote Enrollment in Selective Colleges. *Journal of Policy Analysis and Management*, *40*(2), 453–479.

Hanushek, E. A. (2020). Education Production Functions. In S. Bradley & C. Green (Eds.), *The Economics of Education* (Second Edition, Chap. 13, pp. 161–170).

Hanushek, E. A., Kain, J. F., Rivkin, S. G., & Branch, G. F. (2007). Charter School Quality and Parental Decision Making With School Choice. *Journal of Public Economics*, *91*(5), 823–848.

Hanushek, E. A., Machin, S., & Woessmann, L. (2016). Editors' Introduction. In E. A. Hanushek, S. Machin, & L. Woessmann (Eds.), *Handbook of the Economics of Education* (Vol. 5, pp. xiii–xiv).

Hanushek, E. A., & Welch, F. (2006). Preface. In E. A. Hanushek & F. Welch (Eds.), *Handbook of the Economics of Education* (Vol. 1, pp. xix–xx).

Harter, S. (1981). A New Self-Report Scale of Intrinsic Versus Extrinsic Orientation in the Classroom: Motivational and Informational Components. *Developmental Psychology*, *17*(3), 300–312.

Harter, S., Whitesell, N. R., & Kowalski, P. (1992). Individual Differences in the Effects of Educational Transitions on Young Adolescent's Perceptions of Competence and Motivational Orientation. *American Educational Research Journal*, *29*(4), 777–807.

Havnes, T., & Mogstad, M. (2015). Is Universal Child Care Leveling the Playing Field? *Journal of Public Economics*, *127*, 100 –114. The Nordic Model.

Heckman, J. J. (2006). Skill Formation and the Economics of Investing in Disadvantaged Children. *Science*, *312*(5782), 1900–1902.

Heckman, J. J., Moon, S. H., Pinto, R., Savelyev, P. A., & Yavitz, A. (2010). The Rate of Return to the HighScope Perry Preschool Program. *Journal of Public Economics*, *94*(1), 114 –128.

Holland, P. W. (1986). Statistics and Causal Inference. *Journal of the American Statistical Association*, *81*(396), 945–960.

Horn, L. J., Chen, X., & Chapman, C. (2003). Getting Ready To Pay for College: What Students and Their Parents Know about the Cost of College Tuition and What They Are Doing To Find Out. *National Center for Education Statistics Report No. 2003030*.

Hoxby, C. M. (2003). School Choice and School Productivity: Could School Choice Be a Tide that Lifts All Boats? In C. M. Hoxby (Ed.), *The Economics of School Choice* (Chap. 8, pp. 287–342).

Hoxby, C. M., & Avery, C. (2013). The Missing "One-Offs": The Hidden Supply of High-Achieving, Low Income Students. *Brookings Paper on Economic Activity*, (Spring), 1–66.

Hoxby, C. M., & Turner, S. (2015). What High-Achieving Low-Income Students Know About College. *American Economic Review: Papers & Proceedings*, *105*(5), 514–517.

Hsieh, C.-T., & Urqiuola, M. (2006). The Effects of Generalized School Choice on Achievement and Stratification: Evidence from Chile's Voucher Program. *Journal of Public Economics*, *90*, 1477–1503.

Husain, M., & Millimet, D. L. (2009). The Mythical 'Boy Crisis'? *Economics of Education Review*, *28*(1), 38 –48.

Hyman, J. (2020). Can Light-Touch College-Going Interventions Make a Difference? Evidence from a Statewide Experiment in Michigan. *Journal of Policy Analysis and Management*, *39*(1), 159–190.

Imbens, G. W., & Rubin, D. B. (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences*. New York, NY: Cambridge University Press.

Jensen, R. (2010). The (Perceived) Returns to Education and the Demand for Schooling. *The Quarterly Journal of Economics*, *125*(2), 515–548.

Jindal-Snape (Ed.), D. (2010). *Educational Transitions: Moving Stories From Around the World*. London, UK: Routledge.

Lavy, V. (2010). Effects of Free Choice Among Public Schools. *Review of Economic Studies*, *77*(3), 1164–1191.

Lenes, R., Gonzales, C. R., Størksen, I., & McClelland, M. M. (2020). Children's Self-Regulation in Norway and the United States: The Role of Mother's Education and Child Gender Across Cultural Contexts. *Frontiers in Psychology*, *11*, 2563.

Levitt, S. D., List, J. A., Neckermann, S., & Sadoff, S. (2016). The Behavioralist Goes to School: Leveraging Behavioral Economics to Improve Educational Performance. *American Economic Journal: Economic Policy*, *8*(4), 183–219.

Lindbom, A. (2010). School Choice in Sweden; Effects on Student Performance, School Costs, and Segregation. *Scandinavian Journal of Educational Research*, *54*(6), 615–630.

List, J. A., Petrie, R., & Samek, A. (2021). How Experiments with Children Inform Economics. *NBER Working Paper Series*, (28825).

Loeb, S., Dynarski, S., McFarland, D., Morris, P., Reardon, S., & Reber, S. (2017). *Descriptive Analysis in Education: A Guide For Researchers*. Washington, D.C.: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance.

Mabel, Z., Libassi, C., & Hurwitz, M. (2020). The Value of Using Early-Career Earnings Data in the College Scorecard to Guide College Choices. *Economics of Education Review*, *75*, 101958.

Machin, S. (2014). Developments in Economics of Education Research. *Labour Economics*, *30*, 13–19.

McGuigan, M., McNally, S., & Wyness, G. (2016). Student Awareness of Costs and Benefits of Educational Decisions: Effects of an Information Campaign. *Journal of Human Capital*, *10*(4), 482–519.

Melhuish, E. C. (2011). Preschool Matters. *Science*, *333*(6040), 299–300.

Meyer, B. D. (1995). Natural and Quasi-Experiments in Economics. *Journal of Business & Economic Statistics*, *13*(2), 151–161.

Midgley, C., Anderman, E., & Hicks, L. (1995). Differences Between Elementary and Middle School Teachers and Students: A Goal Theory Approach. *The Journal of Early Adolescence*, *15*(1), 90–113.

Mincer, J. (1958). Investment in Human Capital and Personal Income Distribution. *Journal of Political Economy*, *66*(4), 281–302.

Mizelle, N. B., & Irvin, J. L. (2000). Transition from Middle School into High School. *Middle School Journal*, *31*(5), 57–61.

NOU 2019: 2. (2019). *Fremtidige kompetansebehov II - Utfordringer for kompetansepolitikken*. Oslo, Norway: Ministry of Education.

Napoli, A. R., & Raymond, L. A. (2004). How Reliable Are Our Assessment Data?: A Comparison of the Reliability of Data Produced in Graded and Un-Graded Conditions. *Research in Higher Education*, *45*(8), 921–929.

Oymak, C. (2018). *High School Students' Views on Who Influences Their Thinking About Education and Careers*. Washington, D.C.: U.S. Department of Education, National Center for Education Statistics.

Phillips, D., Lipsey, M., Dodge, K., Haskins, R., Bassok, D., Burchinal, M., ... Weiland, C. (2017). Puzzling It Out: The Current State of Scientific Knowledge on Pre-Kindergarten Effects—A Consensus Statement. *Issues in Pre-Kindergarten Programs and Policy*, 19–30.

Rege, M., Solli, I. F., Størksen, I., & Votruba, M. (2018). Variation in Center Quality in a Universal Publicly Subsidized and Regulated Childcare System. *Labour Economics*, *55*, 230–240.

Rice, F., Frederickson, N., Shelton, K., McManus, C., Riglin, L., & Ng-Knight, T. (2015). *Identifying Factors That Predict Successful and Difficult Transitions to Secondary School*. London, UK: The Nuffield Foundation.

Rice, J. K. (2001). Explaining the Negative Impact of the Transition from Middle to High School on Student Performance in Mathematics and Science. *Educational Administration Quarterly*, *37*(3), 372–400.

Robert, P. (2010). Social Origin, School Choice, and Student Performance. *Educational Research and Evaluation*, *16*(2), 107–129.

Rubin, D. B. (1974). Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies. *Journal of Educational Psychology*, *66*(5), 688–701.

Rubin, D. B. (1977). Assignment to Treatment Group on the Basis of a Covariate. *Journal of Educational Statistics*, *2*(1), 1–26.

Rubin, D. B. (2005). Causal Inference Using Potential Outcomes. *Journal of the American Statistical Association*, *100*(469), 322–331.

Schultz, T. W. (1961). Investment in Human Capital. *American Economic Review*, *51*(1), 1–17.

Scott, L. S., Rock, D. A., Pollack, J. M., & Ingels, S. J. (1995). *Two Years Later: Cognitive Gains and School Transitions of NELS: 88 Eighth Graders*. Washington, D.C.: National Center for Educational Statistics.

Stipek, D. (2012). At What Age Should Children Enter Kindergarten? A Question for Policy Makers and Parents. *Social Policy Report*, *16*(2), 3–16.

Symonds, J. E. (2015). *Understanding School Transition: What Happens to Children and How to Help Them*. England, UK: Routledge Education.

Symonds, J. E., & Galton, M. (2014). Moving to the Next School at Age 10-14 Years: An International Review of Psychological Development at School Transition. *Review of Education*, *2*(1), 1–27.

Synodi, E. (2010). Play in the Kindergarten: The Case of Norway, Sweden, New Zealand and Japan. *International Journal of Early Years Education*, *18*(3), 185–200.

van Rens, M., Haelermans, C., Groot, W., & van den Brink, H. M. (2018). Facilitating a Successful Transition to Secondary School:(How) Does it Work? A Systematic Literature Review. *Adolescent Research Review*, *3*(1), 43–56.

Wise, S. L., & DeMars, C. E. (2005). Low Examinee Effort in Low-Stakes Assessment: Problems and Potential Solutions. *Educational Assessment*, *10*(1), 1–17.

Wolf, L. F., & Smith, J. K. (1995). The Consequence of Consequence: Motivation, Anxiety, and Test Performance. *Applied Measurement in Education*, *8*(3), 227–242.

Chapter 2 – Essay I

Using High-Stakes Grades To Incentivize Learning

Andreas Fidjeland*

Abstract

I investigate whether policymakers can increase human capital production by introducing merit-based enrollment through a natural experiment in Norwegian high schools. By exploiting variation across space and time I compare the performance of students taking the same exit exam in compulsory school, but where the test is high-stakes for only a subset of students. Using a staggered triple-difference framework, I find that exam grades increase in the high-stakes setting if students have a sufficient number of prospective schools within traveling distance. Results from low-stakes ability assessments suggest that actual learning — and not test-taking strategy — could largely explain the observed effect.

JEL Codes: D02, D04, I20, I28

Keywords: Incentives, high-stakes, school choice, learning

*University of Stavanger Business School. E-mail: andreas.fidjeland@uis.no. This paper has benefited from helpful guidance on the part of Mari Rege, Ingeborg Solli, and Eric Bettinger, as well as from excellent comments by Tom Dee, Edwin Leuven, Hans H. Sievertsen, Maximiliaan Thijssen, and the participants in the UiS PhD Workshop in Education Economics. I acknowledge funding from the Norwegian Research Council, Grant No. 270703/H20. All remaining errors are my own.

1 Introduction

Investments in human capital can yield great economic returns both for the individual and for the economy in general. Typical models for the production of human capital posit that it depends both on public inputs such as investments in school resources, facilities, and teachers, and on private inputs such as student effort. However, students often fail to invest properly by making a sustained effort in school, perhaps because the short-term costs are more salient than the rewards, which might materialize only in adulthood (Levitt et al., 2016). Whereas economic research has provided a number of policy prescriptions for the design of the public inputs, it is less clear how policymakers can influence private investments by students. This might be particularly challenging in the case of adolescent students, who tend to be less intrinsically motivated and less engaged in their schoolwork than younger students (Eccles and Midgley, 1989; Eccles et al., 1993), instead increasingly seeking external sources of motivation and validation, typically at ages when they transition to middle and high school (Harter, 1981; Midgley et al., 1995).

Economic theory predicts that we are motivated by incentives. We would therefore expect that grades will provide students with a stronger incentive to learn in cases where they are high-stakes, in the sense that they affect desirable outcomes, than if they are low-stakes (Becker and Rosen, 1992; Grove and Wasserman, 2006; Main and Ost, 2014; Wise and DeMars, 2005). This is illustrated by recent evidence suggesting that low effort could explain why many developed countries produce subpar performances in cross-country ability assessments, despite an overwhelming advantage in educational expenditure (Gneezy et al., 2019; Zamarro et al., 2019). If proper incentives can motivate students to exert a sustained learning effort, their improved effort should also increase human capital production. However, we have very little knowledge on the extent to which adolescent students respond predictably to nonpecuniary incentives in the school setting (Bach and Fischer, 2020).

One way to raise the stakes of grades — and to move the rewards

reaped from investing more effort in school closer to the present — is to adopt merit-based enrollment regimes when allocating students to schools. As exemplified by the debate on school choice, a key argument in favor of such policies rests on the hypothesis that letting students compete for access to schools will incentivize effort if enrolling in specific schools is desirable, thereby promoting academic achievement (Friedman, 1962; Hoxby, 2003). However, we have little direct evidence in support of such a disciplinary effect on students, and particularly on younger students who have yet to exercise choice. This paper is therefore relevant for the many cities and countries that have introduced variants of merit-based high-school enrollment (e.g., Paris, Denmark, Sweden, and the United Kingdom) but lack causal evidence of the effect that such policies have on academic performance in adolescent students.

To investigate the incentivizing effect of high-stakes grades on academic achievement, I exploit a natural experiment created by regional differences in Norwegian high-school admission regimes. Whilst historically the norm has been for students to enroll in their neighborhood high school, several counties have in recent decades chosen to adopt merit-based enrollment regimes, more colloquially referred to as "school choice" policies. In these counties, oversubscription of schools is solved by ranking students according to their compulsory school grade-point average (GPA), admitting those with the highest average first. Given that school placement is thus determined by grades in some counties but not in others, economic theory predicts that students exposed to school choice will attain higher grades, provided that school placement is an outcome they care about. Using rich registry data from a sample period covering six different school-choice reforms across Norway, I exploit the county–year variation in enrollment regimes in a triple-difference framework as my main empirical strategy. To mitigate concerns that county-specific trends or shocks might influence the decision to introduce such reforms, I leverage the supply of schools within traveling distance from a student's home as the third difference. Specifically, I differentiate in terms of whether or not a student, in practice, has a real choice of high schools, defined as having at least three schools within traveling distance. If the prospect of

being able to choose your high school is a driver of student performance, students should not be induced to invest more effort if they have few geographically realistic options to choose from however well they perform. This means that the triple-difference model not only estimates prereform and postreform trends in the reforming counties as compared with nonreforming counties, but also leverages *de facto* nonchoice students as a within-treatment placebo group.

To ensure that changes in grading practices in response to the reforms are not driving my results, I focus my attention on how students perform on the national, centralized exam that all Norwegian students are required to take at the end of compulsory school. On this exam, students are randomly drawn to be tested in one of the three core subjects (mathematics, English, or Norwegian). All students assigned to the same subject take the same exam on the same day. Grading is centralized and double-blinded, and the result enters into students' GPA — which is then used to determine high-school placement in some counties, but not in others. It should be noted that this is the first mandatory national exam faced by Norwegian students and that it represents their last chance to improve their GPA, as the teacher-awarded grades are finalized before the exam (but not revealed to the students until after it). Qualitative studies indicate that Norwegian teenagers experience high-school choice as a critical stage in their schooling, with far-reaching implications for their educational and labor market prospects, and that earning good grades is therefore vital to them (Bakken et al., 2018; Inchley et al., 2013; Ruud, 2018).

According to my results, performance on this final exam suggests that imposing more high-stakes grades has a positive effect on grades earned. I find robust estimates of a treatment effect of 5–6 percent of a standard deviation for those students who are both exposed to a school-choice reform and have a sufficient number of schools within traveling distance — that is, those for whom the exam might actually be experienced as high-stakes. Contrary to common concerns that such merit-based systems might favor certain types of students more than others, I find limited evidence that the reforms had any heterogeneous effects on performance across subgroups.

Rather, subsample analysis suggests that the response to the incentive is fairly uniform, with some suggestive evidence that the effect is stronger for students tested in mathematics.

There are at least two mechanisms that could explain the effect of higher stakes. First, the students' test effort could change. That is, students faced with a high-stakes exam could put in more effort ahead of and during the test itself. This could include adjusting their test-taking strategy (e.g., taking more risks) or making sure to sleep and eat well in the days before the exam. If so, the treatment effect would have limited relevance for human capital development but rather imply that students facing higher stakes will try a little harder on the exam and earn higher grades than others for that reason. The second explanation, which has stronger policy implications, is that students facing high-stakes grades will make a sustained learning effort over time in order to acquire the skills required to succeed on the exam. From a policy perspective, the latter explanation suggests that changes to the incentive structure, in this case stemming from changes to enrollment rules, can be instrumental in increasing students' human capital, with potentially long-lasting effects on subsequent educational and labor market outcomes.

Results from applying the same triple-difference framework to low-stakes national assessment tests conducted in the grade prior to the exit exam indicate that average academic ability increased among exposed students in the wake of the reforms, relative to the control group. This evidence suggests that the learning-effort hypothesis is important for explaining the main effect. This is also corroborated by a dynamic response in the treatment effect, where larger effect sizes are observed for cohorts further removed in time from the reforms. This increasing effect is consistent with the notion that students will adapt to the new regime over time, so that younger cohorts are increasingly aware of the importance of making a sustained effort throughout their schooling and not just toward the end of their final year.

My paper contributes to several strands of literature. First, the results are relevant for the literature examining the links between incentives and academic achievement. A rich accountability literature has documented

how schools, administrators, and teachers might respond to stricter performance standards and outcome-based funding (see Figlio and Loeb, 2011, and Deming and Figlio, 2016, for surveys). However, the present study considers a setting where incentives change for the students only. In contrast to many other studies on related topics (e.g., Gibbons et al., 2008, and Figlio and Hart, 2014), the compulsory schools are unaffected by the reforms to high-school admission and thus have no reason to adjust their behavior or effort. When it comes to student-level effects, a separate but related body of work uses direct financial incentives to increase effort and performance in test-taking situations (e.g., Angrist and Lavy, 2009; Behrman et al., 2015; Bettinger, 2012; Fryer, 2011; Kremer et al., 2009; Leuven et al., 2010). These experimental studies have successfully demonstrated a causal link between extrinsic incentives, motivation, and effort among students, although their effectiveness in moving outcomes has been modest (Levitt et al., 2016). Paying students for their performance is also costly in the long term and may not be feasible on a national scale. Hence, the policy relevance of this body of research remains unclear. My paper is therefore most closely related to Hvidman and Sievertsen (2019) and Bach and Fischer (2020), which consider how students respond to other nonmonetary incentives. The former work considers a grade re-scaling reform in Danish high schools that led to students' GPA being arbitrarily raised or lowered, finding that those students who experienced a fall in their GPA, which determines postsecondary enrollment, responded by performing better in subsequent years, in terms of both teacher-awarded grades and external exams. The authors argue that enhanced study effort is a plausible explanation for this effect. The latter work exploits changes in Germany's tracking system in early primary school. In this case, students face a choice between different ability tracks rather than schools, where some states employ binding recommendations from the teachers based on previous performance. The authors find that relaxing the emphasis on the recommendation in favor of more parental choice reduces student achievement, presumably owing to the reduced incentive to perform well.

On a related note, the paper adds to the literature aimed at under-

standing how competitive behavior implemented through school-choice regimes can influence the efficiency of educational production (e.g., Angrist et al., 2002; Cullen et al., 2006; Figlio and Hart, 2014; Hoxby, 2000; Lavy, 2010). Theoretical studies postulate that allowing parents and students to choose schools freely will improve the quality and productivity on both the supply and the demand side through the disciplinary effect of competition (Becker and Rosen, 1992; Costrell, 1994; Friedman, 1962; Hanushek, 1986). Further, there could also be a positive sorting effect as a result of students (or parents) being allowed to make choices that better fit their needs and preferences, leading to more efficient allocation of students across schools (Epple and Romano, 2003; Hoxby, 2003). However, a weakness of this literature is that outcomes are often measured after the right to choose has been exercised. This makes it difficult to evaluate whether any gains achieved by introducing school choice are indicative of greater learning effort on the part of students or are instead the result of students being in different schools and peer groups. Unlike this literature, I do not study the effect of school choice *per se*, but rather investigate whether the prospect of being *able* to choose, given sufficient academic success, can incentivize students to improve their performance earlier on in their education. Hence my results give a clearer indication of the disciplinary effect of high-stakes grades on student behavior, as opposed to school responses to competitive pressure or the effects of changing peer groups.

Lastly, my paper contributes causal evidence to the interdisciplinary stream of research into the significance of test consequences for performance. The notion that academic tests devoid of consequences will be too low-stakes to make students perform to the best of their abilities is well established in the literature (Wise and DeMars, 2005). Although the results in many cases stem from correlational studies, existing empirical work indicates that motivation and effort are associated with test stakes, while the evidence regarding performance is more mixed (Napoli and Raymond, 2004; Wolf and Smith, 1995). A primary challenge in this literature, as highlighted by a recent vein of research (Gneezy et al., 2019; Segal, 2012; Zamarro et al., 2019), is separating effort and abil-

ity in test-score outcomes. If policymakers are more interested in the students' ability than in their test scores *per se*, the policy relevance of the association may be undermined by the fact that the correlation between test stakes and performance might simply reflect innate differences in factors such as intrinsic motivation and stress resistance (Levitt et al., 2016). To shed more light on the implications of my results, I exploit low-stakes national assessment tests for a supplementary analysis where I argue that test-effort effects are not the primary driver of the results. I thus conclude that there is evidence suggesting that students respond to incentives by exerting effort over time, thereby raising their academic ability. This highlights a channel for policymakers to stimulate private investment in human capital.

The remainder of the paper unfolds as follows: Section 2 describes the institutional and theoretical setting for the analysis; Section 3 details the data, sample, and empirical strategy used in the estimation; Section 4 contains results, while Section 5 investigates potential mechanisms; and Section 6 presents conclusions from the study.

2 Background

2.1 Institutional Setting

The setting for this study is the universal, publicly funded primary and lower-secondary school (henceforth "compulsory school") in Norway, in which attendance is free and mandatory. Norwegian schoolchildren start compulsory school in August in the calendar year of their sixth birthday, and it comprises ten grades and ends in graduation in the year when students turn 16.[1] Private options are limited, with the public-school participation rate exceeding 96% in 2016 (Norwegian Directorate of Education and Training, 2017). The allocation of students to individual compulsory

[1] In the Norwegian educational system, grades 1–7 make up primary school while grades 8–10 make up lower-secondary school, which is roughly equivalent to middle school or junior high school in the United States.

schools is decided on the basis of neighborhood catchment areas. Since having inclusive schools with heterogeneous groups of students is a policy objective, formal parental influence on which school their child attends — except through residential sorting — is limited. In the first seven years, no grades are awarded, as relative performance, ranking, and competition between students are played down in favor of focusing on individual development. Although classroom tests are given, they are typically not scored or ranked in a traditional sense, but primarily serve as a tool for the teacher to chart the progress of individual students. Grades 8 through 10 are seen as a separate stage of compulsory school, and students are typically required to change schools after grade 7; this typically also entails being assigned to a new class.[2] Parental influence on assignment to classes or schools remains limited, and nor is there any tracking at this stage. Indeed, The Education Act (Opplæringslova) (1998) specifies that the classes should reflect the aggregate population, without consideration of ability, gender, or ethnicity, effectively advocating random assignment of students to classes.[3]

Grade 8 also marks the introduction of teacher-assessment grades. In general, grades 8 through 10 represent a more advanced level of study, where subjects are more academically and theoretically oriented and where students are regularly assessed using graded tests and assignments. Every semester, students are given a transcript consisting of a grade on a scale from 1 to 6 for each subject, set by their teachers. However, only those grades received at the end of year 10 will enter their official school record. The final teacher-assessment grades (in all subjects) along with the grades from the above-mentioned final exit exam make up a student's compulsory-school GPA, with all grades given equal weight. Hence the exit-exam grade is one out of approximately 13 grades on the transcript, meaning that the direct impact of the exam on school placement may

[2]In this context, "class" refers to a set group of students within a cohort who you a classroom and attend most subjects together. A class typically stays together for all three years of middle school.

[3]Auestad (2018) shows that within a school, Norwegian students are in fact as-good-as randomly distributed to classes in grade 8. She also finds that Norwegians rarely move house in order to enrol their children in specific schools.

be limited for the student population as a whole. Even so, a two-step increase in the grade earned on the exam will by itself move a student roughly five percentiles up in the GPA distribution, which is more than enough to have a real impact for students who are at the margin of being admitted to their first-choice school rather than their second-choice one. Moreover, what is crucial for whether the incentive represented by the exit exam has a performance-enhancing effect is not so much its objective impact on outcomes as how it is perceived by students. Both Norwegian and cross-country surveys indicate that Norwegian students experience above-average levels of school-related stress toward the end of compulsory school (Bakken et al., 2018; Inchley et al., 2013). Some studies report that students in grade 10 link stress to internal and environmental pressure to perform well, so that they do not spoil their chances of obtaining a good education and having successful careers (Bakken et al., 2018). On an anecdotal note, some students claim that not getting accepted to their preferred school would mean that "everything is ruined" (Ruud, 2018). The final exam represents the last opportunity to better their chances of admission to their preferred school, and it is therefore likely that many students will experience it as high-stakes.

After graduating from compulsory school, students can apply to enroll in high school. While this is not mandatory, students have a statutory right to acceptance for upper-secondary education, and very few end their education before or immediately after finishing compulsory school.[4] When applying to high school, students make their first choice of education track, choosing between a variety of vocational and academic programs.[5] Within programs, the allocation of students between high schools varies from county to county, which is what provided the variation exploited in this study. Administration of the high-school sector is a key task at the county-government level, with decisions on the administration of admissions left to county-level politicians' discretion.

[4]For example, at the start of the 2015/16 academic year, 92% of 16–18-year-olds were enrolled in high school, while only 192 individuals failed to complete their compulsory education (Norwegian Directory of Education and Training, 2017).

[5]The vocational track leads to an apprenticeship within a trade. The primary function of academic-track programs is to prepare students for higher education.

2.2 High School Enrollment Reform

In simplified terms, current processes for high-school admission use one of two opposing regimes. One of them, the neighborhood-catchment (NC) regime, follows the principles of compulsory school in requiring students to attend their nearest school, that is, the high school closest to their place of residence that offers their preferred educational program. Proponents of the NC regime emphasize that this allows students to stay close to home, limiting lengthy commutes and keeping youths attached to their local communities. It also serves to promote heterogeneity within the student body, as it constrains students' ability to self-select into specific schools (on parameters other than program preferences). The opposing type of regime is the school-choice (SC) regime, which allows students to apply to any school within their county, regardless of its geographical location. This includes the option of applying for the same program in several schools, or for several different programs in the same school. In densely populated areas, there will typically be several schools offering the same programs, particularly the academic ones. Where the number of applicants to a high school exceeds its capacity, students are ranked by compulsory-school GPA, with the highest scores being prioritized. Admission is purely merit-based: grades are the sole determinant of student allocation.[6] The cutoff for admission to a particular school is thus equal to the GPA of the last student admitted in that particular year (in the case of ties, admission officials will perform a random draw between those at the cutoff). Cutoffs vary substantially with the popularity and perceived quality of schools and also fluctuate from year to year in accordance with application patterns.[7] Hence the SC regime places a significant emphasis on the grades attained by students in compulsory school, meaning that the final exit exam involves higher stakes for students in SC counties than

[6]For a few programs, such as music and sports, there are additional tests for ability in the domain area. Moreover, in certain instances some counties also take into account a student's travel distance, but this is done on a discretionary case-by-case basis.

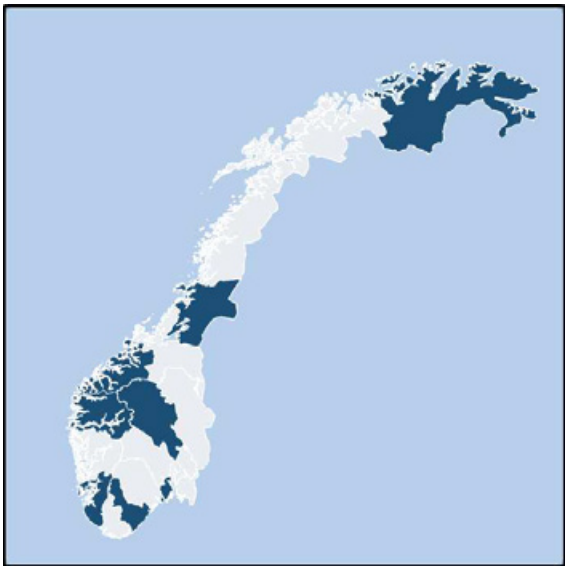
[7]For the least popular schools, admission will typically be uncontested, while it is not uncommon for the cutoff in the most popular urban schools to exceed a 5.0 GPA (out of a possible 6). Information about previous years' cutoffs in specific schools is made available to students.

for students in NC counties.

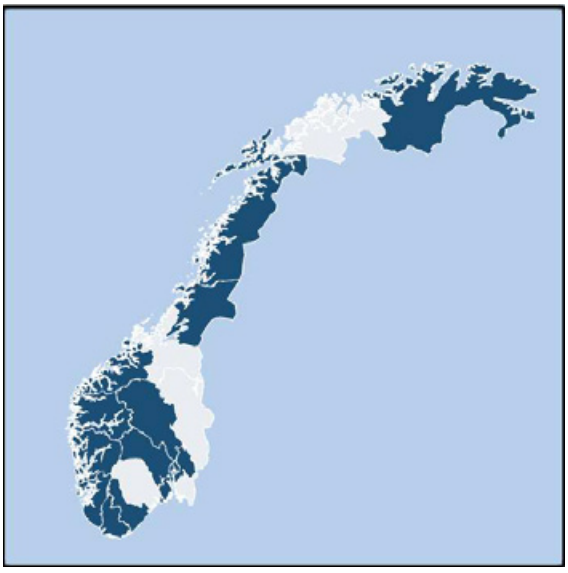
Historically, prior to my sample period, the regimes have changed frequently, with an increasing number of counties adopting SC policies throughout Norway. While eight of nineteen counties were already using an SC regime at the start of my sample period, the variation exploited in this study is provided by the six counties that implemented reforms to introduce SC in their high-school application process during the period from 2002 to 2015.[8] The geographical distribution of admission regimes in the first and last years of my sample period is illustrated in Figure 1. In the final year, 2015, only five counties still applied an NC regime. The SC reform decisions followed a timeline similar to that presented in Figure 2. Thus, students in their final year at the time of the relevant vote had only their last semester to adjust to the new regime.

A county survey of the student population conducted in the wake of one such reform indicated that SC disrupted existing enrollment patterns (Arbeidslaget Analyse, Utgreiing og Dokumentasjon, 2005). In the county of Hordaland, one-quarter of the first cohort affected responded that their preferred high school was not the one they would have been assigned in an NC regime, and 75 percent of those had succeeded in enrolling in their first-choice school. Of the remaining students, who would have preferred to enroll in their geographically closest school, 85 percent were accepted by their first-choice school. In both cases, acceptance rates indicate that enrollment was competitive. However, there is substantial heterogeneity across geography and ability, with the most popular schools being located in city centers. Teacher responses suggest that the primary realignment effect brought about by merit-based enrollment consists in allowing high-ability students in suburban and rural areas to enroll in popular urban schools, displacing low-ability students from the city centers who have to settle for less competitive schools further away.

[8]Specifically, Akershus (2003), Hordaland (2005), Oslo (2009), Vest-Agder (2012), Buskerud (2012) and Nordland (2014).



(a) 2002



(b) 2015

Figure 1: Spread of School Choice Regimes in Norway

Note: Illustration of the increase in school-choice regimes in Norwegian counties during the 2002–2015 period. Dark shading of counties indicates some kind of school choice being in effect for students graduating from compulsory school in that particular year.

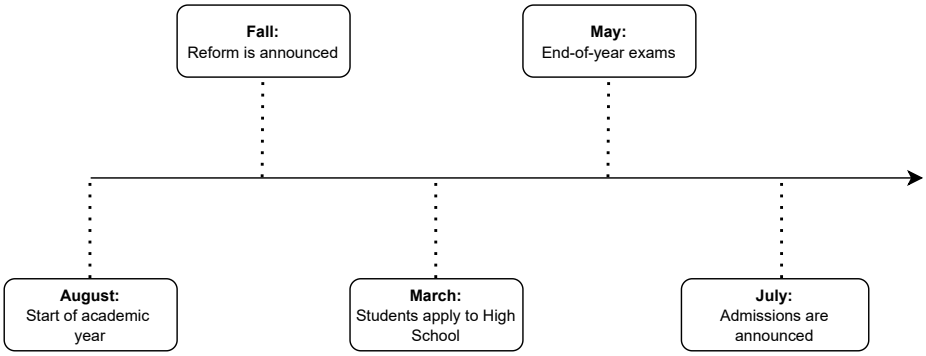


Figure 2: Timeline for the School Choice Reforms

Note: Overview of the series of events of the school choice reforms in the sample.

2.3 Conceptual Framework

This paper investigates the relationship between high-stakes grades and performance. The plausibility of this causal mechanism rests on the hypothesis that linking performance to desirable outcomes creates an incentive that motivates students to exert more effort in school, and that that effort subsequently has a causal effect on learning and human capital accumulation. At the individual level, we would expect an increase in such effort if students perceive, first, that such an increase is clearly related to performance in the relevant domain and, second, that the possible outcomes are of sufficient value to them.[9] In economic terms, we would expect students to invest in school through effort if they expected long-term rewards exceeding the short-term costs of that effort (Levitt et al., 2016). Receiving a grade is not in itself enough to elicit such a response if the associated consequences are not of sufficient magnitude (Grant and Green, 2013). The Norwegian school-choice admission regimes plausibly improve the situation by implicitly providing an extrinsic incentive through a merit-based enrollment regime.

High-stakes grades can be expected to be a more effective incentive

[9]In psychology this is referred to as the expectancy-value theory of motivation (Wigfield and Eccles, 2000).

for some students than for others. Some studies have suggested that motivation to learn, in the sense of striving to acquire the skills demanded by school, is an innate individual characteristic or trait (Brophy, 1987; Segal, 2012). Students who have a strong motivation to learn (whether innate or not) would be expected to work hard and try their best, even in the absence of any extrinsic incentives that policymakers might offer, simply because of their intrinsic drive. Segal (2012) finds that students displaying these traits also perform well on low-stakes assessments, suggesting that they are already properly motivated to capitalize on learning opportunities even when there is no tangible benefit to be gained. Hence high-stakes grades can be expected to provide a more effective incentive for students who do not exhibit those characteristics. Assuming that such students invest strategically in school, effort levels will also vary across individuals, as a function of students' relative probability of achieving their desired outcome (Vroom, 1964). Further, it is often assumed that effort and ability are complementary, and that the marginal effect of effort on human capital production increases with ability (Oettinger, 2002). If this is so, high-stakes grades will primarily improve the performance of low-effort, high-ability students. All else being equal, this would lead to such students being placed in better schools under a school-choice regime, which could lead to adverse segregation effects.

By contrast, effort might be negatively correlated with ability if high-achieving students are able to attain the maximum grade with less effort than average students (Stinebrickner and Stinebrickner, 2008). If so, we might expect a ceiling effect to cause a negative bias in estimates at the top of the ability distribution. Thus there might be little reason to expect a difference in behavior between two such students exposed to opposite regimes. As we can reasonably assume that motivation at least partly maps to performance through effort, it is also reasonable to assume that many high-achievers will already be sufficiently motivated. We might therefore expect a stronger effect among low-achievers for whom faltering motivation could be a root cause of their underperformance. Additionally, some studies have demonstrated that boys respond more than girls to the extrinsic incentives of a competitive environment (Azmat et al., 2016;

Hopland and Nyhus, 2016). Provided that boys outnumber girls in the low-achieving segment of the student population, a stronger treatment effect on boys would indicate a stronger effect for low-ability students.

Tying test performance to desirable outcomes might also change the way students approach the test itself. Since exams map a continuous ability distribution to an arbitrary, discrete scale, the expected marginal benefit of performing better is conditional on a given student's latent ability level prior to the exam in relation to those discrete grades: if a student is not near the margin between grades, the short-term expected marginal benefit of effort is close to zero, while the marginal costs are positive. Thus we would primarily expect to see a performance-enhancing effect on students whose latent ability level is sufficiently close to a point where they could earn a higher (or fall to a lower) grade, and who therefore have positive expected marginal benefits from investing effort. In line with this theoretical argument, some experimental studies have noted that the effect of introducing extrinsic incentives is greatest for a "marginal group" of students, that is, for those who have success within their reach and are neither at the top nor at the bottom of the ability distribution (Angrist and Lavy, 2009). Therefore we might not expect to see any substantial effect on the treatment group as a whole. However, for students who perceive themselves to be at the margin between grades, such an incentive might represent a sufficient nudge to make them put in more effort.

3 Data and Analysis

3.1 Data

The study relies on comprehensive registry data retrieved from the Norwegian National Database of Education, maintained by Statistics Norway. The registry of interest contains compulsory-school outcomes of every student enrolled in a Norwegian school who graduated from grade 10, and it covers the entire student population in the sample period. The sample is limited to 14 adjacent cohorts during the period from 2002 to 2015,

which include a total of 856,040 individuals. Central to the analysis are records detailing the final grades attained by each student in all subjects, both through teacher assessments and through written and oral exams. Additionally, the registry contains information about the subject in which a student was tested on the final exit exam as well as about when and where students graduated. Individual identifiers allow me to link school outcomes to other registries that provide rich details about demographic characteristics, socioeconomic status (SES), and family origin. These identifiers also allow students to be matched with their parents, producing a rich set of potential covariates that can be controlled for in the estimates.

The focus of the analysis is on students graduating from compulsory school, in specific counties exposed to either a school-choice or a neighborhood-catchment regime. As each student is only observed once (at the time of graduation), the data are organized as a repeated cross-section, with dummies indicating from which county and in what year a particular student graduated. Graduation takes place in the spring, and most students subsequently enroll in high school the following August. Cohorts are therefore referred to using the year in which they left compulsory school.[10] Similarly, the reforms are deemed to be in effect starting with the first cohort whose members are able to exercise expanded choice in their high-school applications.[11] Details of current high school admission systems are available in each county's regulations (see www.lovdata.no). Some of these also contain notes about significant changes made to the admission regulations, but typically they do not include detailed information about the timing of reforms. To determine when reforms were implemented, I rely on two investigations carried out at the request of members of Parliament that provide additional details on which counties adhered to which systems at the times in question (Dokument 8:41, 2006; Dokument 8:8, 2003). However, as the most recent of those investigations

[10]For example, the cohort enrolled in Grade 10 in the 2002/2003 academic year is referred to as the 2003 cohort.

[11]If students graduating from compulsory school in the spring of 2003, in a reforming county, can exercise school choice the following fall, the reform is defined as being implemented in 2003.

was carried out in 2006, I have supplemented information from public records of county-parliament sessions for later cohorts. In addition, I have cross-checked those records with newspaper articles from local media in the relevant counties to determine the exact timing of the reforms.

3.2 Measures and Variables

The key outcome variable is a student's grade on the final exit exam in grade 10, standardized to a mean of 0 and a standard deviation of 1 for ease of interpretation. As noted above, in the final semester of compulsory school, all students are randomly drawn for testing in a centrally administrated written exam in either mathematics, English, or Norwegian.[12] The draw is randomized at the class level, so students in different classes within a school will typically be tested in different subjects. It is the responsibility of the individual municipalities in each county to implement the draw in a manner that ensures an even distribution of students across exam subjects, and of exam subjects across schools (Norwegian Directorate for Education and Training, 2018). All students selected for testing in the same subject will take the exact same exam on the same day, and their exam papers will be graded externally by compulsory-school teachers in another part of the country. Both students and teachers remain anonymous throughout the grading process, which uses the same integer scale from 1 (fail) to 6 (top) as teacher-assessment grades and is based on an absolute standard criterion. This anonymity throughout the process and the use of external graders makes the exam grade a more reliable outcome measure than the full GPA, because it is plausible that teachers' grade-setting practices are endogenous to the high-school admission regime applied in a county in the sense that teachers in school-choice counties could be more lenient in an attempt to help their students gain admission to their preferred school. This is clearly less of a concern when the teacher grading an exam does not know who the student is or where

[12] Additionally, students are tested in an oral exam with a similar randomized draw. However, in this case all subjects are eligible for testing, and the exam is carried out locally at each school. The grade from this exam is also added to the student's GPA.

he or she lives.[13]

In order to gauge whether students have a real choice of schools, I construct a measure of the number of high schools within traveling distance from the student's home. To determine whether a school belongs to a particular student's choice set, I use the commuting zones in which students reside.[14] These represent geographically demarcated areas inside county borders within which traveling distances are such that an employee could be expected to commute to work on a daily basis. The variable for the number of schools available to a student thus indicates the number of schools located in his or her commuting zone of residence in the year when he or she graduated from compulsory school. Since there are two main educational tracks to choose from in high school (academic and vocational), I define "real choice" as having at least three high schools within your commuting zone. By doing so, I ensure that at least one of the main tracks will be available in at least two different schools in that region.[15] A total of 599,885 observations (76 percent) satisfy this condition. However, as most Norwegian high-school students will not be able to obtain a driver's license until their final year (the age limit is 18), the commuting zones probably approximate to the *maximum* traveling distance that a student would consider for a daily commute. Because of their reliance on public transport and other means of transportation, this definition will likely overstate the true choice set that a student would consider, which will bias effect sizes toward zero.

In addition to the commuting zone and the cohort-specific fixed effects necessary to estimate DID and triple-difference models, I control for a rich set of conventional covariates. The Central Population Registry provides details on students' gender, nationality, and year of birth. Records of immigration status are used to construct an indicator of im-

[13]For the curious reader, I include results from using a GPA constructed from all nonexam grades as the dependent variable in Table B.3 in the appendix. The effect sizes in this analysis are largely similar to those estimated in the main analysis.

[14]Definitions and demarcations of these zones are given in an overview provided by Statistics Norway — which refers to them as "economic areas."

[15]I assess the sensitivity of the results to this definition in the appendix. Please refer to Section 4.2 for more details

migrant background, defined as being either a first-generation immigrant or born in Norway but having at least one parent born outside of Norway. Using unique identifiers, I link students with their parents, in order to collect data on parental education and income. Education (the highest level of education completed by each parent) is measured on Statistics Norway's nine-point scale.[16] For income, I use the registered taxable income in Norwegian kroner from official tax records for both parents in the year that the student graduated, with household income being the sum of these incomes rounded to the nearest 1,000. Then I divide, for each year separately, households into deciles according to income rank; this is the variable that I include in my analyses.

3.3 Sample Selection

The estimation sample is constructed from the universe of 858,306 individuals having graduated from compulsory school during the years 2002–2015. Of these students, 3,057 were exempted from taking the exit exam (e.g., owing to special education needs) and 835 were confirmed sick on the day of testing. A further 1,863 students did not show up for the exam without providing a reason for their absence. In accordance with Norwegian guidelines, these were not given a failing grade but rather marked as "Not graded." In the present sample, these cases are coded as missing values. An additional 61,605 observations are missing, mostly because of a large teachers' strike in 2008 that caused exams to be canceled. However, attrition analysis — available in Table C.1 in the appendix — shows that grade missingness is not predicted by treatment status. In total, 68,370 observations without exam grades are excluded from the analysis, leaving an estimation sample of 790,936 unique student-level observations. In cases in which a student is registered with multiple graduation years and outcomes (true for 2556 students, 0.29 percent of the gross sample), I use the earliest observed result. In cases where information is missing for covariates, dummies for missing values are constructed and included accordingly, and the covariates are set to zero.

[16] See Statistics Norway (2001) for details.

3.4 Summary Statistics

Table 1 details summary statistics for the estimation sample. Column 1 lists mean values and standard deviations for key variables computed for the treated counties (those that implemented reforms to high-school enrollment during the sample period). Column 2 lists corresponding values for the 13 nonreforming control counties.

Since Norway has a fairly homogeneous population, there are few disparities in the demographic composition of the two groups. One noteworthy exception is the share of immigrants, which is markedly higher in the treated counties. Those counties also have higher levels of average household income than the control counties, despite there being no discernible difference in education level. This is probably due to the fact that some of Norway's largest urban areas, which have a higher frequency of income outliers, are among the reforming counties. This fact is also reflected in the average number of schools available to students as well as in the size of the county cohorts. The average student in the treated counties has thirteen high schools within his or her commuting zone and belongs to an average graduating cohort of 100 students. By contrast, students in control counties have an average of six high schools to choose from and the average graduating cohort per school there consists of 89 students.[17]

Students are — by design — evenly distributed between exam subjects. The only discrepancy found with regard to the exam-subject draw is that the sample share of students tested in Norwegian is roughly 10 percent smaller than that for the other subjects. This is due to the aforementioned teachers' strike in 2008, which overlapped with the predetermined date for the exam in Norwegian, meaning that it was mainly that exam that was canceled. By contrast, exam performance varies substantially. Figure 3 shows that the likelihood of earning the bottom two grades is markedly higher for those selected to be tested in mathematics, all else

[17]While differences in observable characteristics do not bias the results in a DID design *per se* (unless underlying trends overlap with the timing of the reforms, which is particularly unlikely in a triple-difference setting), I do control for a rich set of conventional predictors of school achievement, such as parental background and socioeconomic status, in all my estimations in order to increase the precision of the models.

being equal. In fact, mathematics exams account for three-quarters of all failing students while over half of the students who obtained the top grade were tested in English. One potential concern is that changes in the composition of draws across treatment status and time could threaten the identification strategy. However, considering that the subject draw is randomized within schools across classes, it is unlikely that this would be the case. [18]

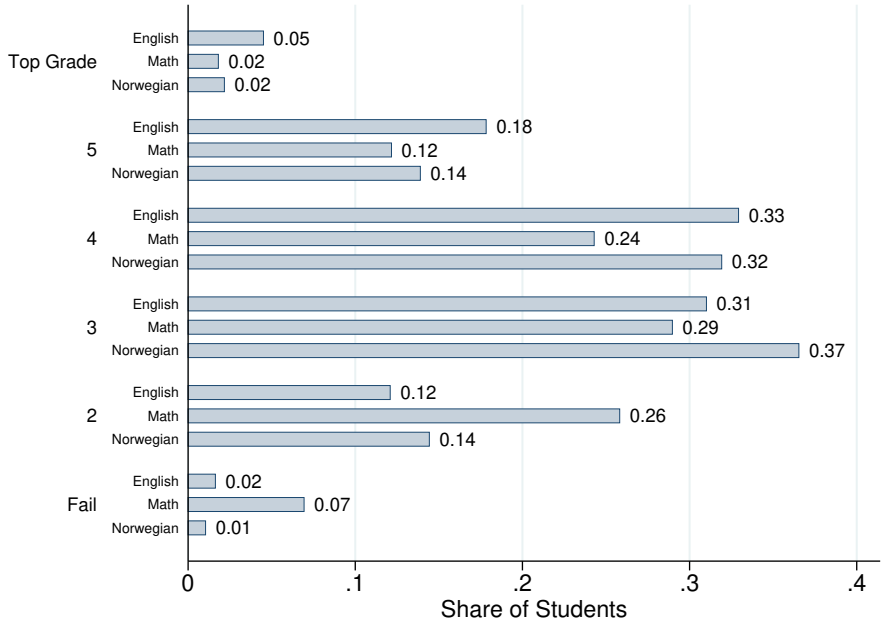


Figure 3: Distribution of Exam Grades by Subject

Note: Grade distribution for each exam subject, measured as the fraction of students tested in that subject attaining a specific grade. The exams are graded on a six point integer scale, where 6 is the top grade, and 1 is a fail.

[18] Based on results not included here, I find that neither treatment status nor covariates are predictive of being tested in mathematics rather than a language.

Table 1—Summary Statistics

	Treated		Control	
	Mean	SD	Mean	SD
Background characteristics				
Female	0.488	(0.50)	0.488	(0.50)
Year of birth	1992.6	(4.11)	1992.6	(4.12)
Age at graduation	16.09	(0.95)	16.09	(1.04)
Immigrant	0.125	(0.33)	0.070	(0.26)
Mother's education	13.32	(2.97)	13.14	(2.673)
Father's education	13.47	(2.87)	13.12	(2.59)
Household income	893.2	(1690.8)	790.8	(785.7)
Educational setting				
Number of HS in CZ	13.10	(9.44)	5.74	(4.67)
Share with >2 HS in region	0.83	(0.38)	0.70	(0.46)
Number of students in school	100.8	(53.26)	88.75	(51.79)
Written exam subject				
Math	0.38	(0.48)	0.38	(0.48)
English	0.36	(0.48)	0.36	(0.48)
Norwegian	0.26	(0.44)	0.26	(0.44)
N	350,858		440,078	

Note: Summary statistics for all students in treated counties compared with the control group. Standard deviations in parentheses. The treatment group consists of the six counties which implemented high-school enrollment reforms during the 2002–2015 period. All nonreforming counties constitutes the control group. Immigrant is defined as having at least one parent who was born outside of Norway. For the education measure, I convert Statistics Norway's nine-point scale for an individual's highest completed degree to years of education using their own conventions. For reference, completing high school is equal to 13 years of education. Household income is reported in nominal NOK/1000. "HS" = high school, "CZ" = commuting zone.

3.5 Empirical Strategy

The Triple Difference Model

The empirical model of interest in this study is the linear relationship between student performance and high-stakes grades (as proxied by the high-school admission regime), as expressed in Equation (1).

$$y_i = \mu D_i + \varepsilon_i \tag{1}$$

If students were randomized to admission regimes, the binary variable D_i in (1) would identify an unbiased causal effect on some outcome y_i of exposure to high-stakes grades. However, it is plausible to claim that students are exposed to either regime in a nonrandom fashion. This gives rise to concerns that (1) would falsely attribute mean differences between the student groups to the regime to which they are exposed.

One way to overcome this identification issue is to exploit the fact that counties implemented school-choice reforms at different points in time, in a difference-in-differences setup (DID). In a potential-outcomes framework, we can consider such a policy-induced change to the admission regime as a treatment, with treatment status assigned by the binary variable D, so that y_{1i} is the outcome of student i exposed to such a school-choice reform, while y_{0i} is the potential outcome for that student in the absence of that reform.

$$\begin{aligned} &\mathrm{D}_i = \{0, 1\} \\ \rightarrow\ &y_{0i} = \text{Outcome for student } i \mid D_i = 0 \\ \rightarrow\ &y_{1i} = \text{Outcome for student } i \mid D_i = 1 \end{aligned}$$

Since the potential outcomes for either condition are unobservable, DID proxies the counterfactual outcomes for those treated by taking the difference between pretreatment and posttreatment observations for a control group, under the assumption that the treatment group would have followed a similar trend if they had not been treated.

$$\mathrm{DID} = E(y_{1i,\,\mathrm{post}} - y_{1i,\,\mathrm{pre}} \mid D_i = 1) - E(y_{0i,\,\mathrm{post}} - y_{0i\,\mathrm{pre}} \mid D_i = 0) \tag{2}$$

The analytical analogue to (2) would then be to estimate[19]

$$y_{ict} = \alpha_c + \lambda_t + \mu D_{c,t} + \varepsilon_{ict} \tag{3}$$

[19] In this brief exposition I exclude nonessential covariates such as student characteristics for the sake of simplicity.

where y_{ict} is the outcome of interest for student i in county c from cohort t, and α_c and λ_t are vectors of indicators controlling for unit- and time-specific fixed effects. The variable of interest,$D_{c,t}$, is a binary indicator that takes the value 1 if county c has been treated by time t and the value 0 otherwise. ε_{ict} is an error term. Within this framework, $\hat{\mu}$ measures the effect of being exposed to a school-choice reform, which is estimated by taking the difference between pretreatment and posttreatment periods for both the treatment group and the control group, and then the difference between these two differences as laid out in (2).

The identifying assumption of the DID model is that of parallel trends; this model posits that, in the absence of an intervention, the trends in outcomes would be equal for treatment and control units, so that any observed deviation from this trend is attributable to the policy change of interest. Thus, in the absence of treatment, the DID framework assumes that

$$E(y_{ict} \mid c, t) = \alpha_c + \lambda_t \tag{4}$$

implying that any observed difference in posttreatment periods is the sum of unit-specific mean differences (α_c), and year-specific effects present among all observations (λ_t). This implies that the potential outcome of the treated cohorts should be unrelated to the timing of the policy change. However, in a setting where the reform is a political decision, this assumption might not hold entirely. For example, there might be unobserved underlying trends in outcomes in the treated units that induced these particular counties to consider school-choice reform in the first place. Further, these reforms could be the result of changes in the political landscape that also led to other changes at the county level around the same time (say, an increase in investment in the educational sector), and those other changes might be correlated with student outcomes.

To assess the viability of the parallel-trends assumption, Figure 4 charts the trends in exam grades, measured as raw averages, for the treatment and control groups. For this plot, I average the exam grade of students in each treated unit in a window around the treatment occurrence, and then average these across units. I then construct a similar

time series for the nontreated students in the same windows, and average over each relative time period. The resulting plot is a trend line centered around the treatment occurrence for all treated units. Under the identifying assumption of the DID model, the trends in exam grades should be parallel in periods prior to the reforms. Figure 4 suggests that this assumption holds only modestly well. While the differences are not large in absolute terms, the trends in the treatment and control groups appear to deviate to a certain extent from one another. At the very least, the plots in Figure 4 do not conclusively allow rejection of the possibility that the treatment group is on a different pretreatment trend than the control group. This raises concerns about the causal nature of DID estimates of the effect of the policy reforms.

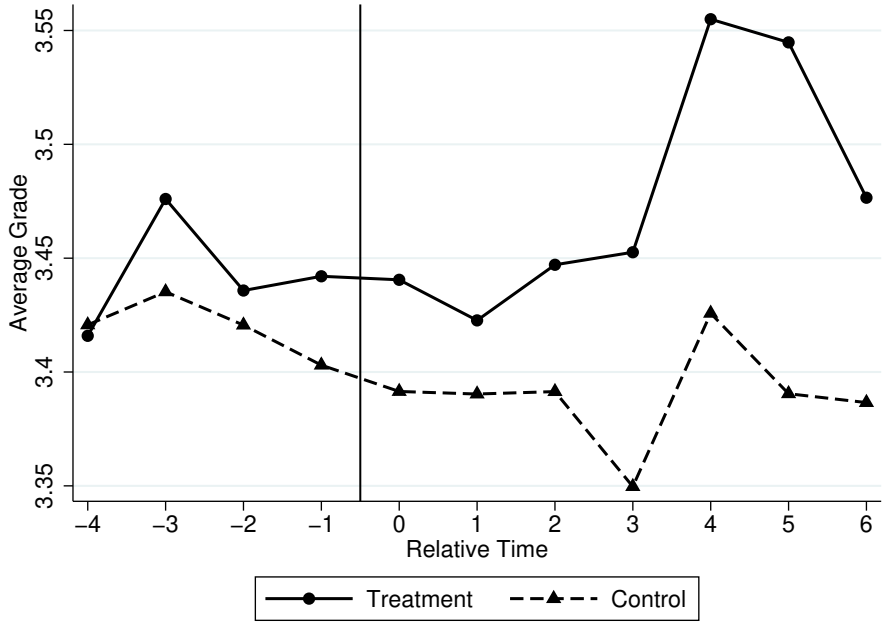


Figure 4: Trend in Average Exam Grades

Note: The figure charts the average grade attained on the written final exit exam, by cohort and treatment status. Circles (triangles) represent averages for students (not) exposed to a school-choice reform in at that relative time point.

To mitigate such concerns, I leverage a third difference that exploits a within-treatment placebo group to construct a triple-difference (DDD) model. Specifically, I consider the supply of schools in a given commuting zone, as detailed in Section 3.1, and make use of those students whom I define as not having a real choice of schools. Those students are in principle treated, because the statutory right to school choice is given to all students in the county, but the minimal supply of feasible options makes them *de facto* nontreated. However, they are exposed to the same confounders and investments potentially underlying the trends depicted in Figure 4 as the other students within a specific treatment unit. A triple-difference model relaxes the parallel-trends assumption by adding a second control group that is on the same trend as the treatment group because they are both part of the same treatment units, thus taking out the variation in outcomes attributable to the trend rather than to the policy change. The triple-difference model thus estimates the exam-performance gap between those with and without choice in the treated units, relative to the corresponding gap in the control units — and, moreover, it determines whether this gap changes in posttreatment periods. That is, we identify a treatment effect if the choice/no choice performance gap increases more posttreatment in the treatment units than in the control units. The identifying assumption in this case is therefore that the trend in the choice/no choice gap in exam performance is parallel between treatment and control groups in the pretreatment period. The triple-difference estimate thus accounts not only for changes that occur within the treatment group before and after treatment relative to the control group, but also for changes within the treatment group between students who should and should not be affected by the treatment.

I assess the validity of this assumption in Figure 5, where I chart the raw difference in grades attained between students defined as having a choice of schools and those defined as having no such choice, separately by time relative to the implementation of school-choice reform and to treatment status. Although there is a slight indication of anticipatory effects in the treatment group in the final pretreatment period (perhaps because students and parents in urban areas are more attuned to ongo-

ing discussions about a possible school-choice reform), the trends in the treatment and control groups prior to the reforms are reasonably parallel — clearly more so than in the double-difference case. It is evident that the difference in performance between students living in commuting zones with a large versus small supply of schools is stable over the sample period in the nonreforming counties (the control group). By contrast, the corresponding gap increases sharply in posttreatment periods in the treatment group, which would suggest a treatment effect.

I estimate the treatment effect more formally by extending equation (3) with the third difference and then estimating the following model using

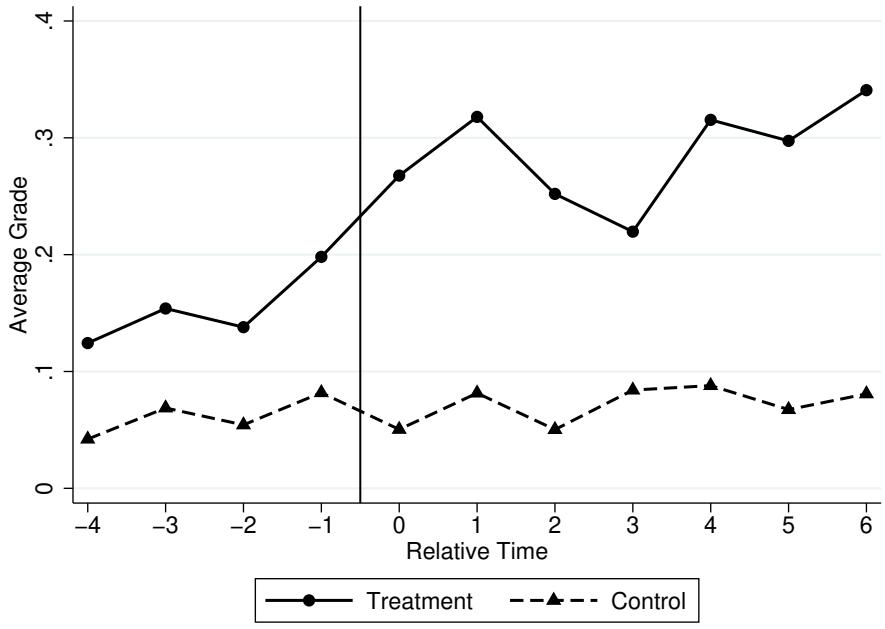


Figure 5: Trend in Choice/No Choice Differential in Average Exam Grades

Note: The figure charts the difference across *choice* status in average grade attained on the written final exit exam, by cohort and treatment status. Circles (triangles) represent averages for students (not) exposed to a school-choice reform in that particular relative time point. Higher values on the y-axis indicate a larger gap in favor of students in *choice* commuting zones.

ordinary least squares:

$$y_{izct} = \alpha_c + \lambda_t + \mu D_{c,t,z}^{Choice} + D_{c,t} + \theta_z \cdot \alpha_c + \theta_z \cdot \lambda_t + \theta_z + \varphi_i + \upsilon_{izct} \quad (5)$$

As before, the dependent variable is the (standardized) grade attained in the written exit exam in compulsory school by student i in commuting zone z in county c, observed in year t, and α_c and λ_t are vectors of unit and time indicators. The binary indicator $D_{c,t}$ takes the value 1 for students graduating in a treated county after a school-choice reform took effect. The third difference is represented by the indicator variable θ_z, which takes the value 1 for students going to school in commuting zone z if and only if that zone has more than two high schools. The variable of interest is thus $D_{c,t,r}^{Choice}$,which is an interaction between $D_{c,t}$ and θ_z where the parameter $\hat{\mu}$ captures the DDD estimate of the effect of imposing high-stakes grades. The triple-difference estimator is thus essentially a three-way interaction between α_c, $\lambda_t$and θ_z. The interaction $\theta_z \cdot \alpha_c$ controls for county-specific differences in outcomes between students living in a commuting zone with real school choice and those not living in such an area, while $\theta_z \cdot \lambda_t$ controls for the possibility that students with real choice have a different linear time trend from those without choice. To control for other predictors of academic achievement, I also add a vector of student-level covariates, represented by φ_i, to all my models. This includes gender, year of birth, immigrant status, parental education, parents' age when the student was born, and household income. In most specifications, I also control for being tested in mathematics as well as for subject-specific time trends.

Event Study Analysis

My primary mode of analysis will involve decomposing the aggregate results obtained with the framework outlined above using an event-study type design. There are two reasons for this approach. First, estimating treatment effects for individual periods leading up to or following the treatment point allows a more formal investigation of the validity of the parallel-trends assumption than merely inspecting descriptive trends in

outcomes. The presence of statistically significant treatment effects in the prereform periods would suggest that other confounding variables could be correlated with either treatment or choice status and thus bias the results.

Second, recent studies have highlighted that, in DID designs where the timing and length of treatment exposure vary between units, estimates of aggregate treatment effects represent a weighted average of all the possible two-by-two DID estimators in the sample, which can yield biased results that are intuitively hard to interpret (Callaway and Sant'Anna, 2020; Goodman-Bacon, 2021). For instance, the implicit weights assigned to each estimator are given by relative unit sizes and by the variance of the treatment indicator, that is, the timing of the treatment relative to the sample period. These weights can be unreasonable; for example, they might have negative values (de Chaisemartin and D'Haultfaeuille, 2020). In such cases, an event study or "stacked" DID design might be a more appropriate approach (Goodman-Bacon, 2021). The potential bias inherent in DID and DDD designs with variation in treatment timing can be particularly problematic if the treatment effect is not homogeneous across units and/or not static over the posttreatment period (Borusyak and Jaravel, 2018; Sun and Abraham, 2020). However, in such cases, even event-study designs can suffer from biased estimates as a result of an unreasonable implicit weighting of the estimators.

To overcome this issue, I follow the procedure introduced by Sun and Abraham (2020) to estimate an interaction-weighted (IW) triple difference model. A conventional event-study design decomposes a binary treatment indicator into a set of leads and lags, each of which is interacted with the treatment to achieve period-specific treatment effects at various points in the window around the treatment occurrence, such as in the following equation.

$$\begin{aligned} y_{izct} = \alpha_c + \lambda_t + \sum_{l=-4}^{-2} \mu_l D_{c,t,z}^{l,Choice} + \sum_{l=0}^{L} \mu_l D_{c,t,z}^{l,Choice} + \sum_{l=-4}^{-2} \mu_l D_{c,t}^{l} \\ + \sum_{l=-4}^{-2} \mu_l D_{c,t}^{l} + \sum_{l=0}^{L} \mu_l D_{c,t}^{l} + \theta_z \cdot \alpha_c + \theta_z \cdot \lambda_t + \theta_z + \varphi_i + \upsilon_{izct} \end{aligned} \tag{6}$$

In Equation (6), the four sets of variants of $\sum_l^L \mu_l D_{c,t,z}^l$ are the binary indicators taking the value 1 if the focal student in commuting zone z in county c in time t graduates l periods from the implementation point of the reform (with Choice denoting whether or not commuting zone z has more than two high schools). Such a specification relaxes the assumption that the treatment effect is static posttreatment, allowing estimates to take a nonparametric functional form across periods. However, note that when we estimate a model such as (6), we also assume that the treatment effect is homogeneous across treatment units for a given l, meaning that the period-specific estimates for all units follow the same dynamic path for $l \geq 0$. If the treatment units are in fact heterogeneous in terms of baseline characteristics, this assumption quickly becomes unreasonable. Sun and Abraham (2020) propose an alternative procedure that allows the treatment effect to vary both across time and across treatment units. Instead of a model specification like (6), they suggest estimating the cohort-specific average treatment effect, $CATT_{e,l}$, for each treated unit $e = 1, \dots, 6$ and then taking the weighted average of the relevant units in l, with the weights determined by the sample share of each unit.[20] Rather than estimating the indicators $\sum_l^L \mu_l D_{c,t,z}^l$ I thus estimate the set of $CATT_{e,l}$ given by $\sum_e \sum_{l \neq -l} \delta_{e,l}(\mathbf{1}\{E_C = e\} \cdot D_{c,t,r}^l)$ (and, correspondingly, by $\sum_e \sum_{l \neq -l} \delta_{e,l}(\mathbf{1}\{E_C = e\} \cdot D_{c,t,r}^{l,Choice})$) in (6), where the resulting coefficient $\hat{\delta}_{e,l}$ is the estimated $CATT_{e,l}$ for unit e in period l. For all l, I then take the sample-share-weighted average across the relevant e to get the IW DDD estimate $\hat{v}_l$ for the observations in the lth period relative to the treatment timing.

[20] In this study, the treated units e are the subsample of counties C that implemented school-choice reform.

4 Results

4.1 Event Study Analysis

I begin my discussion of results by presenting the estimates from the event study outlined in the previous section. First, the results from the IW event-study model are depicted in Figure 6. I report coefficients and standard errors from both this and the conventional event-study model in Table 2.[21]. As per convention, I set the period immediately prior to treatment, $l = -1$, as the reference category. Depicted in the figure is the output of an event study of the period-specific estimates, $\hat{v}_l$, of the treatment effect of taking your exit exam in the lth period relative to the implementation of high-stakes grades. Two things are evident from this figure. First, there is scant evidence of any anticipatory effects. In particular, the estimates for $l = -4$ and $l = -2$ are very close to zero. The point estimate for $l = -3$ is negative and slightly larger in magnitude, but nonetheless it is not statistically significant. This could indicate — but does not provide strong evidence in favor of — slight differences in trends between the treatment and control groups in the early periods, but convergence in the period immediately prior to implementation. In contrast, I find a moderately sized point estimate of 3.9 percent of a standard deviation (0.039σ), significant at the 10% level, for $l = -3$ when using the traditional event-study specification. This suggests that one of the treated units for which the parallel-trends assumption holds less well is overemphasized in the model. However, application of the sample-size re-weighting approach offered by IW DDD makes this anticipatory effect disappear in the aggregate.

Second, there is a clear dynamic response to the implementation of high-stakes grades: first a sharp immediate response, which then fades, but is followed by continually increasing point estimates as we move further away from $l = 0$. The immediate effect is substantial, with a significant estimate of 0.07σ. However, the period-specific estimates peak for the cohorts graduating five years after the reforms, for which I esti-

[21] Full results, including all $\hat{\delta}_{e,l}$, are available in Table D.1 in the appendix.

mate a treatment effect of 0.10σ. Such an increasing effect size suggests that younger cohorts of students adapt to the new incentive over time, perhaps as the culture and focus within schools change as well.[22] The sharp increase in point estimates is in fact apparent only once the fully treated cohorts — that is, those that through grades 8–10 under the new regime — enter the sample. On the other hand, the quickly dissipating immediate effect might suggest that the reforms and their potential effects were highly salient for the first affected cohort (owing to media attention, uncertainty about how it would affect school enrollment in the short term, etc.) but less so for the second and third cohorts.

Despite the concerns outlined in Section 3.5, the coefficients reported in Table 2 do not indicate that the difference between the IW and a conventional event-study approach is large. In the third column, I report p-values from tests of whether the estimates from these different approaches are significantly different. I find that this is the case only for $l = -3$. For all other l, I find broadly similar estimates, suggesting that the conventional event-study approach would be a reasonable approach for this context. Nevertheless, the IW DDD remains my preferred event-study approach throughout the paper, because of its more beneficial properties and assumptions.

[22] An alternative explanation for this pattern of effects could be that the composition of the treatment group changes toward the end of the sample window, as not all treated units are observed in all relative time periods. If the units with the strongest response are also those observed in later relative periods, this could potentially give a false impression of this increasing treatment effect. To assess the validity of this concern, I re-ran the analysis using different compositions of the treatment group; the results are reported in Table B.1 in the appendix. Specifically, I re-estimated the model separately using only the first three cases (the "early adopters") and the last three cases (the "late adopters") in the treatment group, respectively. I also ran a model where I used the middle four cases for which I could create a balanced sample window where all treated units are observed in all relative time periods. The results from these exercises indicate that, although it is apparent that the early and middle adopters are driving the observed effects, they themselves display this dynamic increase in effect sizes. Hence the shape of the event-study model does not seem to be an artifact of a changing composition of the treatment group, but rather a reflection of the dynamics within the units most strongly affected by the reforms.

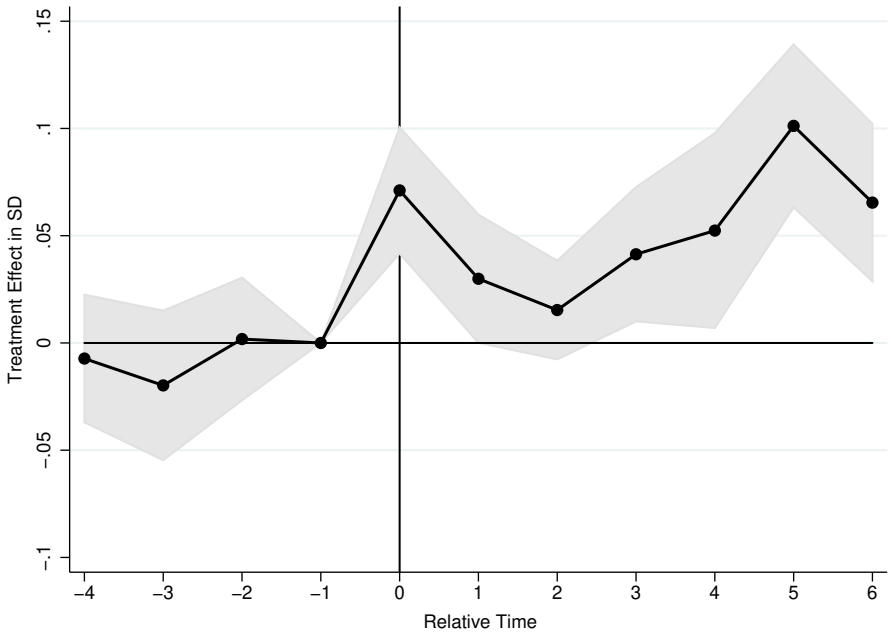


Figure 6: Event Study Estimates of the Effects of School Choice Reforms on Exam Grades

Note: This figure presents the results from estimating an event-study type model decomposing the dynamics of the treatment effect over periods leading up to, and following implementation of the reforms. Reported are the coefficients estimated for indicators for being l periods removed from the treatment, where $l \in \{-4, 6\}$. The model is saturated in period indicators as the indicator for the first and last periods takes the value 1 for all preceding/subsequent periods, respectively. $l = -1$ is omitted as the reference category. The shaded area represents 95% confidence intervals . I report full results in Table D.1 in the appendix.

4.2 Aggregate Results

In this section, I present aggregate estimates of the average treatment effect of implementing high-stakes grades for the posttreatment period as a whole. Results from estimating the triple-difference model (5) using ordinary least squares are presented in Table 3. For ease of exposition, I report only the estimated coefficients for the three key parameters — the indicator for school-choice reform (in essence the *Treat*×*Post* interaction), the indicator for choice, and the triple interaction. First, in Column 1 I present results from my preferred specification where I regress the standardized exam performance for each student on 5, controlling only for student characteristics, parental background, and socioeconomic status (as described in Section 3.1.1). Using this specification, I estimate an average increase in the exam grade attained of 0.053σ. For an intuitive comparison of the effect size, 0.053σ is about half the estimated performance gap between native and immigrant students using this specification. In line with the identifying assumption of my triple-difference model, I cannot reject the null hypothesis of no effect for the school-choice-reform indicator alone. These results imply that imposing high-stakes grades through school-choice reforms is effective in improving student performance if combined with sufficient levels of choice so that the grades are actually perceived as consequential.[23]

In Column 2, I re-estimate the model, adding an indicator of whether a student was tested in mathematics as well as a subject-specific time trend. Using this specification, I estimate a treatment effect of 0.048σ — somewhat smaller, but substantively similar to the result in Column 1.

[23] An alternative to this approach would be to estimate a more conventional double-difference model, and to subsample on the choice condition. Doing so yields broadly similar results, with a DID estimate of the effect of the reforms of 0.042σ, significant at the 5% level, for the *choice* subsample, and a nonsignificant estimate of -0.011 for the *no choice* subsample.

Table 2—Event Study Analysis

Relative time	DDD estimates $\hat{\mu}_l$	IW DDD estimates $\hat{v}_l$	Difference p-value
-4	-0.022	-0.007	0.671
	(0.039)	(0.015)	
-3	0.039*	-0.020	0.001
	(0.020)	(0.018)	
-2	0.013	0.002	0.613
	(0.026)	(0.015)	
-1	Omitted	Omitted	
0	0.065**	0.071***	0.816
	(0.032)	(0.015)	
1	0.046	0.030*	0.580
	(0.036)	(0.015)	
2	0.056	0.015	0.178
	(0.035)	(0.012)	
3	0.034	0.041**	0.811
	(0.036)	(0.016)	
4	0.071*	0.052**	0.574
	(0.037)	(0.023)	
5	0.092	0.101***	0.873
	(0.057)	(0.019)	
6	0.070*	0.065***	0.903
	(0.037)	(0.019)	
N	790,905	790,905	
Adj. R^2	0.214	0.215	

Note: Estimation of the timing of treatment effects using a conventional event-study design and the Sun and Abraham (2020) IW event study approach. For this estimation, treatment status is replaced with an indicator equal to one in that particular year only, except $l = -4$ and $l = 6$, which are one for all preceding/subsequent years. The year prior to implementation is omitted for reference. In the *Difference* column I report p-values from tests of whether $\hat{\mu}_l$ and $\hat{v}_l$ are significantly different. Errors clustered at the commuting-zone level in parentheses. * $p<0.1$, ** $p<0.05$, *** $p<0.01$.

Table 3—Aggregate Results

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
School choice reform ×	0.053**	0.048**	0.067**	0.050**	0.043*	0.054**	0.051***
Choice	(0.024)	(0.025)	(0.030)	(0.023)	(0.025)	(0.026)	(0.13)
School choice reform	-0.010	-0.010	0.003	-0.016	-0.008	-0.007	-0.029**
	(0.014)	(0.016)	(0.020)	(0.012)	(0.016)	(0.015)	(0.011)
Choice	0.013	0.014	0.005	0.158*	-0.011	0.019	0.071
	(0.059)	(0.068)	(0.064)	(0.090)	(0.039)	(0.061)	(0.056)
N	790,905	790,905	790,905	526,303	615,454	750,264	790,905
Adj.R^2	0.173	0.221	0.221	0.181	0.172	0.174	0.215
Covariates	✓	✓	✓	✓	✓	✓	✓
Subject FE + trend		✓	✓				✓
Linear trend			✓				
IW DDD							✓
Excluding:							
Always treated				✓			
Never treated					✓		
Year = 2008						✓	

Note: The table presents estimates of the average treatment effect on exam grade of imposing high-stakes grades through merit-based school-choice admission schemes. The outcome variable is standardized to have a mean of 0 and a standard deviation of 1. Panel A reports results from estimating the DDD model specified in (5). The coefficient of interest is the three-way interaction School choice reform × Choice in the top row, which gives the average treatment effect of being a student graduating from a treated county, in a labor market region with more than two high schools, after the treatment has been implemented. Conversely, the School choice reform variable controls for the conventional two-way fixed effects difference-in-differences estimator of graduating from a treated county in a posttreatment year. Choice is a dummy equal to one for students who have more than two high schools within traveling distance from their home. The triple difference model in practice interacts the DID estimator with this dummy. The models in Column 4 and 5 exclude all observations from always-treated and never-treated counties, respectively. In Column 6 I exclude all observations from the year 2008 from the regression. In Column 7 I aggregate the IW DDD event study results following the procedures suggested by Callaway and Sant'Anna (2020). Cluster-robust standard errors clustered at the commuting-zone level in parenthesis. * $p<0.1$, ** $p<0.05$, *** $p<0.01$.

As a robustness check, in Column 3 I further examine if the treatment counties were on a differential trend before the reforms were implemented by controlling for a treatment-specific linear trend. In doing so, I relax the parallel-trends assumption to see if such differences are driving the results. As is evident from the estimate, controlling for such a trend increases the key point estimates by 0.014σ relative to the preferred specification, while the other parameters remain virtually unchanged. This substantiates the notion that the effect estimated in fact stems from the school-choice reforms, and not from some other underlying trend specific to the treatment counties. If anything, such (unidentified) underlying trends would appear to depress the initial estimate of the treatment effects.

When I use the full sample of students available, all nonreforming counties are designated as controls. This includes both counties that already had school-choice systems in place at the start of the sample period and counties that applied a neighborhood-catchment regime for the duration of that period. Given the dynamic path of the treatment effects illustrated in Figure 6, the presence in the control group of counties that had already implemented similar reforms prior to the start of my sample period (the "always-treated") could potentially bias the results (Goodman-Bacon, 2021). For example, some counties could have implemented similar reforms a short time before the start of my sample period and thus be on a similar dynamic trend. To check if the results are sensitive to the control-group specification, Columns 4 and 5 exclude always-treated and never-treated counties, respectively. Hence in Column 4 the outcomes for students in the six treated counties are considered only in relation to students in those counties that never implemented high-stakes grades. Conversely, Column 5 estimates the same model using only those counties that were already "treated" prior to the start of my sample period. As evident from the results in Table 3, neither approach changes the substance of the results: while point estimates in both cases are smaller, they remain very close to, and are not significantly different from, those of the main model. When the never-treated counties are excluded, the p-value of the estimate does fall below the conventional 5 percent level, but only just. In Column 6, I further consider whether the result is robust

to dropping all observations from 2008 from the sample (in that year, a large teachers' strike caused about one-third of exit exams to be canceled, primarily those in Norwegian; see Section 3 for details). It turns out that dropping the observations from that year does not impact the estimates in any meaningful way.[24]

Finally, in Column 7 I aggregate the event-study treatment effects derived from the IW DDD approach. I follow the approach suggested in Callaway and Sant'Anna (2020) by averaging the group–time specific effects for each treated unit (i.e. averaging over $\hat{\delta}_{e,l}$ for each e) before averaging across units using the sample-share weights derived in the event-study analysis. The resulting parameter is the average effect of being exposed to the reforms experienced by all units that were ever exposed (Callaway and Sant'Anna, 2020, p. 12). As was the case in the event study, using the IW approach does not move the estimates in any way that would cause the conclusions to change.

Overall, the results presented in Table 3 are consistent with the hypothesis that students are incentivized by the prospect of being able to choose high schools given adequate academic performance. They are also consistent with the notion that, since a prerequisite for this mechanism to be effective is having several options within a reasonable commuting distance, students in treated counties but in commuting zones with few choices are viable as a control group. The nonsignificance of the point estimates for the school-choice-reform indicator supports this conjecture. Similarly, having many schools within traveling distance does not in and of itself seem to have an effect on performance. It is only when a sufficiently large supply of schools is combined with school-choice reform that grades are actually perceived as consequential, which boosts students' performance. What is more, these results do not appear to be sensitive to the specific choice of school-supply threshold. Figure B.1 in the appendix indicates that the effects are similar — if anything larger — when the choice

[24] I report results from additional robustness checks in Appendixes B and C. For example, I consider alternative approaches to computing the standard errors, such as clustering at the county level (and performing few-clusters corrections) and using randomization inference rather than conventional t-tests. The results are robust to these alternative approaches.

threshold is set higher. In sum, this exercise suggests that the result is not an artifact of my definition of what constitutes real choice. Rather, it reinforces the notion that the choice set of schools must be sufficiently large to create a competitive market that incentivizes students, suggesting that this effect may increase with the supply of schools. Setting the threshold at three thus represents a conservative constraint.

5 Mechanisms

5.1 Learning vs Test Effort

The results reported in Section 4 suggest that there is a mechanism by which test scores are influenced by the imposition of higher stakes. From a policy perspective, however, our main interest lies not in test scores *per se* but in students' accumulation of human capital. Indeed, one of the main purposes of testing is to measure the extent to which students have learned the skills they are supposed to learn. However, several papers have pointed out that scores on tests involving low stakes will reflect not only students' ability but also their motivation and effort (Gneezy et al., 2019; Heissel et al., 2021; Segal, 2012; Zamarro et al., 2019). One potential explanation for the difference observed in the present study between treated and nontreated students could therefore be that those students do not really differ in human capital but that what distinguishes them is that the treated ones have a stronger incentive than the nontreated ones to put effort into the exit exam and hence are likely to obtain better grades.

To explore whether the results reflect a sustained learning effort or mere test effort, I exploit the fact that, for the past decade, the Norwegian Ministry of Education has required all students to take a national standardized assessment test in grades 5, 8, and 9, the latter test being specifically implemented to measure students' improvement over the first year of the second stage of compulsory school. These tests are meant to provide a comprehensive assessment of a student's ability level at that point in time, providing school managers and policymakers with a tool

enabling them to determine where resources and measures should be directed in order to improve student outcomes. For the students, however, there are no formal consequences associated with the tests. Their test scores do not factor into their grades, do not appear on any transcript, and are available only to their teacher and to their parents. Hence these tests are low-stakes in nature. According to economic theory, the rational decision for a student, assuming that effort is costly, is therefore to devote less effort to such tests than to high-stakes test such as the exit exam. This, in turn, would imply that scores on these assessment tests may not adequately reflect students' true ability. Importantly, this does not change as a result of school-choice reforms. Consequently, if turning the final exit exam in grade 10 into a high-stakes test affects the effort students make to learn throughout the second stage (grades 8–10; "lower-secondary school") and not just their effort ahead of and during that exam, this should be observable in the development of scores on the national assessment test. In other words, if students subjected to high-stakes grades put in more effort to learn, at least from the start of grade 8, they should have improved their ability level between grades 8 and 9 more than other students. If this is so, this would imply that the incentives provided in order to increase effort have actually worked by placing those students on a higher learning trajectory than they would otherwise be on.

I test the hypothesis outlined above by estimating triple-difference models similar to those used in the main analysis as described in Section 4, with scores on the national assessment test in grade 9, that is, in the year prior to the year of graduation, as the outcome of interest. As these tests were introduced for ninth-graders in 2010, the analysis is restricted to the 2010–2015 cohorts. I match grade 9 observations to the same students' scores in grade 8, so that I can control for previous performance. I include students missing tests from eighth grade by constructing an indicator equal to one if the subject score is missing and setting the score to zero. Within the sample period, three counties implemented school-choice reforms (in 2012 and 2014, respectively). This provides a staggered DDD framework similar to that previously used. All students are tested in both mathematics and Norwegian language/reading in both grade 8 and

Table 4—National Assessment Test Event Study Results

Relative time	DDD estimates $\hat{\mu}_l$	IW DDD estimates $\hat{v}_l$	Difference p-value
-2	0.023	0.039	0.571
	(0.033)	(0.033)	
-1	Omitted	Omitted	
0	-0.011	0.010	0.376
	(0.027)	(0.023)	
1	0.030	0.053***	0.290
	(0.024)	(0.019)	
2	0.053	0.071*	0.362
	(0.035)	(0.042)	
3	0.055	0.070**	0.685
	(0.038)	(0.027)	
N	249,602	249,602	
Adj. R^2	0.767	0.753	

Note: The table presents results from a triple difference event-study analysis using performance on the standardized national assessment tests in mathematics and reading in 9th-grade as the outcome. The event study decomposes the results over the years leading up to, and following, the implementation of the reforms using both the conventional, and the Sun and Abraham (2020) IW event-study approach. I standardize the score of each test, take the mean, and standardize the resulting composite score. The outcome is thus a representation of the general skill level of the student in subjects applicable for the final exam. For these estimations, treatment status is replaced with an indicator equal to one in that particular year only. The year prior to implementation is omitted for reference. In the *Difference* column I report p-values from tests of whether $\hat{\mu}_l$ and $\hat{v}_l$ are significantly different. Cluster-robust standard errors clustered at the commuting-zone level in parentheses. * $p<0.1$, ** $p<0.05$, *** $p<0.01$.

grade 9.[25] To construct my outcome measure, I standardize the scores on each test, average them across the tests, and standardize the resulting average score once more. This composite score is thus a measure of a student's general skill level in the subjects covered by the final exit exam.

I present the results from this analysis in Table 4. That table includes estimates from event studies similar to those described in Section 4.1, decomposing the triple-difference results into leads and lags using both

[25]Students are also tested in English in grade 8, and I include those scores as well in the controls.

the Sun and Abraham (2020) interaction-weighted design and the conventional event-study specification. As previously, I set $l = -1$ as the reference category. For these grade 9 assessment tests, the period-specific estimates display a similar dynamic evolution in terms of effect size as was observed for the exit-exam grades (Figure 6). This is inconsistent with the idea that the improvements in test scores result only from changes in the amount of effort spent on the assessment tests themselves, as such an effect should be observable immediately upon implementation and then remain stable. In fact, I find no effect on the scores of those students who took the assessment tests immediately after the implementation of high-stakes grades. On the other hand, for the cohort of students who where in grade 8 when the reforms were implemented, meaning that they had ample time to adjust their effort levels to the new regime, I find a substantial increase in the composite-score measure. Strong effects are also evident for subsequent cohorts, amounting to approximately 0.070σ (unfortunately, the sample period does not allow me to extend the analysis further into the posttreatment period). The fact that these effect sizes appear with a similar dynamic rhythm as the increases in effect sizes in the main analysis lends support to the claim that the main treatment effect observed in scores on the final exit exam is not solely attributable to test effort, but is also explained by an increase in what students have actually learned — that is, in their ability level.[26]

5.2 Interactions Analysis

The channels through which the effect of this incentive might work could also be illuminated by its differential effects across subsamples. For example, a widely accepted notion is that a more competitive environment in schools will benefit boys, who tend to thrive more than girls under such

[26]In the appendix I also report results from a similar analysis using the test scores in grade 8 as the outcome. In this case, there is no clear pattern to the results — if anything students appear to do somewhat worse after reform implementation, suggesting that the change in behavior starts upon entry to lower-secondary school, not in earlier grades. This is consistent with the notion that lower-secondary school marks a new stage in the students' trajectory, where grades and future academic paths are more strongly emphasized.

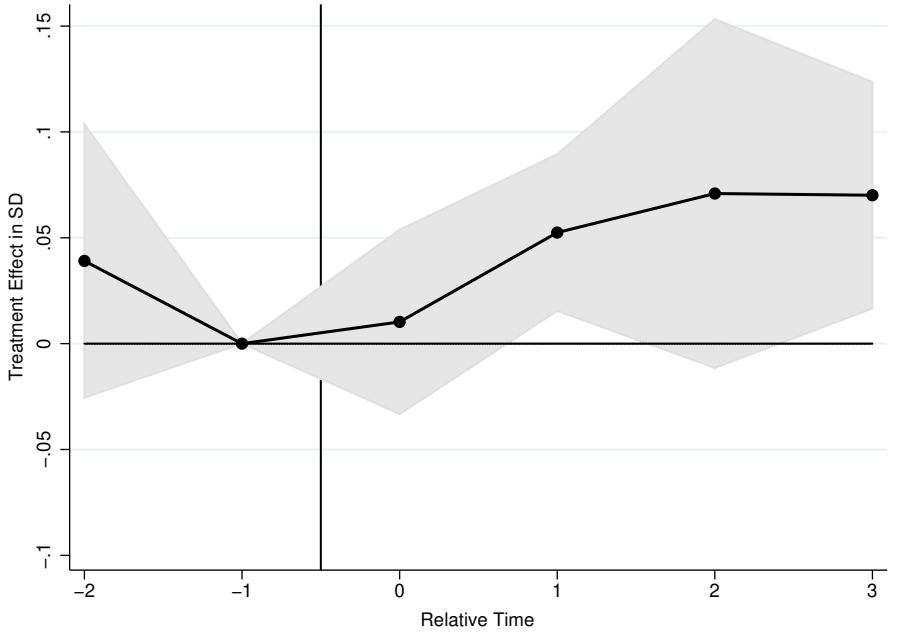


Figure 7: Event Study Results for Assessment Test Scores in 9th-grade

Note: This figure presents the results from estimating an event-study type model decomposing the dynamics of the treatment effect of introducing high-stakes grades in 10th grade on low-stakes assessment tests conducted in 9th grade. Reported are the coefficients estimated for indicators for being l periods removed from implementation, where $l \in \{-2, 3\}$. The model is saturated in period indicators as the sample period is constrained to the 6-year window in question. $l = -1$ is omitted as the reference category. The shaded area represents 95% confidence intervals. Full results are available in Table D.2 in the Appendix.

conditions (Almås et al., 2016; Azmat et al., 2016; Hopland and Nyhus, 2016). Certain other subsamples are also of particular policy interest, including students from a low socioeconomic background. Socioeconomic status (SES) is a major predictor of educational achievement, and there is a large body of research into interventions at the compulsory-school level aimed at improving the performance of students from low-SES households (Dietrichson et al., 2017). Evidence that such typically at-risk students respond positively to high-stakes grades — learning more in the process — would therefore have obvious policy implications. Moreover, Almås et al. (2016) demonstrate that there is a strong socioeconomic gradient in terms

of competition preferences. In particular, boys from lower-SES households are less willing to compete than boys from higher-SES backgrounds. If we believe that the competitive pressure created by high-stakes grades is the driving mechanism behind the observed increase in performance, that increase could therefore also reflect an adverse segregational effect across parental background in that boys from richer homes may benefit to a particularly large extent.

In the following analyses I also consider whether students who were tested in mathematics at the exit exam are more impacted by the treatment than others. As students take only one exam, the subject they are allocated can greatly influence their performance, all else being equal. Generally, students tested in mathematics perform far worse than those tested in a language subject, as illustrated in Figure 3. In this particular case, it is plausible that mathematical skills can be improved more by high-effort behavior such as cramming and repetition, and may thus be more responsive to high-stakes grades. Conversely, language skills may be harder to improve through effort alone, in that they require a longer-term maturation process. This hypothesis takes into account evidence suggesting that students' vocabulary and language skills are strongly tied to their parental background (Buckingham et al., 2013; Dustmann, 1997), and that scores on language tests often appear to be less receptive to interventions than scores on mathematics tests (Bettinger, 2012).

For this purpose, I extend (5) to incorporate either gender or SES as a fourth dimension, to estimate a quadruple-type model of the form [27]

$$\begin{aligned} y_{igzct} = \mu_1 D_{c,t,z,g}^{Choice} &+ \sum_{\beta=2}^{5} \mu_\beta (4 \text{ triple interactions}) \\ &+ \sum_{\beta=6}^{11} \mu_\beta (6 \text{ double interactions}) \\ &+ \sum_{\beta=12}^{15} \mu_\beta (4 \text{ linear terms}) + \varphi_i + \upsilon_{izct} \end{aligned} \tag{7}$$

[27] This presentation and specification of the quadruple-difference estimator follows the approach used by Muralidharan and Prakash (2017).

$D_{c,t,r,g}^{Choice}$ takes the value 1 if student i of gender (SES) g in commuting zone z in county c in cohort t takes her exam in a treated county after a school-choice reform has been implemented there, and her commuting zone has more than two high schools. In the model I control for all possible interactions among the four main variables, represented by $\sum_{\beta=2}^{15} \mu_\beta$, and for student-level characteristics φ_i. I estimate the model separately for the full sample and for the subsamples tested in mathematics and language, respectively.

Table 5 presents the results from estimations of the quadruple-difference models. Panel A reports results for the gender specifications. Evidently, the estimates do not indicate any gender-specific differential effects of the admission reforms. While I find large and significant point estimates for the effect of the reform in general, the estimates for the differential effect on girls are small and statistically insignificant. This is the case both for the overall sample and across exam subjects. As the top row reports the marginal effect of being a treated girl, the coefficients for *School choice reform* × *Choice* give the average treatment effect for treated boys. Columns 1 and 2 indicate that boys randomly drawn to be tested in mathematics respond more strongly to the reforms than those tested in language, but these estimates are imprecise and not significantly different from each other.

Panel B considers low-SES students, defined as having a mother whose highest completed level of education is at most compulsory school (which is true for 22.7% of the sample). Following Almås et al. (2016), we would expect these students to respond less strongly to a competitive incentive and hence to manifest smaller treatment effects. However, as with gender, I find limited evidence of such a differential effect using the quadruple-difference model. As reported in Table 5, I find small positive coefficients for both the total sample and the language subsample. Although neither is close to being statistically significant, in both cases the direction of the estimate is the opposite of what the literature would have us expect. This is also the case for the mathematics subsample, for which I find a moderately sized point estimate of 0.054σ. At face value, such an estimate suggests that treated low-SES students who were tested in mathematics

increased their performance more than treated students with other socioeconomic backgrounds who were also tested in mathematics. While this estimate is also imprecisely estimated, it provides a suggestive piece of evidence that, if anything, the reforms served to *reduce* the SES gap in performance on the mathematics exam.

Overall, however, the conclusion to be drawn from the analysis presented in this section is that I find limited evidence of differential treatment effects across important subsamples. Instead, the positive effect of the admission reforms on student performance seems to be rather uniform across the subsamples considered here, with some evidence that the effect is stronger for students tested in mathematics, in particular for those with a low-SES background. It would appear that these results should mitigate our concern regarding the possibility of strong segregational effects of school-choice policies such as those studied in this paper.

6 Concluding Remarks

In this paper, I investigate the incentivizing effect of high-stakes grades on student learning. I exploit a natural experiment created by regional differences in Norwegian high-school admission regimes to compare scores on the final exit exam of compulsory school, which is a high-stakes exam for some students but not for others. I use the supply of schools within students' traveling distance as a third source of variation, to distinguish students who have a real choice of schools from those who have such a choice only in theory. In line with theory-based predictions, my triple-difference model reveals that tying the final exit exam of compulsory school to salient outcomes improves the grades attained, with an effect size of 5–6 percent of a standard deviation. The effect size is moderate, but it is still economically meaningful. For example, the magnitude is equal to about 20% of the unconditional gender gap in exam performance, and to 10% of the SES gap. While several papers have demonstrated a causal link between test stakes and performance, either through smaller field experiments or by using financial incentives, this paper provides evidence for the

Table 5—Interactions Analysis

	All	Math	Language
	(1)	(2)	(3)
Panel A: Gender			
School choice reform × Choice × Female	-0.028	-0.015	-0.017
	(0.022)	(0.040)	(0.022)
School choice reform × Choice	0.080***	0.099*	0.064**
	(0.027)	(0.056)	(0.028)
School choice reform	-0.037**	-0.024	-0.048**
	(0.016)	(0.037)	(0.024)
School choice reform × Female	0.005	-0.021	0.006
	(0.019)	(0.038)	(0.018)
Choice × Female	0.042	0.045	-0.021
	(0.030)	(0.039)	(0.028)
Female	0.396***	0.180***	0.494***
	(0.022)	(0.036)	(0.024)
N	790,905	297,414	493,491
Adj.R^2	0.174	0.214	0.181
Panel B: Socioeconomic Status			
School choice reform × Choice × Low SES	0.013	0.054	0.014
	(0.028)	(0.033)	(0.036)
School choice reform × Choice	0.056**	0.075	0.043
	(0.028)	(0.056)	(0.029)
School choice reform	-0.015	0.004	-0.034
	(0.014)	(0.033)	(0.022)
School choice reform × Low SES	-0.020	-0.087***	0.010
	(0.021)	(0.031)	(0.024)
Choice × Low SES	0.001	0.029	-0.025
	(0.023)	(0.037)	(0.031)
Low SES	-0.802***	-1.012***	-0.735***
	(0.035)	(0.045)	(0.040)
N	771,445	289,554	481,891
Adj.R^2	0.163	0.208	0.168

Note: This table reports results from subsample analyses of differential treatment effects across gender and socioeconomic status. Column 1 estimates effects for the full sample, while columns 2 and 3 estimate identical models for those tested in mathematics or a language separately, using the preferred specification from Table 3. In panel A I consider differential effects between boys and girls. In Panel B I consider whether the effects interact with socioeconomic background. Here I use the mother's education to determine socioeconomic status, where low SES indicates that her highest level of completed education is at most compulsory school (10 years). Errors clustered at the commuting-zone level in parentheses. * $p<0.1$, ** $p<0.05$, *** $p<0.01$.

viability of exploiting such a mechanism to stimulate students' investment of effort in school. Indeed, the results indicate that the change to a merit-based enrollment regime in high school in and of itself improves performance in younger students. That is, performance improves at a stage where no tracking or sorting of any kind is conducted. However, a crucial prerequisite is that the supply of schools must be sufficient to create a sense of real choice. Introducing school choice has little impact if students have only one or two schools within a reasonable traveling distance. Further, my analysis does not find any significant heterogeneity in treatment effect across exam subject, socioeconomic status or gender – — a result that contrasts with the results of earlier studies suggesting that school-choice enrollment regimes might have adverse segregational effects (Altonji et al., 2015; Hsieh and Urqiuola, 2006; Lindbom, 2010)

Building on a growing body of work exploring the relationship between effort and performance in low-stakes assessments (Gneezy et al., 2019; Segal, 2012; Zamarro et al., 2019), I assess the extent to which my results can be explained by a sustained learning effort, as opposed to a more punctual test-taking effort, on the part of students. By contrasting performance on the final exit exam with scores on comprehensive ability assessments conducted in earlier grades, I demonstrate that students exposed to a school-choice enrollment regime appear to be on a higher learning trajectory than students in the control group. These results imply that the main treatment effect is not only a result of increased test effort but is also indicative of a higher, sustained learning effort throughout the final years of compulsory school. Evidence of students making a long-term investment in their schooling should increase the relevance of this study for policymakers. The effect sizes are nontrivial, but nevertheless moderate, which suggests that some students respond more to these incentives than others. While identifying those students lies beyond the scope of the present study, policymakers can be expected to be interested in finding out who they are, in order to thoroughly assess the distributional effects of implementing high-stakes grades.

References

Almås, I, Cappelen, A. W., Salvanes, K. G., Sørensen, E. Ø., & Tungodden, B. (2016). Willingness to Compete: Family Matters. *Management Science*, *62*(8), 2149–2162.

Altonji, J. G., Huang, C.-I., & Taber, C. R. (2015). Estimating the Cream Skimming Effect of School Choice. *Journal of Political Economy*, *123*(2), 266–324.

Angrist, J. D., & Lavy, V. (2009). The Effects of High Stakes High School Achievement Awards: Evidence From a Randomized Trial. *American Economic Review*, *99*(4), 1384–1414.

Angrist, J., Bettinger, E., Bloom, E., King, E., & Kremer, M. (2002). Vouchers for Private Schooling in Colombia: Evidence From a Randomized Natural Experiment. *American Economic Review*, *92*(5), 1535–1558.

Arbeidslaget Analyse, Utgreiing og Dokumentasjon. (2005). *Evaluering av fritt skolevalg*. Hordaland Fylkeskommune.

Athey, S., & Imbens, G. (2017). The Econometrics of Randomized Experiments. In A. V. Banerjee & E. Duflo (Eds.), *Handbook of Economic Field Experiments* (Chap. 3, Vol. 1, pp. 73 –140).

Auestad, M. L. (2018). The Effect of Low-Achieving Peers. *Labour Economics*, *55*, 178–214.

Azmat, G., Calsamiglia, C., & Iriberri, N. (2016). Gender Differences in Response to Big Stakes. *Journal of the European Economic Association*, *14*(6), 1372–1400.

Bach, M., & Fischer, M. (2020). Understanding the Response to High Stakes Incentives in Primary Education. *IZA Discussion Paper Series*, (13845).

Bakken, A., Sletten, M. A., & Eriksen, I. M. (2018). Generasjon prestasjon? Ungdoms opplevelse av press og stress. *Tidsskrift for ungdomsforskning*, (2).

Becker, W. E., & Rosen, S. (1992). The Learning Effect of Assesment and Evaluation in High School. *Economics of Education Review*, *11*(2), 107–118.

Behrman, J. R., Parker, S. W., Todd, P. E., & Wolpin, K. I. (2015). Aligning Learning Incentives of Students and Teachers: Results From a Social Experiment in Mexican High Schools. *Journal of Political Economy*, *123*(2), 325–364.

Bettinger, E. P. (2012). Paying to Learn: The Effect of Financial Incentives on Elementary School Test Scores. *The Review of Economics and Statistics*, *94*(3), 686–698.

Bind, M.-A. C., & Rubin, D. B. (2020). When Possible, Report a Fisher-Exact P-Value and Display its Underlying Null Randomization Distribution. *Proceedings of the National Academy of Sciences*, *117*(32), 19151–19158.

Borusyak, K., & Jaravel, X. (2018). Revisiting Event Study Designs, With an Application to the Estimation of the Marginal Propensity to Consume. *Working Paper*.

Brophy, J. (1987). Synthesis of Research on Strategies for Motivating Students to Learn. *Educational Leadership*, *45*(2), 40–48.

Buckingham, J., Wheldall, K., & Beaman-Wheldall, R. (2013). Why Poor Children Are More Likely to Become Poor Readers: The School Adulthood. *Australian Journal of Education*, *57*(3), 190–213.

Callaway, B., & Sant'Anna, P. H. C. (2020). Difference-in-Differences with Multiple Time Periods. *Journal of Econom*terics*, *forthcoming*.

Cameron, A., & Miller, D. (2015). A Pactitioner's Guide to Cluster-Robust Inference. *Journal of Human Resources*, *50*(2), 317–372.

Costrell, R. M. (1994). A Simple Model of Educational Standards. *American Economic Review*, *84*(4), 956–971.

Cullen, J. B., Jacob, B. A., & Levitt, S. (2006). The Effect of School Choice on Participants: Evidence From Randomized Lotteries. *Econometrica*, *74*(5), 1191–1230.

de Chaisemartin, C., & D'Haultfaeuille, X. (2020). Two-Way Fixed Effects Estimators with Heterogeneous Treatment Effects. *American Economic Review*, *110*(9), 2964–96.

Deming, D. J., & Figlio, D. (2016). Accountability in US Education: Applying Lessons from K-12 Experience to Higher Education. *Journal of Economic Perspectives*, *30*(3), 33–56.

Dietrichson, J., Bøg, M., Filges, T., & Klint Jørgensen, A.-M. (2017). Academic Interventions for Elementary and Middle School Students with Low Socioeconomic Status: A Systematic Review and Meta-Analysis. *Review of Educational Research*, *87*(2), 243–282.

Dokument 8:41. (2006). *Forslag fra stortingsrepresentantene Anders Anundsen, Jon Jæger Gåsvatn og Åse M. Schmidt om å innføre fritt skolevalg for alle elever i videregående skole*. Oslo, Norway: Stortingets Utredningsseksjon.

Dokument 8:8. (2003). *Forslag fra stortingsrepresentantene Ulf Erik Knudsen, Arne Sortevik og Karin S. Woldseth om innføring av elevers rett til fritt skolevalg på videregående skole*. Oslo, Norway: Stortingets Utredningsseksjon.

Dustmann, C. (1997). The Effects of Education, Parental Background and Ethnic Concentration on Language. *The Quarterly Review of Economics and Finance*, *37*, 245 –262.

Eccles, J. S., & Midgley, C. (1989). Stage-Environment Fit: Developmentally Appropriate Classrooms for Young Adolescents. In *Research on Motivation in Education* (Vol. 3, pp. 139–186). New York, NY: Academic Press.

Eccles, J. S., Wigfield, A., Midgley, C., Reuman, D., Iver, D. M., & Feldlaufer, H. (1993). Negative Effects of Traditional Middle Schools on Students' Motivation. *The Elementary School Journal*, *93*(5), 553–574.

Epple, D., & Romano, R. E. (2003). Neighborhood Schools, Choice, and the Distribution of Educational Benefits. In C. M. Hoxby (Ed.), *The Economics of School Choice* (Chap. 7, pp. 227–286).

Figlio, D., & Hart, C. M. D. (2014). Competitive Effects of Means-Tested School Vouchers. *American Economic Journal: Applied Economics*, *6*(1), 133–156.

Figlio, D., & Loeb, S. (2011). School Accountability. In E. A. Hanushek, S. Machin, & L. Woessmann (Eds.), *Handbook of the Economics of Education* (Chap. 8, Vol. 3, pp. 383–421).

Friedman, M. (1962). *Capitalism and Freedom*. Chicago, IL: University of Chicago Press.

Fryer, R. (2011). Financial Incentives and Student Achievement: Evidence from Randomized Trials. *The Quarterly Journal of Economics*, *126*(4), 1755–1798.

Gibbons, S., Machin, S., & Silva, O. (2008). Choice, Competition and Pupil Achievment. *Journal of the European Economic Association*, *6*(4), 912–947.

Gneezy, U., List, J. A., Livingston, J. A., Qin, X., Sadoff, S., & Xu, Y. (2019). Measuring Success in Education: The Role of Effort on the Test Itself. *American Economic Review: Insights*, *1*(3), 291–308.

Goodman-Bacon, A. (2021). Difference-in-Differences With Variation in Treatment Timing. *Journal of Econometrics*, *forthcoming*.

Grant, D., & Green, W. B. (2013). Grades as Incentives. *Empirical Economics*, *44*, 1563–1592.

Grove, W. A., & Wasserman, T. (2006). Incentives and Student Learning: A Natural Experiment with Economics Problem Sets. *American Economic Review*, *96*(2), 447–452.

Hanushek, E. A. (1986). The Economics of Schooling: Production and Efficiency in Public Schools. *Journal of Economic Literature*, *24*(3), 11–41–1177.

Harter, S. (1981). A New Self-Report Scale of Intrinsic Versus Extrinsic Orientation in the Classroom: Motivational and Informational Components. *Developmental Psychology*, *17*(3), 300–312.

Heissel, J. A., Adam, E. K., Doleac, J. L., Figlio, D. N., & Meer, J. (2021). Testing, Stress and Performance: How Students Respond Physiologically to High-Stakes Testing. *Education Finance and Policy*, *16*(2), 183–208.

Hopland, A. O., & Nyhus, O. H. (2016). Gender Differences in Competitiveness: Evidence from Educational Admission Reforms. *The B.E. Journal of Economic Analysis & Policy*, *16*(1).

Hoxby, C. M. (2003). School Choice and School Productivity: Could School Choice Be a Tide that Lifts All Boats? In C. M. Hoxby (Ed.), *The Economics of School Choice* (Chap. 8, pp. 287–342).

Hoxby, C. (2000). Does Competition Among Public Schools Benefit Students and Taxpayers? *American Economic Review*, *90*(5), 1209–1238.

Hsieh, C.-T., & Urqiuola, M. (2006). The Effects of Generalized School Choice on Achievement and Stratification: Evidence from Chile's Voucher Program. *Journal of Public Economics*, *90*, 1477–1503.

Hvidman, U., & Sievertsen, H. H. (2019). High-Stakes Grades and Student Behavior. *Journal of Human Resources*, *forthcoming*, 0718–9620R2.

Inchley, J., Currie, D., Young, T., Samdal, O., Torsheim, T., Augustson, L., ... Barnekow, V. (Eds.). (2013). *Growing Up Unequal: Gender and Socioeconomic Differences in Young People's Health and Well-Being*. WHO Regional Office for Europe: World Health Organization.

Kremer, M., Miguel, E., & Thornton, R. (2009). Incentives to Learn. *The Review of Economics and Statistics*, *91*(3), 437–456.

Lavy, V. (2010). Effects of Free Choice Among Public Schools. *Review of Economic Studies*, *77*(3), 1164–1191.

Leuven, E., Oosterbeek, H., & van der Klaauw, B. (2010). The Effect of Financial Rewards on Students' Achievement: Evidence from a Randomized Experiment. *Journal of the European Economic Association*, *8*(6), 1243–1265.

Levitt, S. D., List, J. A., Neckermann, S., & Sadoff, S. (2016). The Behavioralist Goes to School: Leveraging Behavioral Economics to Improve Educational Performance. *American Economic Journal: Economic Policy*, *8*(4), 183–219.

Lindbom, A. (2010). School Choice in Sweden; Effects on Student Performance, School Costs, and Segregation. *Scandinavian Journal of Educational Research*, *54*(6), 615–630.

MacKinnon, J. G., & Webb, M. D. (2020). Randomization Inference for Difference-in-Differences With Few Treated Clusters. *Journal of Econometrics*, *218*(2), 435–450.

Main, J. B., & Ost, B. (2014). The Impact of Letter Grades on Student Effort, Course Selection, and Major Choice: A Regression-Discontinuity Analysis. *The Journal of Economic Education*, *45*(1), 1–10.

Midgley, C., Anderman, E., & Hicks, L. (1995). Differences between Elementary and Middle School Teachers and Students: A Goal Theory Approach. *The Journal of Early Adolescence*, *15*(1), 90–113.

Muralidharan, K., & Prakash, N. (2017). Cycling to School: Increasing Secondary School Enrollment for Girls in India. *American Economic Journal: Applied Economics*, *9*(3), 321–350.

Napoli, A. R., & Raymond, L. A. (2004). How Reliable Are Our Assessment Data?: A Comparison of the Reliability of Data Produced in Graded and Un-Graded Conditions. *Research in Higher Education*, *45*(8), 921–929.

Norwegian Directorate for Education and Training. (2018). Trekkordning ved eksamen for grunnskole og videregående opplæring. [Online]. Accessed: 2018-09-07 Available from: https://www.udir.no/regelverk-og-tilsyn/finn-regelverk/etter-tema/eksamen/trekkordning-ved-eksamen-for-grunnskole-og-videregaende-opplaring-udir-2-2018/.

Norwegian Directorate of Education and Training. (2017). *The Education Mirror 2016*. Oslo, Norway: Utdanningsdirektoratet.

Oettinger, G. S. (2002). The Effect of Nonlinear Incentives on Performance: Evidence from "ECON 101". *The Review of Economics and Statistics*, *84*(3), 509–517.

Ruud, M. (2018). Skolepress og stress øker, særlig blant jenter. Retrieved from https://www.utdanningsnytt.no/helse-psykisk-helse-skolemiljo/skolepress-og-stress-oker-saerlig-blant-jenter/152221

Segal, C. (2012). Working When No One is Watching: Motivation, Test Scores, and Economic Success. *Management Science*, *58*(8), 1438–1457.

Statistics Norway. (2001). *Norwegian Standard Classification of Education*. Statistics Norway.

Stinebrickner, R., & Stinebrickner, T. (2008). The Causal Effect of Studying on Academic Performance. *The B.E. Journal of Economic Analysis & Policy*, *8*(1).

Sun, L., & Abraham, S. (2020). Estimating Dynamic Treatment Effects in Event Studies with Heterogeneous Treatment Effects. *Journal of Econometrics*, *forthcoming*.

The Education Act (Opplæringslova). (1998). *Lov om grunnskolen og den videregående opplæringa (LOV-1998-07-17-61)*.

Vroom, V. H. (1964). *Work and Motivation*. New York, NY: Wiley.

Wigfield, A., & Eccles, J. S. (2000). Expectancy-Value Theory of Achievment Motivation. *Contemporary Educational Psychology*, *25*, 68–81.

Wise, S. L., & DeMars, C. E. (2005). Low Examinee Effort in Low-Stakes Assessment: Problems and Potential Solutions. *Educational Assessment*, *10*(1), 1–17.

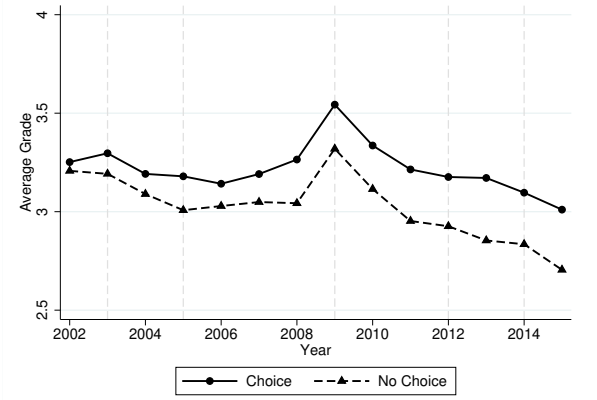
Wolf, L. F., & Smith, J. K. (1995). The Consequence of Consequence: Motivation, Anxiety, and Test Performance. *Applied Measurement in Education*, *8*(3), 227–242.

Zamarro, G., Hitt, C., & Mendez, I. (2019). When Students Don't Care: Reexamining International Differences in Achievement and Student Effort. *Journal of Human Capital*, *13*(4), 519–552.

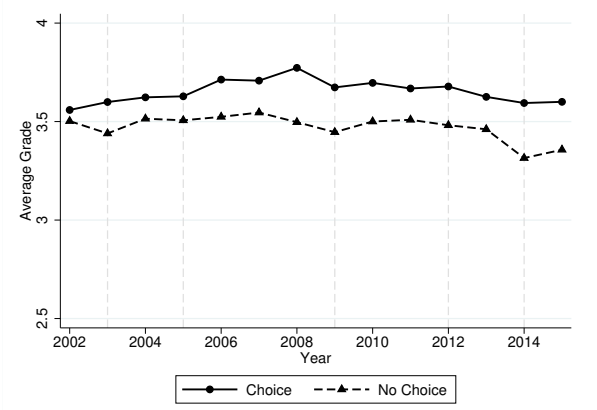
Appendix

Appendix A:

Descriptive Trends in Average Exam Grade by Year and Subject



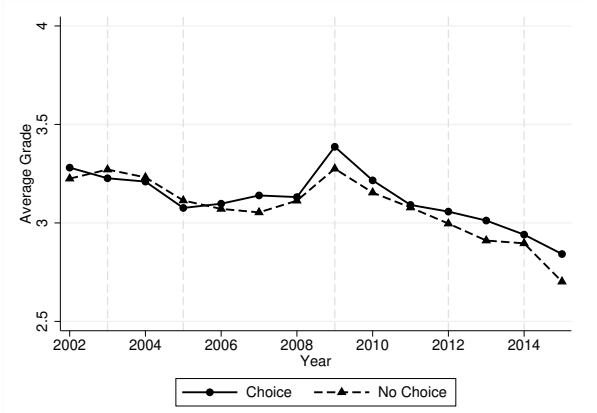
(a) Math



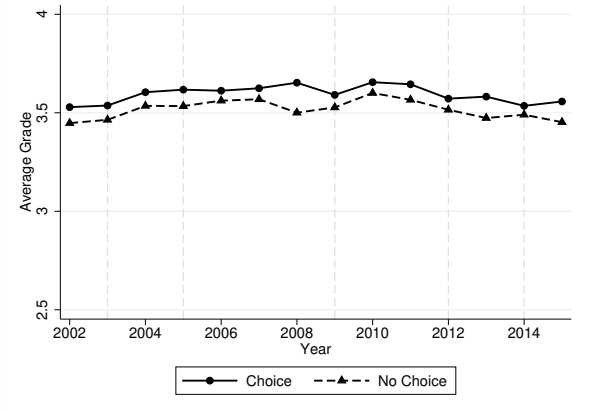
(b) Language

Figure A.1: Time Series of Average Exam Grade – Treatment

Note: The figure charts the average grade attained in the written final exit exam for students in the Choice and No-choice students respectively, by subject and treatment status. For reference, exam grades run on a scale of 1 (fail) to 6 (top grade). A circle (triangle) indicates the average for students with more (fewer) than two high schools within traveling distance in that particular year. Treatment group refers to the six counties that introduced school choice reforms during the sample period, indicated by the gray dashed reference lines.



(a) Math



(b) Language

Figure A.2: Time Series of Average Exam grade – Control

Note: The figure charts the average grade attained in the written final exit exam for students in the Choice and No-choice students respectively, by subject and treatment status. For reference, exam grades run on a scale of 1 (fail) to 6 (top grade). A circle (triangle) indicates the average for students with more (fewer) than two high schools within traveling distance in that particular year. Treatment group refers to the six counties that introduced school choice reforms during the sample period, indicated by the gray dashed reference lines.

Appendix B:

Results Robustness

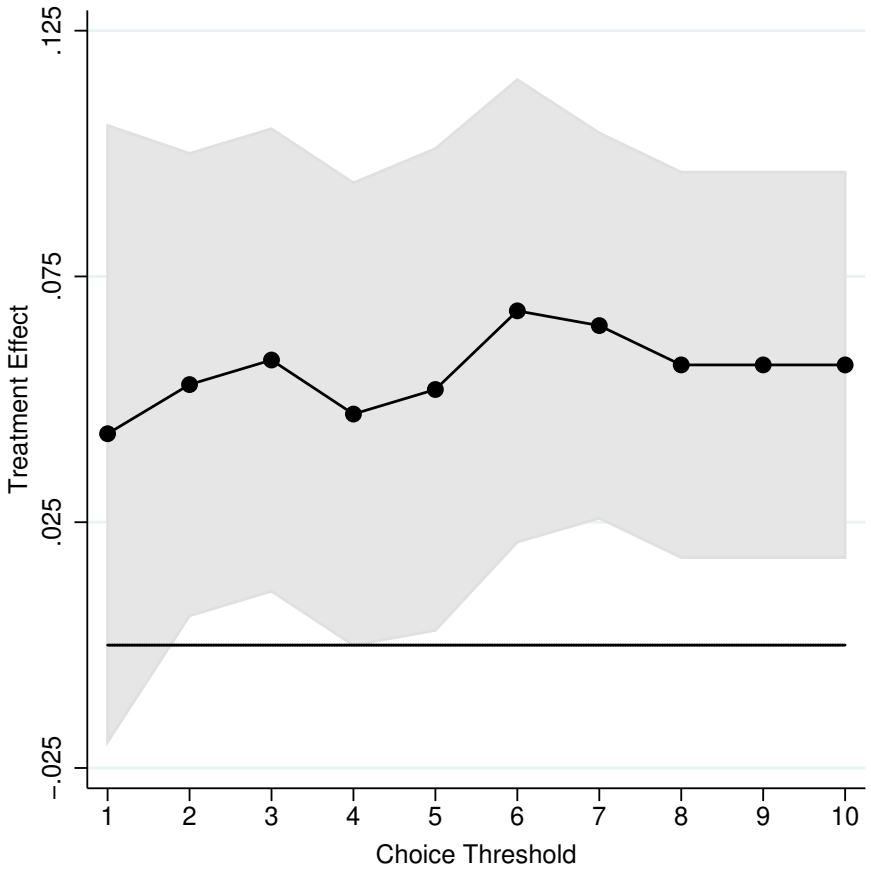


Figure B.1: Treatment Effect Estimates Across Choice Thresholds

Note: The figure shows results from estimation of my main triple difference model with various thresholds for what constitutes a real choice of schools. The numbers on the x-axis indicate the minimum number of schools required for the choice indicator to take the value 1. My preferred specification used throughout the paper, sets this threshold at 3. In the case where the threshold is set to 1, the triple difference model collapses to a conventional difference-in-difference model. The dots represent point estimates from separate regressions, with the shaded area indicating the 95% confidence interval of the estimates. The outcome variable is standardized to have a mean of 0 and a standard deviation of 1.

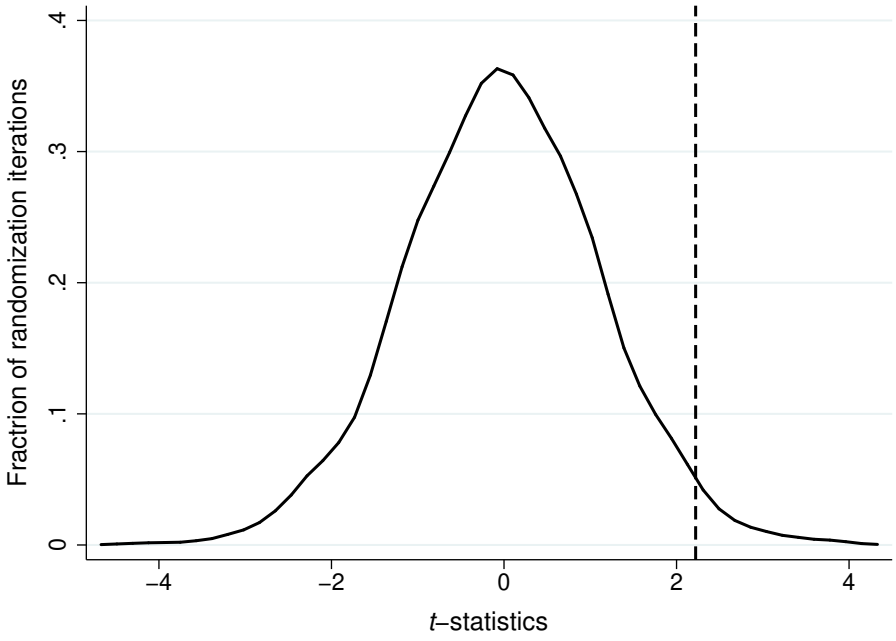


Figure B.2: Randomization-Based Inference

Note: The figure shows the results from conducting a randomization-based inference test, as prescribed by Athey and Imbens (2017) and Bind and Rubin (2020). The test simulates and approximates the likelihood of a treatment effect appearing by random chance due to the fact that a fixed number of units were assigned to treatment, and not a random sample of the population. Thus there might be unobserved differences that would make this set of units particularly likely to benefit from the treatment, or have other differences in baseline characteristics that would produce a false positive treatment effect estimate. Randomization inference randomly re-assigns treatment status, and re-estimates the treatment effects arising from these placebo assignments. By re-iterating this process multiple times the algorithm produces a distribution of placebo effects from assignments under which the null hypothesis should be true. By comparing the fraction of iterations in which the absolute value of the estimate exceeds the "true" estimate to the total number of iterations, randomization inferences produces an intuitive p-value. In the figure presented here I report results from a procedure where I re-assign treatment 5000 times. I follow MacKinnon and Webb (2020) and base the inference on the t-statistic rather than the β-coefficient as the treated units vary in size. The dashed line indicate the t-statistic obtained in the "true" model, while the solid line gives the distribution of placebo results. The procedure produced 260 placebo assignments in which the absolute value of the estimated t-statistic exceeded that of the true model. This corresponds with a p-value of $260/5000 = 0.052$.

Table B.1—Early vs Late Adopters Event Study Results

Relative time	Early adopters	Late adopters	Balanced panel
-4	0.010	-0.065**	
	(0.014)	(0.029)	
-3	-0.057**	0.030	-0.015
	(0.028)	(0.020)	(0.019)
-2	0.002	0.002	-0.003
	(0.019)	(0.025)	(0.016)
-1	Omitted	Omitted	Omitted
0	0.106***	0.003	0.068***
	(0.018)	(0.023)	(0.012)
1	0.082***	-0.065***	0.042**
	(0.017)	(0.023)	(0.015)
2	-0.017	0.028	0.016
	(0.019)	(0.032)	(0.013)
3	0.076***	-0.053	0.073***
	(0.017)	(0.034)	(0.014)
4	0.050**		
	(0.023)		
5	0.107***		
	(0.019)		
6	0.063***		
	(0.017)		
N	763,030	607,346	698,235
Adj. R^2	0.213	0.210	0.212

Note: The table presents results from a triple difference event study analysis where I split the treatment in various composition based on their relative timing of reform. In the first column run the analyzis using the 'early adopters', i.e. the first 3 counties in the sample period to implement school choice reforms as the treatment group. Similarly, in column two I run the analyses using only the latter 3 counties, the 'late adopters', in the treatment. In column 3 I consider a 'balanced' panel using the middle 4 reform cases for which I am able to construct a sample window where I observe the entire treatment group in all relative time periods (3 pretreatment and 4 posttreatment). In each of the analyses I exclude observations from the nonfocal treatment counties as including them in the control group potentially could bias the results. The event study decomposes the results over the years leading up to, and following the implementation of the reforms using both the conventional, and the Sun and Abraham (2020) interaction weighted event study approach. For these estimations treatment status is replaced with an indicator equal to one in that particular year only. The year prior to implementation is omitted for reference. Cluster-robust standard errors clustered at the commuting zone level in parentheses. * p<0.1, ** p<0.05, *** p<0.01.

Table B.2—Alternative Approaches to Standard Errors

	(1)	(2)	(3)	(4)	(5)	(6)
School choice reform × Choice	0.053**	0.053**	0.053**	0.053**	0.053*	0.053*
	(0.022)	(0.022)	(0.024)	(0.024)		
t-statistic	2.39	2.42	2.21	2.22	2.22	2.22
p-value	0.017	0.016	0.030	0.039	0.067	0.052
School choice reform	-0.010	-0.010	-0.010	-0.010		
	(0.019)	(0.015)	(0.014)	(0.016)		
Choice	0.013	0.013	0.013	0.013		
	(0.052)	(0.056)	(0.059)	(0.045)		
N	790,905	790,905	790,905	790,905	790,905	790,905
Adj.R^2	0.174	0.174	0.174	0.174	0.174	0.174
Cluster on:						
School	✓					
Municipality		✓				
Commuting Zone			✓			
County				✓	✓	✓
Few-clusters correction:						
Wild T Bootstrap					✓	
Randomization Inference						✓

Note: The table presents estimates of the average treatment effect, using the preferred DDD-specification from Table 3. Each column represent a separate estimation of this model with various approaches for computing the standard errors. In columns 1–4 I vary the level of clustering of the errors, starting with the school level and ending up at the county level. Beneath the point estimates and standard errors I also report the t-statistic and p-value of the main coefficient of interest, i.e. the three-way interaction School choice reform × Choice. In columns 5 and 6 I consider the fact that clustering at the county level implies that I only have 19 clusters, a number that is arguably too small to provide valid inference (Cameron and Miller, 2015). Therefore I perform two corrections to the few-clusters problem to assess the sensitivity of the results to this issue. In column 5 I perform a Wild-T cluster bootstrap, as suggested by Cameron and Miller (2015). In column 6 I perform a randomization inference procedure, which has been shown to yield appropriate rejection rates even with very few clusters (see e.g. Athey and Imbens (2017) and Bind and Rubin (2020) for a discussion of the method's merits, and MacKinnon and Webb (2020) for a practical illustration). For the latter, I keep the number of treated units within a given year fixed to the sample number in that year, but randomly assign treatment status to counties. This procedure is then repeated 5000 times to produce a distribution of placebo treatment results, which the estimated treatment effect of the true treatment assignment is then compared against. Evidently, the overall result of these exercises is that the main aggregate result of the study is robust to these standard error considerations. * p<0.1, ** p<0.05, *** p<0.01.

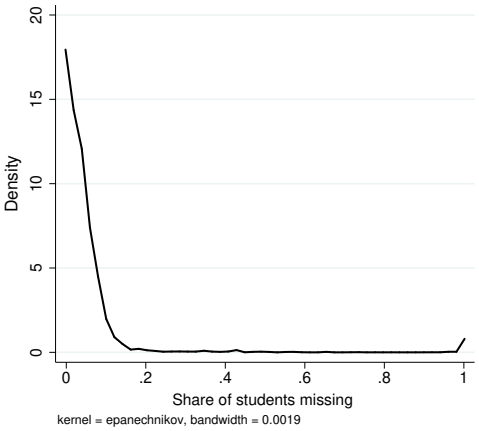
Table B.3—Aggregate Results Using GPA as Outcome

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
School choice reform × Choice	0.056**	0.052**	0.060**	0.055**	0.040*	0.055**	0.037**
	(0.024)	(0.025)	(0.027)	(0.025)	(0.026)	(0.023)	(0.014)
School choice reform	-0.009	-0.004	0.010	-0.013	0.002	-0.009	0.012
	(0.020)	(0.021)	(0.024)	(0.021)	(0.022)	(0.019)	(0.014)
Choice	0.005	0.012	0.003	0.168***	0.023	0.011	0.034
	(0.051)	(0.051)	(0.052)	(0.055)	(0.036)	(0.049)	(0.047)
N	851,857	851,857	851,857	567,755	661,252	789,519	851,857
Adj.R^2	0.275	0.277	0.277	0.280	0.273	0.276	0.276
Covariates	✓	✓	✓	✓	✓	✓	✓
Subject FE + trend		✓	✓				✓
Linear trend			✓				
IW DDD							✓
Excluding:							
Always treated				✓			
Never treated					✓		
Year = 2008						✓	

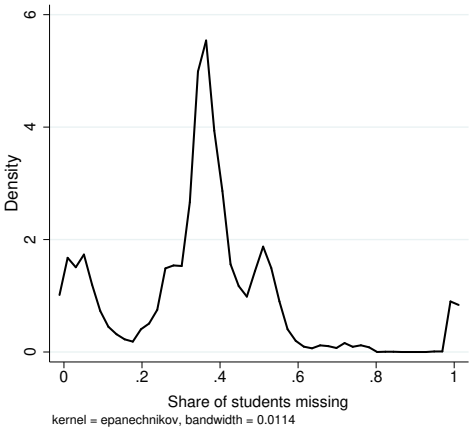
Note: The table presents estimates of the average treatment effect on GPA of imposing high-stakes grades through merit-based school choice admission schemes. The GPA is calculated using all nonexam grades, i.e. those set by the student's teachers, and is standardized to have a mean of 0 and a standard deviation of 1. Results stem from estimating the DDD model specified in (5). The coefficient of interest is three-way interaction School choice reform × Choice in the top row which gives the average treatment effect of being a student graduating from a treated county, in a labor market region with more than two high schools, after the treatment has been implemented. Conversely, the School choice reform variable controls for the conventional two-way fixed effects difference-in-differences estimator of graduating from a treated county in a posttreatment year. Choice is a dummy equal to one for students who have more than two high schools within traveling distance from their home. The triple difference model in practice interacts the DID-estimator with this dummy. The models in Column 4 and 5 exclude all observations from always-treated and never-treated counties respectively. In Column 6 I exclude all observations from the year 2008 from the regression. In Column 7 I report an aggregation of unit-time specific effects derived from an IW DDD model using the approach suggested by Callaway and Sant'Anna (2020). Cluster-robust standard errors clustered at the commuting zone level in parenthesis. * p<0.1, ** p<0.05, *** p<0.01.

Appendix C:

Missingness



(a) All Years Except 2008



(b) 2008 only

Figure C.1: Share of Students Missing

Note: The figure reports kernel density plots for the fraction of students whose exam grade is missing in a given school in a given year, calculated using all students for whom the school ID is observed (4,682 missing values are excluded). Panel (a) shows the kernel density for all years, with the exception of 2008 which is excluded. The distribution for 2008 is shown by itself in Panel (b).

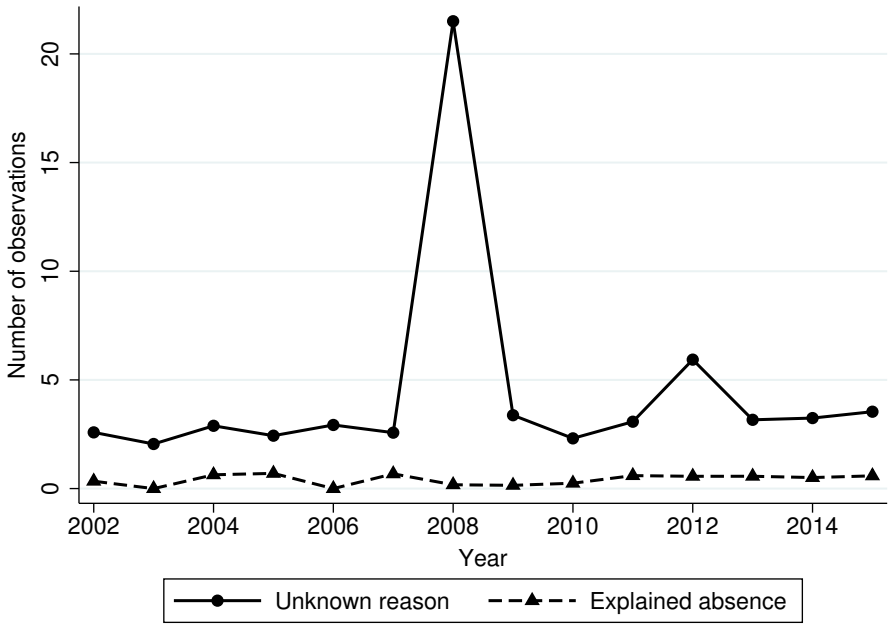


Figure C.2: Missing Values by Year

Note: The figure charts the number of students with an exam grade missing for each cohort. The dashed line counts students whose absence is explained in the registry (exempted, sick or no-shows) while the solid line counts unexplained missing values. For the former group exam subject is known. Y-axis values are in thousands.

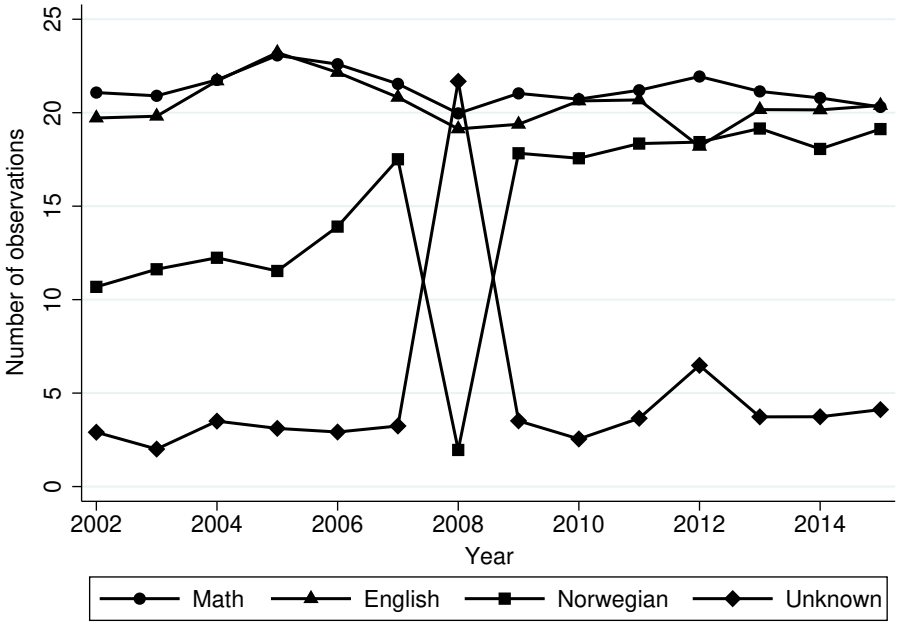


Figure C.3: Missing Values by Year

Note: This figure charts the number of students who were tested in a particular subject, for each cohort. Y-axis values are in thousands. Those for whom information on subject is missing are listed as 'Unknown'.

Table C.1—Predicting Missing Values

	All missing	Sick	Exempt	No-show
	(1)	(2)	(3)	(4)
School choice reform × Choice	0.008	-0.001	0.001	0.001
	(0.008)	(0.000)	(0.001)	(0.001)
School choice reform	-0.004	0.001	-0.000	-0.001
	(0.005)	(0.000)	(0.001)	(0.001)
Choice	0.029	0.000	-0.001	0.001
	(0.014)	(0.001)	(0.003)	(0.001)
N	854,539	854,539	854,539	854,539
Adj.R^2	0.141	0.006	0.009	0.005

Note: This table lists results from estimating identical triple difference models to those in the main analysis, but using various categories of missing values as the outcome variable. Column 1 pools all categories (and those without explicit reason for 'missingness') and regresses a dummy equal to one if the exam grade is missing for the preferred specification as described in Section 3.2. Columns 2–4 repeats the estimation for each known reason for absence separately. Cluster-robust standard errors clustered at the commuting zone level in parentheses. * p<0.1, ** p<0.05, *** p<0.01.

Appendix D:

Full Event Study Results

Table D.1—Main Event Study Results

Relative time	DDD estimates $\hat{\mu}_l$	IW DDD estimates $\hat{v}_l$	Cohort-specific ATT estimates after l periods $\hat{\delta}_{1,l}$	$\hat{\delta}_{2,l}$	$\hat{\delta}_{3,l}$	$\hat{\delta}_{4,l}$	$\hat{\delta}_{5,l}$	$\hat{\delta}_{6,l}$
-4	-0.022	-0.007	.	0.015	-0.240	0.107	.	-0.037
	(0.039)	(0.015)	.	(0.019)	(0.070)	(0.034)	.	(0.036)
-3	0.039*	-0.020	.	0.037	-0.001	0.009	-0.129	0.056
	(0.020)	(0.018)	.	(0.022)	(0.041)	(0.045)	(0.047)	(0.025)
-2	0.013	0.002	.	-0.000	-0.104	-0.025	0.017	0.092
	(0.026)	(0.015)	.	(0.027)	(0.031)	(0.078)	(0.030)	(0.028)
-1	Omitted	Omitted	.	.	.	.	.	.
0	0.065**	0.071***	0.177	0.080	-0.059	-0.014	0.058	0.058
	(0.032)	(0.015)	(0.024)	(0.025)	(0.027)	(0.045)	(0.040)	(0.040)
1	0.046	0.030*	0.100	0.103	-0.075	-0.112	0.037	-0.025
	(0.036)	(0.015)	(0.028)	(0.032)	(0.036)	(0.048)	(0.035)	(0.035)
2	0.056	0.015	0.028	0.052	-0.001	0.057	-0.110	.
	(0.035)	(0.012)	(0.044)	(0.027)	(0.035)	(0.044)	(0.020)	.
3	0.034	0.041**	0.099	0.071	-0.107	0.010	0.041	.
	(0.036)	(0.016)	(0.020)	(0.026)	(0.046)	(0.036)	(0.039)	.
4	0.071*	0.052**	0.144	0.113	.	.	-0.095	.
	(0.037)	(0.023)	(0.036)	(0.028)	.	.	(0.035)	.
5	0.092	0.101***	0.097	0.102	.	.	0.105	.
	(0.057)	(0.019)	(0.030)	(0.029)	.	.	(0.035)	.
6	0.070*	0.065***	0.127	0.130	.	.	-0.050	.
	(0.037)	(0.019)	(0.023)	(0.035)	.	.	(0.022)	.
N	790,905	790,905						
Adj. R^2	0.214	0.215						

Note: Estimation of the timing of treatment effects using the both the conventional event study design, and the Sun and Abraham (2020) IW cohort design. For these estimations treatment status is replaced with an indicator equal to one in that particular year only, except $l = -4$ and $l = 6$ which is one for all preceding/subsequent years. The year prior to implementation is omitted for reference. $\hat{\delta}_{1,l}$ - $\hat{\delta}_{6,l}$ gives the results for the cohort-specific treatment effect at each relative time point, if the cohort is treated at that point. The IW DDD estimate is the weighted average of the observed CATTs at any given l. Errors clustered at the commuting zone level in parentheses. * p<0.1, ** p<0.05, *** p<0.01.

Table D.2—National Assessment Test Event Study Results (9th Grade)

Relative time	DDD estimates $\hat{\mu}_l$	IW DDD estimates $\hat{v}_l$	CATT estimate after l periods $\hat{\delta}_{1,l}$	$\hat{\delta}_{2,l}$	$\hat{\delta}_{3,l}$
-2	0.023	0.039	0.157	-0.078	0.014
	(0.033)	(0.033)	(0.085)	(0.031)	(0.017)
-1	Omitted	Omitted			
0	-0.011	0.010	0.075	-0.073	0.011
	(0.027)	(0.023)	(0.047)	(0.040)	(0.017)
1	0.030	0.053***	0.091	0.026	0.035
	(0.024)	(0.019)	(0.040)	(0.018)	(0.021)
2	0.053	0.071*	0.113	0.018	.
	(0.035)	(0.042)	(0.064)	(0.044)	
3	0.055	0.070**	0.178	-0.066	.
	(0.038)	(0.027)	(0.041)	(0.019)	
Aggregate result	0.037**				
	(0.018)				
N	249,602	249,602			
Adj. R^2	0.767	0.753			

Note: Estimation of the timing of treatment effects using the both the conventional event study design, and the Sun and Abraham (2020) IW cohort design. For these estimations treatment status is replaced with an indicator equal to one in that particular year only, except $l = -4$ and $l = 6$ which is one for all preceding/subsequent years. The year prior to implementation is omitted for reference. $\hat{\delta}_{1,l}$ - $\hat{\delta}_{6,l}$ gives the results for the cohort-specific treatment effect at each relative time point, if the cohort is treated at that point. The IW DDD estimate is the weighted average of the observed CATTs at any given l. Errors clustered at the commuting zone level in parentheses. * p<0.1, ** p<0.05, *** p<0.01.

Table D.3—National Assessment Test Event Study Results (8th Grade)

	DDD estimates	IW DDD estimates	CATT estimate after l periods		
Relative time	$\hat{\mu}_l$	$\hat{v}_l$	$\hat{\delta}_{1,l}$	$\hat{\delta}_{2,l}$	$\hat{\delta}_{3,l}$
-2	-0.009	0.024	0.016	0.068	0.001
	(0.034)	(0.029)	(0.056)	(0.059)	(0.030)
-1	Omitted	Omitted			
0	-0.007	0.004	0.007	0.061	-0.040
	(0.024)	(0.024)	(0.032)	(0.058)	(0.030)
1	-0.084**	-0.058*	-0.092	0.068	-0.116
	(0.037)	(0.032)	(0.053)	(0.070)	(0.036)
2	0.050	0.084	0.046	0.138	.
	(0.053)	(0.057)	(0.087)	(0.053)	.
3	-0.101***	-0.069**	-0.088	-0.045	.
	(0.029)	(0.034)	(0.040)	(0.048)	.
Aggregate result	-0.036				
	(0.029)				
N	247,421	247,421			
Adj. R^2	0.174	0.174			

Note: Estimation of the timing of treatment effects using the both the conventional event study design, and the Sun and Abraham (2020) IW cohort design. For these estimations treatment status is replaced with an indicator equal to one in that particular year only, except $l = -4$ and $l = 6$ which is one for all preceding/subsequent years. The year prior to implementation is omitted for reference. $\hat{\delta}_{1,l}$ - $\hat{\delta}_{6,l}$ gives the results for the cohort-specific treatment effect at each relative time point, if the cohort is treated at that point. The IW DDD estimate is the weighted average of the observed CATTs at any given l. Errors clustered at the commuting zone level in parentheses. * p<0.1, ** p<0.05, *** p<0.01.

Chapter 3 – Essay II

Reducing the Gender Gap in Early Learning: Evidence From a Field Experiment in Norwegian Preschools

Andreas Fidjeland[1]*, Mari Rege[1], Ingeborg F. Solli[1], and Ingunn Størksen[2]

Abstract

Early-childhood programs attract considerable policy interest as a tool to prepare children for formal schooling. However, we have limited knowledge about whether conditions for early learning are similar for all children. Program attendance seems to have less effect on boys, but so far there is little evidence suggesting reasons for why this might be the case. In this field experiment, we investigate whether a more structured curriculum can reduce the gender gap in early learning. While girls have higher skills at baseline, we find that the intervention primarily benefits boys, with effects persisting into formal schooling.

JEL Codes: I20, H42

Keywords: Gender gap, field experiment, school readiness, child development

*Corresponding author. Email: andreas.fidjeland@uis.no. We are thankful for helpful comments and suggestions from Hans H. Sievertsen, Henning Hermes, John Eric Humphries, Maximiliaan Thijssen, May Linn Auestad, and from seminar participants at the University of Stavanger, as well as conference participants at the AEA meetings, the AEFP conference, the Lund PhD Workshop in the Economics of Education, and the annual meeting of the Norwegian Association for Economists. Funding from the Norwegian Research Council, grant numbers 237973 and 270703, is gratefully acknowledged.

[1] University of Stavanger Business School, Department of Economics and Finance
[2] University of Stavanger, Norwegian Center for Learning Environment and Behavioral Research in Education

1 Introduction

An extensive literature documents persistent gender gaps in academic achievement, across a variety of outcomes and educational contexts (e.g., Autor and Wasserman, 2013; Bedard and Cho, 2010; DiPrete and Buchmann, 2013; OECD, 2015). Boys are not only outperformed by girls in test scores, but also disproportionately represented in negative statistics such as school dropout, behavioral problems, and special needs, the effects of which spill over into adulthood with links to college enrollment, unemployment, and even crime (Fortin et al., 2015; Goldin et al., 2006; Vincent-Lancrin, 2008).[1] Still, the origins of these gaps are not fully understood. Many hypotheses have been proposed as to why they might emerge once children have started school, but there is also evidence of substantial differences across gender in terms of skills crucial for learning even before children start formal schooling. (Brandlistuen et al., 2020; Magnuson and Duncan, 2016).

Such early gender differences have relevance for policy debates about the provision of early childhood education and care (ECEC). In earlier decades, the perception that boys are more immature than girls has caused an increasing prevalence of delayed school entry for boys, but there is little evidence to suggest that such "academic redshirting" has long-term positive effects on child development and educational attainment (Deming and Dynarski, 2008). In contrast, ECEC programs have attracted increasing interest from policymakers as research has demonstrated how participation in such programs can improve school readiness, with potential long-term gains in academic and labor market outcomes (Berlinski

[1] Even though, over the past decades, women have surpassed men by substantial (and increasing) margins in terms of educational attainment, it is still the case that women, on average, earn less than men and are disproportionately less likely to hold powerful positions in society (DiPrete and Buchmann, 2013). However, while women might still face barriers to capitalizing on their education in the labor market, particularly older cohorts of women now in the latter half of their labor-market career, the earnings gap has halved over the past 40 years despite increasing wage inequality overall (Blau and Kahn, 2007).

et al., 2008; Cornelissen et al., 2018; Felfe and Lalive, 2018; Felfe et al., 2015; Heckman et al., 2010; Melhuish, 2011). However, the variety of program contexts, designs, and features makes the literature far from unified with respect to the conditions necessary for child development (Phillips et al., 2017; White et al., 2015). Even less is known about the distribution of potential effects, and about whether the conditions for realizing these benefits are similar for all children (Duncan and Magnuson, 2013; Phillips et al., 2017). The evidence of gender-specific returns to ECEC enrollment is mixed (Magnuson et al., 2016), but several studies report results indicating that girls might benefit more in terms of skill development than boys (e.g., Anderson, 2008; Cornelissen et al., 2018; Felfe et al., 2015; Fessler and Schneebaum, 2019; Goodman and Sianesi, 2005; Havnes and Mogstad, 2015). However, so far the literature provides little evidence — or even discussion — when it comes to why this might be the case.

One potential explanation is that girls and boys seemingly spend their time in childcare very differently, particularly in unstructured settings (Early et al., 2010; Tonyan and Howes, 2003). When given the opportunity, girls are much more likely to engage in activities that promote school readiness and skill development (Stangeland et al., 2018; Størksen et al., 2015). In contrast, boys engage more in spontaneous and physical behavior, shifting attention between activities rapidly and interacting with adults to a lesser extent. This suggests that boys may be less exposed to many of the stimulating learning activities that girls seem inclined to engage in of their own accord.

These observational insights suggest that a more structured curriculum could decrease gender gaps in early learning. We investigate this hypothesis using data from a randomized controlled trial (RCT) carried out in the context of the universal preschool system in Norway (see Rege et al., 2021, for aggregate treatment results).[2] The intervention introduced a more intentional practice through a structured, comprehensive curricu-

[2]The project was preregistered in the AEA registry (code AEARCTR-0002241), where gender heterogeneity was specified as part of our analysis plan.

lum for groups of five-year-olds in their final year of preschool, with the goal of improving their school readiness. This practice contrasts with the prevalent Norwegian ECEC pedagogical philosophy, which largely centers on child-initiated free play.

To that end we recruited 71 ECEC centers, where the teachers in the treatment group were provided with a curriculum encompassing age-appropriate intentional skill-building activities to be implemented for all five-year-olds. This curriculum was coupled with a thorough professional-development program as well as further support throughout the intervention. The activities were embedded within a playful learning approach and targeted key school-readiness skills in the areas of mathematics, language, and executive functioning.

We rely on data collected through detailed one-to-one assessments by certified testers blind to treatment status to investigate the effects of the intervention. We find that there is a substantial gender gap in school readiness at baseline, and that this gap is not mitigated by business-as-usual pedagogical practice. Moreover, consistent with our hypothesis, we find that the average improvement in school readiness brought about by the intervention — as was first reported in Rege et al. (2021) — is almost entirely driven by a treatment effect of about 20 percent of a standard deviation on the boys. In contrast, we find little evidence that the new curriculum had any effect on the girls. This is true both for the posttreatment assessment and for a one-year follow-up at the end of first grade. The positive effects seen for boys persist across the transition to formal schooling. We also find suggestive evidence that the boys at the bottom of the skill distribution at baseline are the ones who improve the most. In a heterogeneity analysis we estimate decreasing treatment effects on boys as we move up the rank distribution of baseline scores, suggesting that the efficacy of the intervention decreases with the initial skill level. For girls, we find no such relationship at all.

Our results have important policy implications. With many countries experiencing a push toward universal provision of preschool programs,

there is a need for robust evidence on how curricular design and pedagogical practice might have heterogeneous impacts on child subgroups. To date, much of the curriculum in universal preschool programs is fairly non-specific and unstructured. This is particularly the case in many Nordic and Central-European countries, where childcare pedagogy emphasizes the value of free play, autonomy and spontaneous engagement between teacher and child (Engel et al., 2015; White et al., 2015). This holistic approach to child development offers individual ECEC centers substantial discretion with regard to pedagogical content. However, the unstructured nature of learning activities could also give rise to heterogeneity in school readiness. Hence, implementing curricula such as that featured in our intervention could potentially reduce gender gaps in early learning by having a positive impact on the development of boys, while not being detrimental to girls. Furthermore, the persistence of the treatment effects as the children transition to formal schooling suggests that these curricula have the potential to help reduce gender gaps in later academic achievement as well.

We also contribute to the literature on several fronts. First, the paper highlights the need for a better understanding of what constitutes *process* quality in early childhood education (Phillips et al., 2017) — as opposed to structural quality, which has been the focus of much economic research. We present evidence underscoring the importance of curriculum design and intentional practice as a channel for promoting child development. Second, our paper also addresses the knowledge gap on heterogeneous impacts of childcare participation (Duncan and Magnuson, 2013; Phillips et al., 2017). We provide causal evidence for a plausible channel through which boys and girls could be affected differently. Thus our study also contributes to the literature investigating the origins of gender gaps in educational outcomes. Our high-quality assessment data allow a precise characterization of the scope of early gender gaps, while also allowing us to follow the development of those gaps in the crucial transition from early childhood care to formal education.

2 The Scope and Origins of Gender Gaps in Early Learning

The extant literature on gender gaps in early learning suggests that girls start formal schooling with a significant advantage in school-readiness skills (e.g., Brandlistuen et al., 2020; DiPrete and Jennings, 2012; Magnuson and Duncan, 2016). Although gender gaps at school start are not necessarily caused by the preschool programs the children may have attended, several studies report results suggesting that such programs might be particularly beneficial for the development of girls, although gender differences in program effectiveness are rarely an explicit focus in those studies (Magnuson et al., 2016). Anderson (2008) provides a prominent example by revisiting the Perry Preschool, Abecedarian, and Early Training projects and finding that those early model programs benefited girls over boys by a difference of about 40 percent of a standard deviation. While differential effects of such a magnitude have rarely been replicated elsewhere in the literature, a meta-analysis of 23 early preschool programs revealed that girls benefited significantly more than boys in terms of cognitive, achievement, behavioral, and mental-health measures, although the differences were small (Magnuson et al., 2016). In contrast, boys benefited much more than girls in terms of other school outcomes such as detention and need for special education. There is also some evidence that programs affect boys more in the long term (Domond et al., 2020; Gray-Lobe et al., 2021).

There are at least three plausible mechanisms through which preschool programs might affect boys and girls differently. First, if girls enter preschool age with better-developed pre-academic skills — and skills beget skills (Cunha and Heckman, 2007) — we would expect developmental trajectories during preschool to be different for boys and girls. However, not only do boys and girls typically enter preschool at a roughly equal level of development (Magnuson et al., 2016), but there is also seemingly no gender difference in the aptitude for developing early language and mathematics skills (Spelke, 2005).

Second, preschool teachers might be inclined to foster a learning en-

vironment better suited for girls' development. For example, the lack of male role models in ECEC might be particularly detrimental for boys (Sumsion, 2005), and there is indeed some evidence that increasing the share of male teachers would be beneficial for their development (Drange and Rønning, 2020; Gørtz et al., 2018). Other studies have indicated that teachers have different expectations for the behavior of boys and girls, where the latter are to a larger extent expected to behave in self-regulated manners, such as by sitting still, waiting for their turn, and playing quietly (Lenes et al., 2020).

Third, girls and boys might spend their time in preschool differently, particularly in unstructured settings. Broadly, girls have in fact been found to be more likely to spend their time in cognitively stimulating activities during free-play time, while boys are more likely to engage more in spontaneous and physical behavior (Early et al., 2010; Tonyan and Howes, 2003). Further, girls are more likely to interact with adults, hence developing higher-quality relationships with teachers. In turn, the degree of teacher–child interaction has been found to be a consistent indicator of classroom quality as well as a predictor of child development in preschool settings (Mashburn et al., 2008). Indeed, Howes et al. (2008) find that effective preschool classrooms are characterized by intentionality in the way teachers engage and interact with the children, and that opportunities to learn are more plentiful in classrooms where teachers manage time more actively.

In this context, there is a strand of research that highlights the potential benefits of more structured curricula for child development (Clements and Sarama, 2011; Diamond et al., 2007; Dillon et al., 2017; Schmitt et al., 2015; Weiland and Yoshikawa, 2013). These studies argue that preschool staff can target school-readiness competences through a more intentional and systematic approach to learning situations. This could be particularly important for boys, who might need more support and scaffolding from teachers to engage in stimulating activities (Størksen et al., 2015). In this paper, we therefore ask whether providing staff with more structured learning activities to carry out with *all* the children improves boys' school readiness, so that they start formal schooling on a more equal footing with girls.

3 Institutional Background

Norway invests heavily in ECEC by subsidizing all preschool centers that adhere to governmental regulations. Parental payments are capped at approximately USD 300 per month per child, which reflects about 15 percent of the total cost of childcare enrollment (Norwegian Directorate for Education and Training, 2019). Price reductions are given for siblings, and free enrollment is offered to households with incomes below certain thresholds. As a result, Norway is among the OECD countries with the highest public spending on ECEC. Children are typically first enrolled in ECEC between ages 1 and 2, and all children are guaranteed enrollment in a center in their municipality. These heavy subsidies have led to a near-universal take-up. As of 2020, 92.2 percent of all children aged 1–5 years were enrolled in formal childcare, and 97.5 percent of all five-year-olds were.[3] Once enrolled, most children remain in ECEC until they start compulsory schooling in August of the calendar year in which they turn six.

The Norwegian center-based ECEC is founded on a social pedagogical tradition emphasizing free play and child-initiated activities in preschool child groups (Engel et al., 2015). While play is seen as an activity that may facilitate learning, a cornerstone of the Norwegian philosophy is that play also has an intrinsic value and is a goal in and of itself. There is no set curriculum to guide the provision of ECEC. Rather, individual centers are given substantial discretion in how to structure daily activities for the children in order to meet the goals of a framework plan that loosely outlines the purpose, values, and learning areas of Norwegian ECEC (OECD, 2015).[4] Against this backdrop, current pedagogical practice emphasizes unstructured and spontaneous play to such an extent that it is often given priority over adult-driven activities even when those are preplanned and scheduled (Synodi, 2010). Indeed, Karlsen and Lekhal (2019) find in their case studies that 60 percent of the ECEC day consisted of free-play activ-

[3] Aggregate data on participation is available at https://www.ssb.no/en/utdanning/barnehager/statistikk/barnehager

[4] An English-language version is available at https://www.udir.no/globalassets/filer/barnehage/rammeplan/framework-plan-for-kindergartens2-2017.pdf

ities and that center staff spent almost half of that time away from play situations, indicating that children spend a significant portion of their time at the ECEC center without interacting with adults.

This emphasis on free and autonomous play contrasts with the Anglo-American school-readiness tradition found in the United States and the United Kingdom, where preparing children for formal schooling is a more explicit pedagogical objective. The structured yet playful curriculum intervention investigated here should be seen as a step toward a more intentional school-readiness perspective in Norway. While the intervention does not abandon the tenets of the social-pedagogical tradition, it does incorporate some of the intentional and structured practices of the school-readiness philosophy.

4 Experimental Design and Measures

4.1 Experimental Design

This paper investigates gender-specific effects on school readiness in a randomized controlled trial conducted in Norwegian preschools. We recruited participants from two Norwegian counties. First we invited all 30 municipalities in those counties to sign up for the project, of which 15 did. Then we invited all publicly regulated childcare centers operating in those 15 municipalities to participate. Out of 190 centers, 72 signed up. One center in the control group later withdrew from the project, leaving us with a sample of 71 participating centers.

For the randomization procedure, we split the centers in 15 blocks, matched for size and geographic location. The resulting blocks consisted of 4 to 6 centers, with the total number of children ranging from 29 to 92. Parental consent was collected prior to randomization, but we accepted late consenters because of the lengthy time between initial collection and the start of the intervention. Of 701 parental consents collected (92% consent rate), 18.8 percent were submitted after our initial deadline. The late consenters were skewed toward the treatment group. This could be

because the teachers, who were in charge of collecting the consent sheets, might have been more invested in the project and worked harder to get parents to sign up once they became aware that they would be in the treatment group. For this reason, we include an indicator for late consent in all our estimations. However, excluding all late consenters from the sample yields broadly similar results (see Table D.5 in the appendix).

As the curriculum was developed by the project team and hence is not covered by existing preschool-teacher training, the project period started with teachers in the treatment group receiving training in the form of a 15 ECTS (i.e., roughly corresponding to half a semester full-time) university course on the pedagogics and practices of *playful learning*.[5] This training also allowed us to obtain extensive feedback from the practitioners on the curriculum and to revise it accordingly. The teachers subsequently implemented the curriculum starting in the fall of 2016. We conducted our *baseline* assessment immediately prior to implementation (T1). The treatment group then proceeded with the intervention for nine months, before we conducted the *postintervention* assessment in late spring 2017 (T2). We reconnected with the children one year later for a *follow-up* asessment once they were nearing the end of first grade (T3).

Throughout the intervention period, the control group followed a business-as-usual condition, implementing the curricular content that they would normally implement. However, the teachers in the control centers were informed that they would receive the training, the curriculum documents, and any accompanying materials when the participating children left for primary school. This was made clear early in the process, in an effort to mitigate any discouragement effects in the control group that might bias the results.[6]

[5]Størksen et al. (2021) detail the theories, concepts, and processes underpinning the practical implementation of the project.

[6]Our project did not involve the parents, and it was up to the centers to keep them informed about daily activities. As we did not survey the parents, we cannot speak to how aware they were of the project or whether they may have responded in a compensatory manner, for example upon learning that their child was in the control group. However, because the activities in the new curriculum did not differ radically from existing practices and represented only a small part of the regular schedule, we believe that such adjustments from parents are unlikely. It should also be noted that there are very limited short-term opportunities to move children between centers, for

4.2 Intervention Content

Our intervention is a bundle of several components. The main feature is the preschool curriculum, which consists of structured and intentional skill-building activities centered around *playful learning*. The *playful* aspect requires activities to be interactive and engaging at a level appropriate for the age group to master (Weisberg et al., 2013). Concretely, the curriculum consisted of a booklet with 130 activities, most of which would already be familiar to many Norwegian preschools.[7] However, they differed in both design and content, as well as in the level of intentionality with which they were to be implemented in order to stimulate school-readiness skills. Examples of activities include puzzles and games to cultivate number and quantitative thinking, dialogical reading to stimulate language, and stories and images where the children had to identify emotions.[8]

We encouraged teachers to develop their own approach to the curriculum. The activities were flexible, allowing teachers to adapt their difficulty and complexity to best fit the needs of their child group. The only requirement was for teachers to commit to spending at least 8 hours a week doing activities with their five-year-olds, Given that nearly all Norwegian children of this age spend 30–40 hours a week in childcare, our intervention took up only a modest proportion of the preschool schedule. Even so, it represented a substantial increase in the time devoted specifically to the stimulation of school-readiness skills, which, according teacher reports, ranged from 0 to 3 hours a week prior to the intervention.

To improve implementation quality, we assisted the teachers with supervision and guidance throughout the intervention. Three times per semester, a team member would call the teachers to answer questions or discuss challenges they faced. The teachers were also required to answer a weekly questionnaire in which they reported on implementation fidelity.

example to enrol them in one of the treatment centers, as centers do not typically have any available slots in the middle of the academic year.

[7]The booklet was also published as a book after the end of the project, and it is now widely available to practitioners and researchers (Størksen et al., 2018).

[8]See appendix material in Rege et al. (2021) for a more comprehensive description of the curricular content.

Because child groups are typically mixed-age, the five-year-olds had to be separated from the rest of their group for the activities. The teachers also needed time to prepare the week's activities. To cover the centers' costs associated with hiring additional and temporary staff we provided them with lump-sum transfers equivalent to the cost of a part-time (50 percent) position for nine months. Similarly, we covered the costs of a 50-percent position for four months as the teachers participated in the training course.

4.3 Measures

We center our assessment around the construct of *school readiness* (Bennett and Tayler, 2006). Conceptually, this encapsulates (both cognitive and socioemotional) skills that support a successful transition to formal schooling, as well as constituting the foundation on which later learning is achieved. The assessment consisted of age-appropriate, validated tests for measuring early skills. All tests were conducted in one-to-one sessions using a tablet computer, both for the child to interact directly with and for the tester to record answers and scores on.

We used the following measures to assess the children's skills:
Numeracy——To measure early mathematical skills, we used the Ani Banani Math Test (ABMT), which assesses the understanding of numeracy, geometry, and problem solving using a playful tablet application. The children help a monkey with different tasks, such as counting bananas and setting the table with enough plates for their birthday-party guests. ten Braak and Størksen (2021) assess the psychometric properties of the ABMT and find strong predictability of later mathematical achievement as well as discriminant validity against related constructs. They do note signs of gender bias in three items, but this was not consistent across samples, and nor was gender predictive of the latent construct. Still, for robustness we run our analyses both with and without those three items, finding that this does not affect results (see Table D.6 for details).

Language——To assess vocabulary, we use a short version of the Norwegian Vocabulary Test (NVT; Størksen et al., 2013). The children were presented with images on the tablet and asked to name the object de-

picted. To assess phonological awareness, we used a 12-item blending task that is part of the official literacy screening battery of the Norwegian Directorate for Education and Training. A word is presented in by phonemes (language sounds), and the child has to select the one out of four options on the tablet that corresponds to that word.

Executive functioning—To assess executive functioning, we used three tests: Wechsler's Digit Span Test (Wechsler, 2003) for measuring working memory; the Head-Toes-Knees-Shoulders task (McClelland et al., 2014) to assess behavioral self-regulation; and the Hearts and Flowers task (Davidson et al., 2006), which is widely used to measure cognitive flexibility in young children.

From each of these six tests we created three outcome measures by standardizing scores within each wave (T1–T3) to a mean of 0 and a standard deviation of 1. We then averaged across tests within each domain (numeracy, language, and executive functioning) and re-standardized the resulting index. We also averaged across indices and re-standardized to construct a sum-score measure.

We conducted the T1 and T2 assessments at local science museums. Participating centers were invited to spend a day at the museum, and we assigned a time slot for assessment. At that point, the children were lined up at the assessment station by the center staff. Children standing in line were continuously assigned to the next available tester, meaning that child-to-tester matching was naturally randomized. Testing took place over several days, and centers from both the treatment and control groups were invited each day. The testers were blind to treatment status, and they had been trained and certified prior to data collection. At T3, the children had moved on to primary school and hence were spread across multiple sites. For this wave, testers traveled to schools, where staff let children leave the classroom to be assessed.

5 Data and Empirical Strategy

5.1 Sample

Out of the 701 children for whom we collected parental consent, 658 participated in the T1 baseline assessment while 650 participated in the T2 postintervention assessment. For the T3 follow-up we were able to locate and assess 661 children. Although we did not explicitly balance the sample on gender, shares of boys and girls were equal in each wave, with the difference in absolute numbers ranging from 2 at T1 to 9 at T3.

We construct our analytical samples from the children observed in the T2 and T3 waves, respectively, and run our analysis on these samples separately. Although this means that the T2 and T3 samples will be slightly different, there is a substantial overlap as 620 children participated in both waves. For those missing at baseline, we impute test scores using predicted values based on child and parent characteristics (gender, birth month, mother's and father's education and earnings, immigrant status, and indicators for the preschool center). We add an indicator for missing baseline scores in all our estimations.[9]

As is evident from our contact rate across waves, attrition was generally low. Even more importantly, as we show in Table D.2, attrition rates were balanced across gender and treatment status.

5.2 Summary Statistics

We combine our assessment data with registry data from Statistics Norway relating to child and parent characteristics. The key variables used in our analyses are listed in Table 1, where we report means and standard deviations for the T3 analytical sample separately across gender and treatment status.[10] Birth month is a running variable taking a value of 1 (December) to 12 (January), so that a higher value indicates an older child. Immigrant status is denoted by an indicator taking the value 1 if the child's mother or father is a non-Western immigrant. Mother's and

[9] In Table D.7 in the appendix we replicate our analysis after excluding observations with imputed pre-scores. It does not affect our overall results and conclusions.

[10] The appendix provides similar information for the T2 sample.

father's education is measured in years of schooling, and their annual earnings are measured in Norwegian kroner on a running scale rounded to the nearest 50,000. We also report summary statistics on the baseline scores of the children. The values presented correspond to the average of the subgroup relative to a sample mean of 0 and expressed in standard deviation units. In the final row, we report the proportion of children without baseline scores.

Table 1 also presents results (in the columns labeled *Difference*) from a test to determine whether child and parent characteristics and baseline scores are balanced across treatment status within each gender. The test consists of regressing the covariate on treatment status while controlling for randomization block. For both genders, background characteristics are sufficiently balanced: no differences that are significant at conventional levels are uncovered. In addition, the magnitudes are also too small to be economically meaningful. We find that treated boys, on average, score somewhat higher at baseline than those in the control group, but their scores are not significantly different. However, we do find a gap in the girls' language score which is of a meaningful magnitude. Such imbalances, even though they might occur by random chance, highlight the importance of controlling for baseline performance, which we do in all our preferred specifications.

5.3 Empirical Strategy

We leverage the randomization to treatment to identify the gender-specific effects of our intervention. To quantify these effects, we use ordinary least squares to estimate models of the form

$$y_{i,c} = \alpha + \gamma_1(Boy_i \times T_c) + \gamma_2(Girl_i \times T_c) + \delta Girl_i + \boldsymbol{\beta X_i} + \epsilon_{i,c} \quad (1)$$

where $y_{i,c}$ is the score for the outcome of child i enrolled in center c. Treatment status is denoted by the indicator T_c taking the value 1 if the child's center was randomized to treatment. We interact the treatment indicator with gender so that γ_1 and γ_2 capture the average treatment effect of being treated for boys and girls separately, enabling us to test whether these effects are statistically different from 0. We also report results from tests

Table 1—Descriptive Statistics and Balance Test

	Boys			Girls		
	Control	Treat	Difference	Control	Treat	Difference
Child characteristics						
Birth Month	6.380	6.249	-0.265	6.139	6.091	0.028
	(3.153)	(3.260)	(0.386)	(3.307)	(3.091)	(0.305)
Immigrant	0.114	0.161	0.041	0.128	0.224	0.091
	(0.319)	(0.369)	(0.045)	(0.336)	(0.418)	(0.059)
Mother Education	14.333	14.128	-0.115	14.433	14.123	-0.224
	(2.495)	(2.520)	(0.260)	(2.602)	(2.635)	(0.291)
Father Education	13.896	13.656	0.260	13.676	13.786	0.007
	(2.426)	(2.422)	(0.291)	(2.640)	(2.532)	(0.310)
Mother Earnings	344,680	329,301	-17,695	345,774	333,160	-12,373
	(225,887)	(200,712)	(31,008)	(216,475)	(200,750)	(30,897)
Father Earnings	571,014	565,968	1,805	525,675	559,337	38,296
	(268,260)	(262,916)	(27,804)	(256,313)	(284,157)	(29,256)
Baseline Scores						
T1 Sum Score	-0.123	-0.068	0.016	0.172	0.039	-0.116
	(1.050)	(0.970)	(0.133)	(0.964)	(1.021)	(0.081)
T1 Math	-0.117	-0.090	0.010	0.140	0.080	-0.055
	(0.982)	(1.019)	(0.113)	(0.915)	(1.024)	(0.130)
T1 EF	-0.115	-0.044	0.047	0.064	0.089	0.045
	(1.055)	(0.982)	(0.141)	(0.958)	(1.019)	(0.090)
T1 Language	-0.065	-0.030	-0.017	0.212	-0.075	-0.270*
	(0.986)	(0.941)	(0.110)	(1.069)	(1.018)	(0.126)
Missing T1 Scores	0.070	0.052	-0.020	0.046	0.029	-0.024
	(0.257)	(0.222)	(0.032)	(0.211)	(0.167)	(0.026)
N	142	193	335	151	175	326

Note: The columns provide means (standard deviations) for child characteristics and T1 test scores separately by gender and treatment status for the T3 analytic sample. The columns labeled *Difference* represent the estimated coefficient (standard error) from regressing each covariate against treatment status, while controlling for randomization block. Regressions are also clustered on the block level.

to determine whether the effects are statistically different from each other. Our preferred specification includes controls for baseline test scores, block fixed effects, and a vector of child and parent background characteristics.

We also add indicators for turning in the consent sheet on time and for being assessed at baseline. $\epsilon_{i,c}$ is the error term. We estimate the models separately for each of the outcome measures, and for T2 and T3 scores.

For our skill-heterogeneity analysis, we extend (1) to include indicators for specific segments of the test-score distribution at baseline. Hence we estimate the model

$$\begin{aligned} y_{i,c} = &\alpha + \phi_1(Boy_i \times T_c \times BS_i^{Boy}) + \phi_2(Girl_i \times T_c \times BS_i^{Girl}) \\ &+ \gamma_1(Boy_i \times T_c) + \gamma_2(Girl_i \times T_c) + \theta_1 BS_i^{Boy} \\ &+ \theta_2 BS_i^{Girl} + \delta Girl_i + \boldsymbol{\beta X_i} + \epsilon_{i,c} \end{aligned} \tag{2}$$

where BS_i^{Boy} is an indicator taking the value 1 if child i is a boy with a baseline score in the relevant segment, and BS_i^{Girl} is the female equivalent. We focus primarily on those scoring in the bottom 10, bottom 25, bottom 50, top 25 and top 10 percent. The coefficient ϕ captures the marginal treatment effect of being a child in the particular segment relative to the rest of the treated children of the same gender.

We compute standard errors that are robust to serial correlation by clustering at the level of randomization (the blocks). A potential concern with this approach is that 15 clusters are too few to provide reliable inference (Cameron and Miller, 2015). To assess whether this concern is warranted here, we also use two alternative approaches that are more robust to small-sample issues. First, we account for the small number of clusters by performing a Wild T bootstrap procedure. Second, we perform a permutation test (randomization inference) where we randomly reassign treatment status within the blocks to estimate a distribution of placebo treatment effects with which we can compare are our true effect estimate (Abadie et al., 2020; Athey and Imbens, 2017). In our results, we report p-values obtained from estimations both with and without these corrections.

6 Results

6.1 Descriptive Evidence

We begin our presentation of the results with a discussion of the descriptive evidence relating to early gender gaps in our sample. In Figure 1 we present differences in school-readiness skills measured by test-score performance at baseline. The bars indicate the gap in average scores between girls and boys in standard deviation units (σ). Assuming that we can interpret the *sum score* as a measure of the child's overall school readiness, we find that girls are, on average, about 0.15σ more ready for school than boys. Moreover, we find that girls score better on all measures, with the largest discrepancy found for mathematics (0.2σ).

While average scores may provide useful information, they may also mask important variation over the skill distribution. In Figure 2 we present density plots for the distribution of baseline scores across the four skill measures. We find that boys are more likely to score in the bottom half of the distribution than girls. For all measures, there are about 60 percent boys among those scoring 1.5σ or more below the mean. However, there are equal numbers of boys and girls scoring 1.5σ or more *above* the mean. This suggests that the discrepancies are driven in large part by low-achieving boys at the bottom of the distribution. To determine whether the boys in the lower parts of the skill distribution are particularly responsive to our intervention, we perform a skill-heterogeneity analysis in Section 6.3.

While our data do not allow us to investigate the extent to which the centers actually *cause* these gender gaps, we find little evidence that business-as-usual mitigates them. Indeed, from T1 to T3 the gender gap in the *sum score* actually increases from 0.29σ to 0.34σ in the control group (although the difference is not statistically significant). This is driven by growing gender gaps in executive functioning and language, while the mathematics gap decreases, particularly once the children start formal schooling.

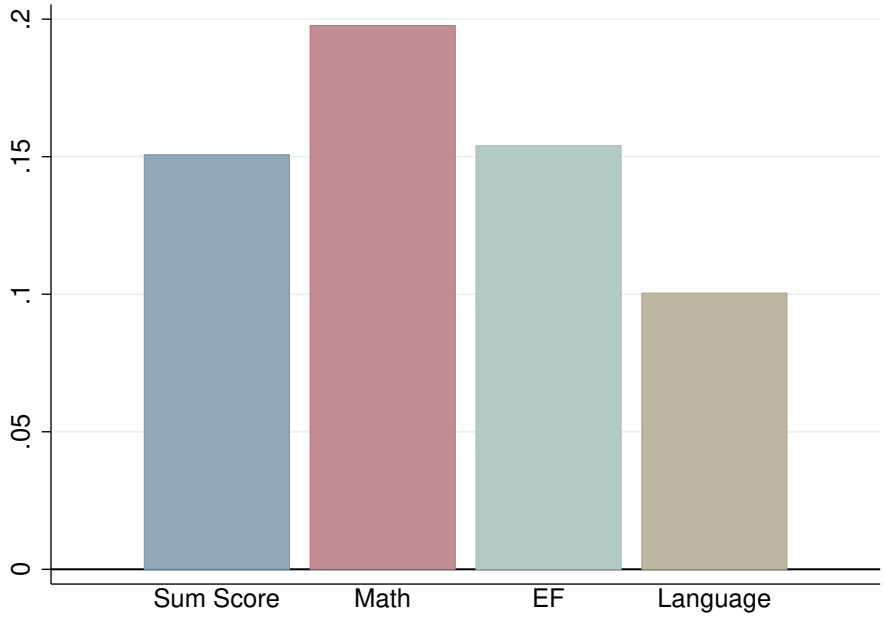


Figure 1: Gender Skill Gap Prior to Treatment

Note: The bars represent the difference in average scores between boys and girls on our four school-readiness measures at baseline. Higher values on the y-axis indicate a larger gap in favor of girls. Scores are standardized so that values represent difference in skills in standard deviation units.

6.2 Main Results

We report the main results from estimating Equation (1) in Table 2. In the leftmost panel we find large and positive point estimates for the treatment effect on boys at the T2 assessment. For the *sum score* measure we estimate that the intervention improved boys' school readiness by 0.19σ. For an intuitive comparison, this effect size is equivalent to 4 months of development for the boys in the control group. For girls, we find much smaller and statistically insignificant estimates. Such a pattern of large discrepancies is found consistently across all our outcome measures. The intervention seems to have a limited effect on girls, while boys seem to be the primary beneficiaries, with effects of meaningful magnitudes across all

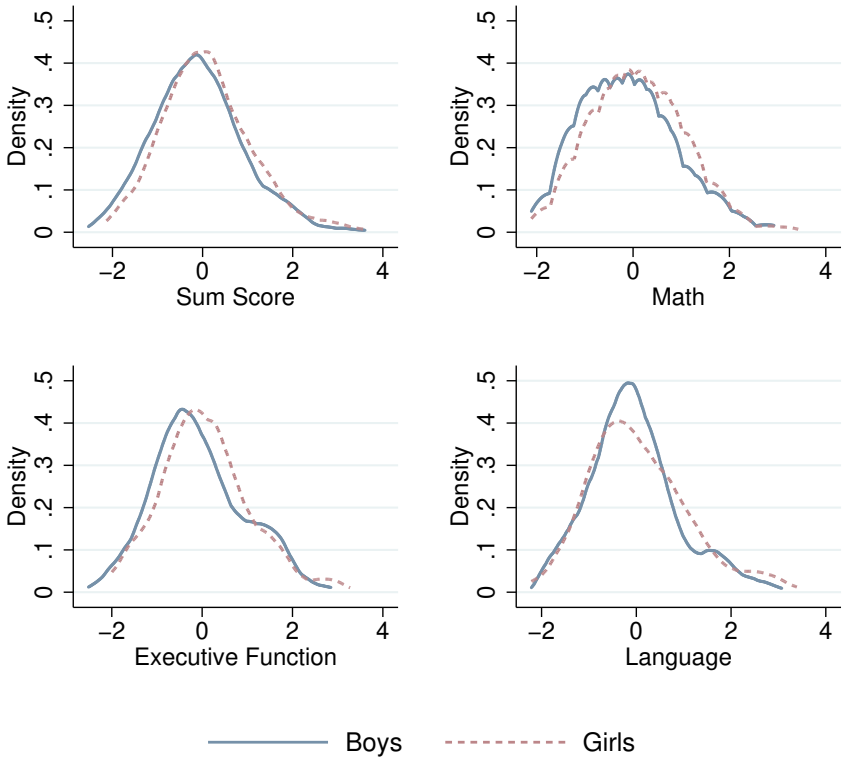


Figure 2: Distribution of Baseline Scores by Gender

Note: The figure presents density plots of the baseline test scores on our four school-readiness measures. The scores are standardized so that 0 corresponds to the sample mean on that measure. We use epanechnikov kernels in all plots.

measures, except language.

In row 3 we report the results of testing for significant differences between the two gender-specific estimates. While the difference in point estimates for all measures is substantial—ranging from 0.087σ to 0.16σ—we often lack the precision necessary to attain significance at conventional levels for outcomes other than the *sum score* measure. Even so, we argue that the meaningful and consistent differences found between the estimates strongly suggest that the boys are the primary beneficiaries of our

intervention, and also that our results provide moderately strong evidence that the difference is not zero (Romer, 2020).

In the rightmost panel we find a similar pattern for treatment effects at the one-year follow-up. In fact, T3 estimates actually exceed T2 ones. For boys, we find large, positive effects, while for girls we find small effects that are both statistically and substantively insignificant. The gender difference in the *sum score* measure is 0.20σ. The largest effect is on boys' mathematical skills, for which we estimate a treatment effect of 0.33σ, which also reflects a substantial increase from the T2 assessment.

Table 2—Gender Specific Treatment Effects

	Post-Intervention (T2)				*Follow-Up (T3)*			
	Sum Score	Math	EF	Language	Sum Score	Math	EF	Language
Treatment Effect	0.191*	0.197	0.201*	0.059	0.235*	0.330**	0.137	0.108
Boys	(0.079)	(0.120)	(0.071)	(0.083)	(0.097)	(0.089)	(0.110)	(0.099)
Treatment Effect	0.046	0.110	0.041	-0.041	0.025	0.117	-0.028	-0.027
Girls	(0.069)	(0.090)	(0.083)	(0.086)	(0.089)	(0.114)	(0.055)	(0.105)
Difference	-0.145+	-0.087	-0.160	-0.010	-0.201+	-0.213	-0.165	-0.135
	(0.082)	(0.104)	(0.102)	(0.101)	(0.118)	(0.160)	(0.137)	(0.084)
p-value	0.098	0.420	0.140	0.340	0.097	0.206	0.248	0.132
Wild Cluster	0.127	0.431	0.150	0.419	0.109	0.241	0.263	0.138
RI	0.133	0.433	0.150	0.419	0.072	0.093	0.188	0.278
N	652	650	652	648	661	661	660	659
Adj. R^2	0.61	0.44	0.49	0.53	0.55	0.40	0.41	0.52

Note: Each column in each panel presents the regression coefficient of treated (standard error) interacted with gender using ordinary least squares. The coefficient for the interaction gives the total treatment effect on that gender, so that the difference between them represents the marginal effect. The *Difference* panel reports coefficients and errors for tests on significant differences between the gender specific estimates. Below we report three sets of p-values computed using clustering on block, the Wild T bootstrap procedure, and randomization inference respectively. For both assessment periods we regress outcome on the treatment–gender interaction, controlling for baseline test scores, gender, birth month, parental characteristics (mother and father's education level, earnings, and indicator for non-Western country of birth), and indicators for late consent and not having participated in the T1 assessment. All regressions are clustered on, and control for, randomization block. $^{+}$ $p<0.1$, * $p<0.05$, ** $p<0.01$.

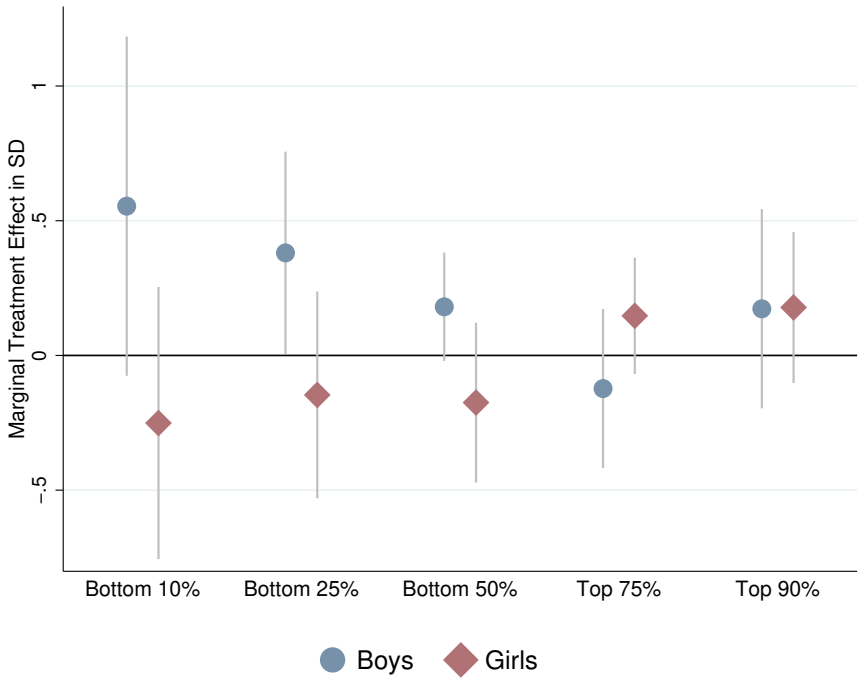


Figure 3: Heterogeneous Effects by Baseline Skills

Note: In this figure we plot estimates of the marginal treatment effect of placing in a specific segment of the *Sum Score* distribution for your gender at baseline. Each of the five circle/diamond pairs represents a separate regression, where the plot gives the estimated coefficient of the three-way gender×treatment×baseline score segment interaction. For example, the left-most circle gives the marginal treatment effect for boys scoring in the lowest 10% of boys at baseline, relative to the treatment effect for all other boys. The grey lines indicate 90% confidence bands.

6.3 Treatment Effect Heterogeneity by Baseline Skill

Policy discussions often have a particular focus on *low-achieving* boys, who are deemed to be at greatest risk and the most vulnerable (see, e.g., Chetty et al., 2016, and Autor et al., 2020). Against this backdrop, we consider it relevant to investigate whether treatment effects are heterogeneous to the baseline skill level. To do so, we plot the results from estimating (2) in Figure 3. What are presented in the figure are five pairs of plots representing results for boys and girls, respectively, stemming

from separate regressions. As an example, the leftmost circle (diamond) represents the marginal treatment effect for boys (girls) scoring in the lowest 10 percent of boys (girls) at baseline, relative to the treatment effect for all other boys (girls). Moving along the horizontal axis implies moving upward in the baseline skill distribution.

We see a clear downward trend in the marginal effects for boys. Starting with those scoring in the lowest 10 percent at baseline, we find a marginal treatment effect of over 0.5σ, although imprecisely estimated. It should be noted that this marginal effect is in addition to the average treatment effect on treated boys, meaning that the total effect sums up to about 0.7σ. The marginal effect then declines rapidly as we move to higher-achieving boys, converging to zero for the top of the distribution. This pattern suggests that the majority of the treatment effect is concentrated in the group of initially low-achieving boys.

For girls, we do not find a similar pattern. If anything, we find some indication that the highest-achieving girls may have benefited the most from the intervention. Overall, however, the small (if any) treatment effect on girls seems fairly homogeneous across the distribution.

7 Implications

Our overall finding is that there are substantial gender gaps in early skills crucial for future learning, measured prior to enrollment in formal schooling.[11] This implies that Norwegian girls start school with a significant skill advantage over boys. This difference in school readiness may explain, at least partly, why boys tend to fall behind as they progress through the education system.

The structured curriculum investigated in the present study boosts school readiness on average, and we find that these positive effects are

[11] In a related paper, Thijssen et al. (2021) show that the boost in executive function stemming from the intervention is particularly crucial as it also bolsters the development of language and mathematical skills after the children move on to primary school.

strong primarily for boys. Moreover, we present suggestive evidence that the lowest-achieving boys benefit the most from this curriculum. Our intervention could therefore improve conditions for later learning in this group. In sum, these results imply that introducing more structured activities could be beneficial for boys in contexts where ECEC practice is primarily centered around free play.

Our findings are therefore relevant for the design of early childhood curricula and pedagogical practice. The extant literature has emphasized the need for a better understanding of the process and heterogeneous impacts of early childhood education. Our study suggests that failing to provide boys in particular with the appropriate amount of scaffolding in their ECEC environment might exacerbate gender gaps in early learning. In general, further prying open the black box that is ECEC quality, so as to understand its impact on other child subgroups, is an important task for future research.

References

Abadie, A., Athey, S., Imbens, G. W., & Wooldridge, J. M. (2020). Sampling-Based versus Design-Based Uncertainty in Regression Analysis. *Econometrica*, *88*(1), 265–296.

Anderson, M. L. (2008). Multiple Inference and Gender Differences in the Effects of Early Intervention: A Reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects. *Journal of the American Statistical Association*, *103*(484), 1481–1495.

Athey, S., & Imbens, G. (2017). The Econometrics of Randomized Experiments. In A. V. Banerjee & E. Duflo (Eds.), *Handbook of Economic Field Experiments* (Chap. 3, Vol. 1, pp. 73 –140).

Autor, D., Figlio, D. N., Karbownik, K., Roth, J., & Wasserman, M. (2020). Males at the Tails: How Socioeconomic Status Shapes the Gender Gap. *NBER Working Paper*, (27196).

Autor, D., & Wasserman, M. (2013). Wayward Sons: The Emerging Gender Gap in Education and Labor Markets. *NEXT Series*.

Bedard, K., & Cho, I. (2010). Early Gender Test Score Gaps Across OECD Countries. *Economics of Education Review*, *29*(3), 348 –363.

Bennett, J., & Tayler, C. (2006). *Starting Strong II: Early Childhood Education and Care*. Paris, France: OECD Publishing.

Berlinski, S., Galiani, S., & Manacorda, M. (2008). Giving Children a Better Start: Preschool Attendance and School-Age Profiles. *Journal of Public Economics*, *92*(5), 1416 –1440.

Blau, F. D., & Kahn, L. M. (2007). The Gender Pay Gap. *Academy of Management Perspectives*, *21*(1), 7–23.

Brandlistuen, R. E., Flatø, M., Stoltenberg, C., Helland, S. S., & Wang, M. V. (2020). Gender Gaps in Preschool Age: A Study of Behavior, Neurodevelopment and Pre-Academic Skills. *Scandinavian Journal of Public Health*, 140349482094474.

Cameron, A., & Miller, D. (2015). A Pactitioner's Guide to Cluster-Robust Inference. *Journal of Human Resources*, *50*(2), 317–372.

Chetty, R., Hendren, N., Lin, F., Majerovitz, J., & Scuderi, B. (2016). Childhood Environment and Gender Gaps in Adulthood. *American Economic Review*, *106*(5), 282–288.

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

Cornelissen, T., Dustmann, C., Raute, A., & Schönberg, U. (2018). Who Benefits From Universal Child Care? Estimating Marginal Returns to Early Child Care Attendance. *Journal of Political Economy*, *126*(6), 2356–2409.

Cunha, F., & Heckman, J. (2007). The Technology of Skill Formation. *American Economic Review*, *97*(2), 31–47.

Davidson, M. C., Amso, D., Anderson, L. C., & Diamond, A. (2006). Development of Cognitive Control and Executive Functions From 4 to 13 Years: Evidence From Manipulations of Memory, Inhibition, and Task Switching. *Neuropsychologia*, *44*(11), 2037 –2078. Advances in Developmental Cognitive Neuroscience.

Deming, D., & Dynarski, S. (2008). The Lengthening of Childhood. *Journal of Economic Perspectives*, *22*(3), 71–92.

DiPrete, T. A., & Buchmann, C. (2013). *Rise of Women, The: The Growing Gender Gap in Education and What it Means for American Schools*. Russell Sage Foundation.

DiPrete, T. A., & Jennings, J. L. (2012). Social and Behavioral Skills and the Gender Gap in Early Educational Achievement. *Social Science Research*, *41*(1), 1 –15.

Diamond, A., Barnett, W. S., Thomas, J., & Munro, S. (2007). Preschool Program Improves Cognitive Control. *Science*, *318*(5855), 1387–1388.

Dillon, M. R., Kannan, H., Dean, J. T., Spelke, E. S., & Duflo, E. (2017). Cognitive Science in the Field: A Preschool Intervention Durably Enhances Intuitive but not Formal Mathematics. *Science*, *357*(6346), 47–55.

Domond, P., Orri, M., Algan, Y., Findlay, L., Kohen, D., Vitaro, F., ... Côté, S. M. (2020). Child Care Attendance and Educational and Economic Outcomes in Adulthood. *Pediatrics*, *146*(1).

Drange, N., & Rønning, M. (2020). Child Care Center Quality and Early Child Development. *Journal of Public Economics*, *188*, 104204.

Duncan, G. J., & Magnuson, K. (2013). Investing in Preschool Programs. *Journal of Economic Perspectives*, *27*(2), 109–132.

Early, D. M., Iruka, I. U., Ritchie, S., Barbarin, O. A., Winn, D.-M. C., Crawford, G. M., ... Pianta, R. C. (2010). How do Pre-Kindergarteners Spend Their Time? Gender, Ethnicity, and Income as Predictors of Experiences in Pre-Kindergarten Classrooms. *Early Childhood Research Quarterly*, *25*(2), 177 –193.

Engel, A., Barnett, W. S., Anders, Y., & Taguma, M. (2015). *Early Childhood Education and Care Policy Review*. Paris, France: OECD Publishing.

Felfe, C., & Lalive, R. (2018). Does Early Child Care Affect Children's Development? *Journal of Public Economics*, *159*, 33–53.

Felfe, C., Nollenberger, N., & Rodríguez-Planas, N. (2015). Can't Buy Mommy's Love? Universal Childcare and Children's Long-Term Cognitive Development. *Journal of Population Economics*, *28*(2), 393–422.

Fessler, P., & Schneebaum, A. (2019). The Educational and Labor Market Returns to Preschool Attendance in Austria. *Applied Economics*, *51*(32), 3531–3550.

Fortin, N. M., Oreopoulos, P., & Phipps, S. (2015). Leaving Boys Behind Gender Disparities in High Academic Achievement. *Journal of Human Resources*, *50*(3), 549–579.

Goldin, C., Katz, L. F., & Kuziemko, I. (2006). The Homecoming of American College Women: The Reversal of the College Gender Gap. *Journal of Economic Perspectives*, *20*(4), 133–156.

Goodman, A., & Sianesi, B. (2005). Early Education and Children's Outcomes: How Long Do the Impacts Last? *Fiscal Studies*, *26*(4), 513–548.

Gray-Lobe, G., Pathak, P. A., & Walters, C. R. (2021). The Long-Term Effects of Universal Preschool in Boston. *NBER Working Paper Series*, (28756).

Gørtz, M., Rye Johansen, E., & Simonsen, M. (2018). Academic Achievement and the Gender Composition of Preschool Staff. *Labour Economics*, *55*, 241–258.

Havnes, T., & Mogstad, M. (2015). Is Universal Child Care Leveling the Playing Field? *Journal of Public Economics*, *127*, 100 –114. The Nordic Model.

Heckman, J. J., Moon, S. H., Pinto, R., Savelyev, P. A., & Yavitz, A. (2010). The Rate of Return to the HighScope Perry Preschool Program. *Journal of Public Economics*, *94*(1), 114 –128.

Howes, C., Burchinal, M., Pianta, R., Bryant, D., Early, D., Clifford, R., & Barbarin, O. (2008). Ready to Learn? Children's Pre-Academic Achievement in Pre-Kindergarten Programs. *Early Childhood Research Quarterly*, *23*(1), 27–50.

Karlsen, L., & Lekhal, R. (2019). Practitioner Involvement and Support in Children's Learning During Free Play in Two Norwegian Kindergartens. *Journal of Early Childhood Research*, *17*(3), 233–246.

Lenes, R., Gonzales, C. R., Størksen, I., & McClelland, M. M. (2020). Children's Self-Regulation in Norway and the United States: The Role of Mother's Education and Child Gender Across Cultural Contexts. *Frontiers in Psychology*, *11*, 2563.

Magnuson, K. A., Kelchen, R., Duncan, G. J., Schindler, H. S., Shager, H., & Yoshikawa, H. (2016). Do the Effects of Early Childhood Education Programs Differ by Gender? A Meta-Analysis. *Early Childhood Research Quarterly*, *36*, 521–536.

Magnuson, K., & Duncan, G. J. (2016). Can Early Childhood Interventions Decrease Inequality of Economic Opportunity? *RSF: The Russell Sage Foundation Journal of the Social Sciences*, *2*(2), 123–141.

Mashburn, A. J., Pianta, R. C., Hamre, B. K., Downer, J. T., Barbarin, O. A., Bryant, D., ... Howes, C. (2008). Measures of Classroom Quality in Prekindergarten and Children's Development of Academic, Language, and Social Skills. *Child Development*, *79*(3), 732–749.

McClelland, M. M., Cameron, C. E., Duncan, R., Bowles, R. P., Acock, A. C., Miao, A., & Pratt, M. E. (2014). Predictors of Early Growth in Academic Achievement: The Head-Toes-Knees-Shoulders Task. *Frontiers in Psychology*, *5*, 599.

Melhuish, E. C. (2011). Preschool Matters. *Science*, *333*(6040), 299–300.

Norwegian Directorate for Education and Training. (2019). *Utdanningsspeilet [The Education Mirror]*. Oslo, Norway.

OECD. (2015). *The ABC of Gender Equality in Education: Aptitude, Behaviour, Confidence*.

Phillips, D., Lipsey, M., Dodge, K., Haskins, R., Bassok, D., Burchinal, M., ... Weiland, C. (2017). Puzzling It Out: The Current State of Scientific Knowledge on Pre-Kindergarten Effects—A Consensus Statement. *Issues in Pre-Kindergarten Programs and Policy*, 19–30.

Rege, M., Størksen, I., Solli, I. F., Kalil, A., McClelland, M., ten Braak, D., ... Hundeland, P. S. (2021). The Effects of a Structured Curriculum on Preschool Effectiveness: A Field Experiment. *CESifo Working Paper*, (7775).

Romer, D. (2020). In Praise of Confidence Intervals. *AEA Papers and Proceedings*, *110*, 55–60.

Schmitt, S. A., McClelland, M. M., Tominey, S. L., & Acock, A. C. (2015). Strengthening School Readiness for Head Start Children: Evaluation of a Self-Regulation Intervention. *Early Childhood Research Quarterly*, *30*, 20 –31.

Spelke, E. S. (2005). Sex Differences in Intrinsic Aptitude for Mathematics and Science?: A Critical Review. *American psychologist.*, *60*(9), 950–958.

Stangeland, E. B., Lundetræ, K., & Reikerås, E. (2018). Gender Differences in Toddlers' Language and Participation in Language Activities in Norwegian ECEC Institutions. *European Early Childhood Education Research Journal*, *26*(3), 375–392.

Størksen, I., Ellingsen, I. T., Tvedt, M. S., & Idsøe, E. M. (2013). Norsk vokabulartest (nvt) for barn i overgangen mellom barnehage og skole: Psykometrisk vurdering av en nettbrettbasert test. *Spesialpedagogikks forskningsdel*, *4*(13), 40–54.

Størksen, I., ten Braak, D., Breive, S., Lenes, R., Lunde, S., Carlsen, M., ... Rege, M. (2018). *Playful learning - A Research Based Preschool Curriculum from the Agder Project*. GAN Aschehoug.

Størksen, I., Ellingsen, I. T., Wanless, S. B., & McClelland, M. M. (2015). The Influence of Parental Socioeconomic Background and Gender on Self-Regulation Among 5-Year-Old Children in Norway. *Early Education and Development*, *26*(5-6), 663–684.

Størksen, I., Ertesvåg, S. K., & Rege, M. (2021). Implementing Implementation Science in a Randomized Controlled Trial in Norwegian Early Childhood Education and Care. *International Journal of Educational Research*, *108*, 101782.

Sumsion, J. (2005). Male Teachers in Early Childhood Education: Issues and Case Study. *Early Childhood Research Quarterly*, *20*(1), 109 –123.

Synodi, E. (2010). Play in the Kindergarten: The Case of Norway, Sweden, New Zealand and Japan. *International Journal of Early Years Education*, *18*(3), 185–200.

ten Braak, D., & Størksen, I. (2021). Psychometric Properties of the Ani Banani Math Test. *European Journal of Developmental Psychology*, *Forthcoming*.

Thijssen, M. W. P., Rege, M., Solli, I. F., & Størksen, I. (2021). On the Cross-Productivity of Executive Functions: Results from a Randomized Early Childhood Program. *Working Paper*.

Tonyan, H. A., & Howes, C. (2003). Exploring Patterns in Time Children Spend in a Variety of Child Care Activities: Associations with Environmental Quality, Ethnicity, and Gender. *Early Childhood Research Quarterly*, *18*(1), 121 –142.

Vincent-Lancrin, S. (2008). The Reversal of Gender Inequalities in Higher Educationn: An On-Going Trend. In OECD (Ed.), *Higher Education to 2030: Demography* (Vol. 1).

Wechsler, D. (2003). *Wechsler Intelligence Scale for Children–Fourth Edition (WISC-IV)*. San Antonio, TX: The Psychological Corporation.

Weiland, C., & Yoshikawa, H. (2013). Impacts of a Prekindergarten Program on Children's Mathematics, Language, Literacy, Executive Function, and Emotional Skills. *Child Development*, *84*(6), 2112–2130.

Weisberg, D. S., Hirsh-Pasek, K., & Golinkoff, R. M. (2013). Guided Play: Where Curricular Goals Meet a Playful Pedagogy. *Mind, Brain, and Education*, *7*(2), 104–112.

White, L. A., Prentice, S., & Perlman, M. (2015). The Evidence Base for Early Childhood Education and Care Programme Investment: What We Know, What We Don't Know. *Evidence & Policy*, *11*(4), 529–546.

Appendix

Appendix A:

Experimental Design and Compliance

This section provides additional details on the experimental design. After randomization and collection of informed consent the intervention followed the timeline laid out in Figure A.1. Starting in the fall of 2015 the participating teachers in the treatment group received training in the form of a credit-earning course at the University on how to incorporate the curriculum in their daily practice. Participants earned 15 ECTS for completing the course, equivalent to 1/4 of a full course load for one academic year in the Norwegian university system. The class consisted of insights from the theoretical and empirical research literature on the pedagogics behind playful learning, and was to a large extent practice-oriented. Required learning activities included four two-day lecture gatherings over a period of eight months, and teachers were expected to practice practice playful learning activities with their current five-year-olds (who are not part of the sample used in the analysis) between class gatherings. The course also allowed us to collect feedback from the teachers on the feasibility and usefulness of the curriculum, and adjust accordingly.

The teachers subsequently implemented the curriculum starting in the fall of 2016 This coincided with the start of the Norwegian academic year, which runs from medio August to late June. We conducted the baseline assessment, referred to as *Assessment T1* in the figure, in the final week of August, immediately prior to implementation. The treatment group then proceeded with the intervention the following nine months, before we conducted the postintervention (T2) assessment in June of 2017. We then reconnected with the children one year later to conduct our follow-up assessment (T3) in March of 2018. The teachers in the control group were offered the same teacher training as the treatment group received in the fall of 2017 once the participating children had started primary school.

We assessed compliance through weekly surveys of the participating teachers. In electronic questionnaires the teachers reported, among other things, how many hours they had spent conducting activities from the curriculum. In Figure A.2 we present the distribution of time spent per

week as reported by the teachers. We requested that they spent at least 8 hours a week doing so, and find in the surveys that 60 percent of centers spent an average of 7 or higher over the implementation period. In total the teachers were requested to submit 34 weekly reports, with the majority of centers successfully doing so for all weeks (see the bottom panel of Figure A.2 for the distribution). Of the 974 reports collected in total 67 percent report having met the 8 hour request the previous week, while in contrast only 16 percent reported spending less than 6. Based on these teacher reports we evaluate compliance to be satisfactory.

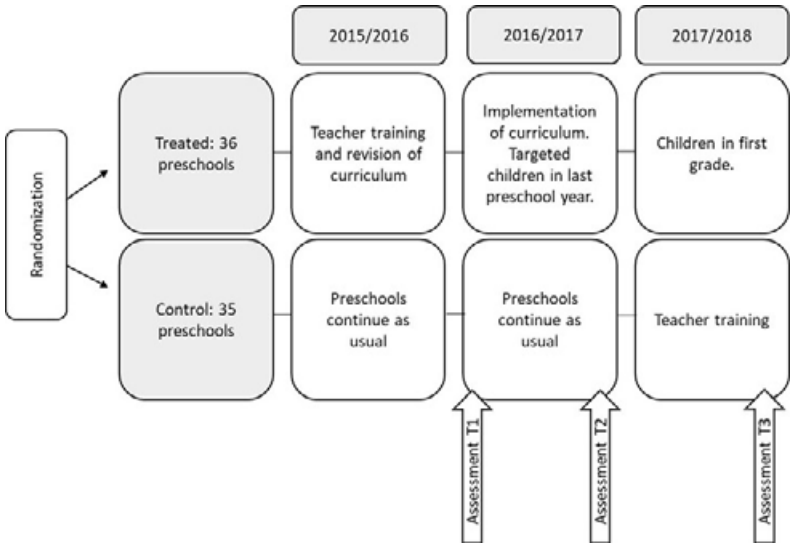


Figure A.1: Experimental Design

Note: 71 preschool centers randomly split between control and treatment. Preschool year 2015/2016: Teachers in treated centers attended the teacher training and helped revise the curriculum. 2016/2017: Teachers in treated implemented the structured curriculum with the five-year-olds in their center. 2017/2018: Teachers in control received the teacher training. We assessed in August 2016 (baseline, T1), June 2017 (postintervention, T2), and March 2018 (follow-up, T3). The figure was first presented in Rege et al. (2020).

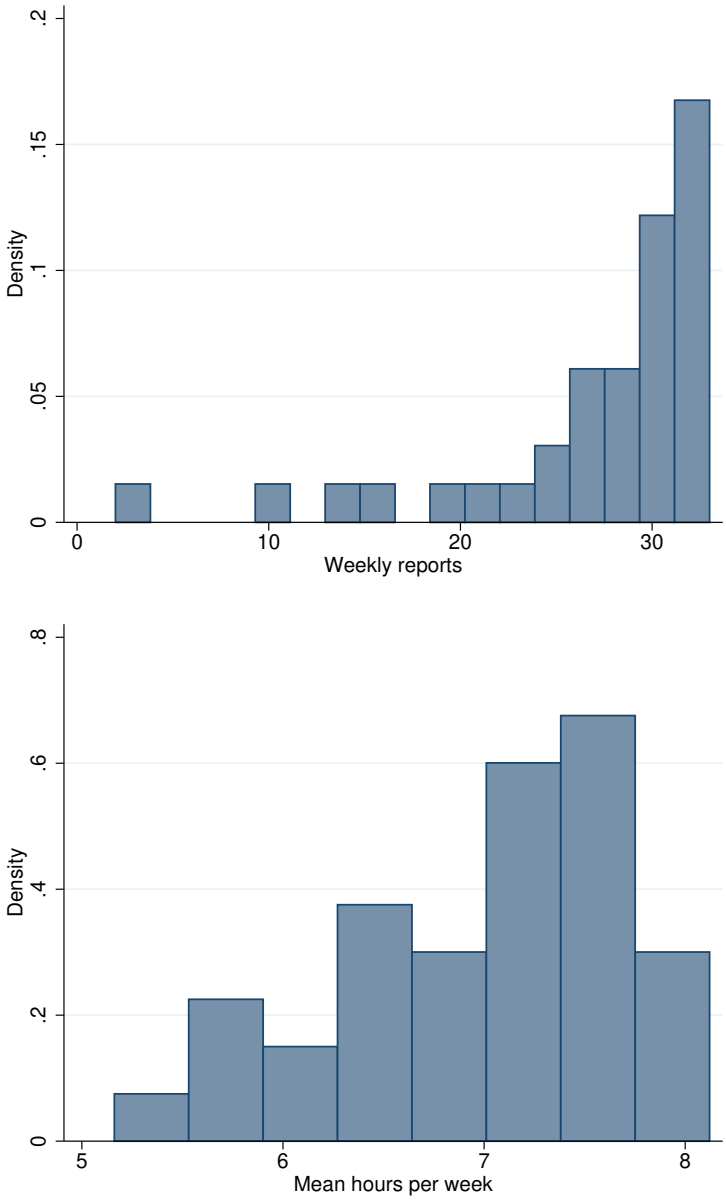


Figure A.2: Number of Weekly Reports (Top) and Hours Spent on Activities (Bottom) by Centers

Note: These figures were first presented in Rege et al. (2020).

Appendix B:

Heterogeneous Effects by Baseline Skills

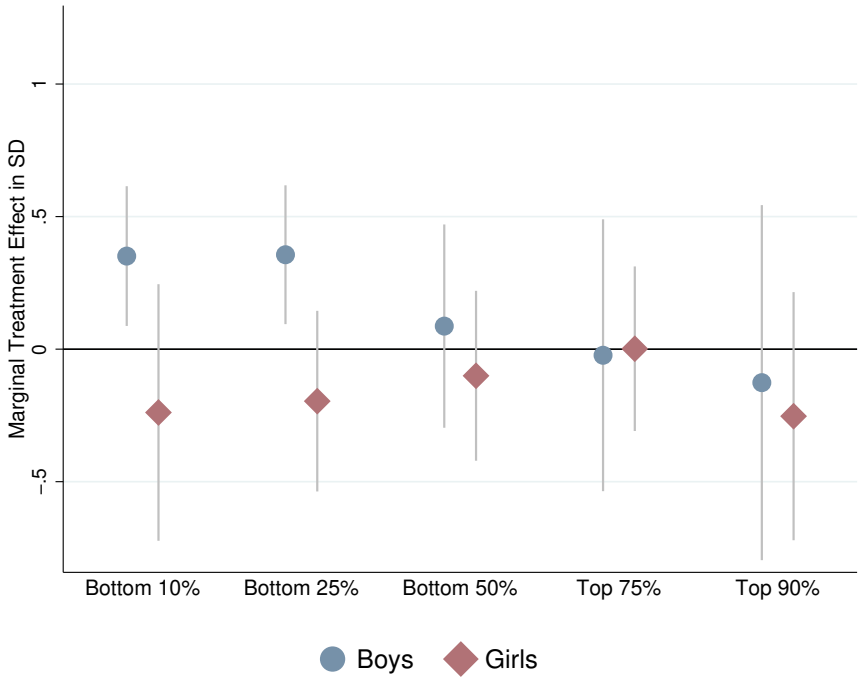


Figure B.1: Math

Note: In figures B.1–B.3 we plot estimates of the marginal treatment effect of placing in a specific segment of the score distribution for your gender at baseline, for separately for each of the three skill domains. Each of the five circle/diamond pairs represent a separate regression, where the plot gives the estimated coefficient of the three-way gender×treatment×baseline score segment interaction. For example, the left-most circle gives the marginal treatment effect for boys scoring in the lowest 10% of boys at baseline, relative to the treatment effect for all other boys. The grey lines indicate 90% confidence bands.

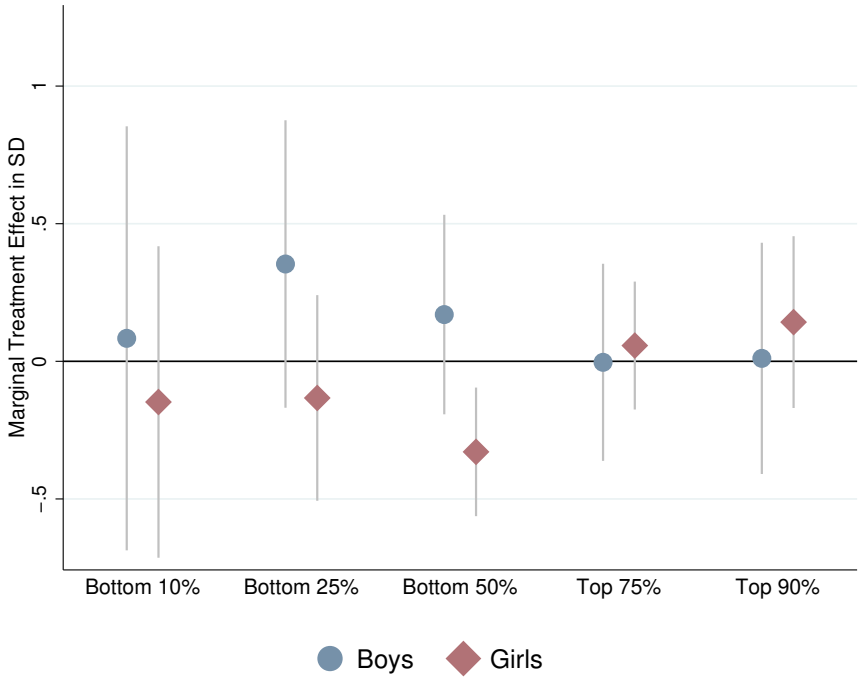


Figure B.2: Executive Function

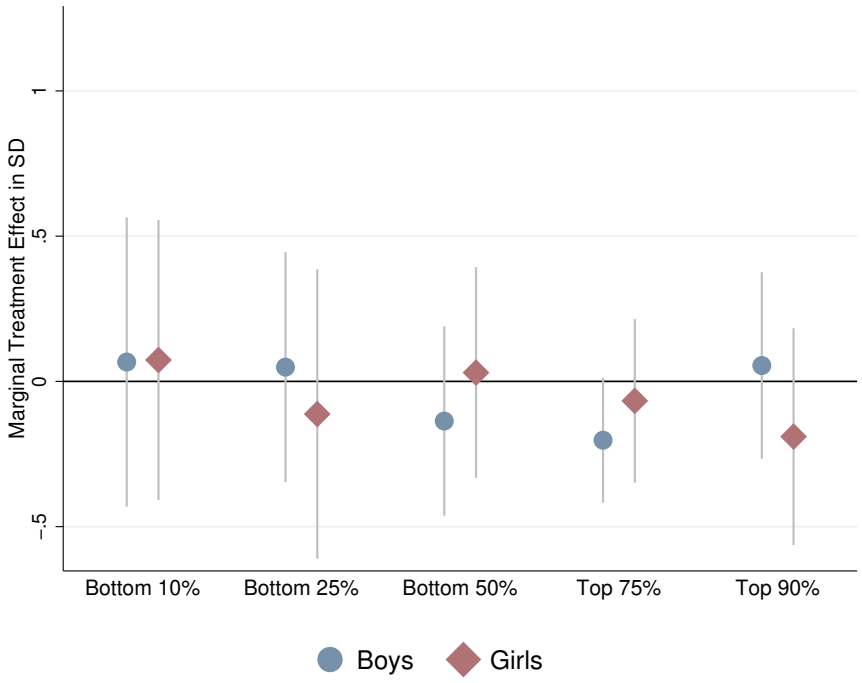


Figure B.3: Language

Table B1.A—Sum Score

	Post-Intervention (T2)					*Follow-Up (T3)*				
	Bottom 10%	Bottom 25%	Bottom 50%	Top 25%	Top 10%	Bottom 10%	Bottom 25%	Bottom 50%	Top 25%	Top 10%
Treatment × Male	0.283	0.124	0.087	-0.100	-0.294	0.554	0.381	0.181	-0.123	0.173
× Rank	(0.220)	(0.156)	(0.181)	(0.164)	(0.293)	(0.357)	(0.213)	(0.114)	(0.168)	(0.210)
Treatment × Male	0.161	0.163	0.144	0.218	0.224	0.171	0.147	0.144	0.266	0.218
	(0.082)	(0.084)	(0.106)	(0.103)	(0.102)	(0.101)	(0.102)	(0.116)	(0.100)	(0.104)
Treatment × Female	-0.025	-0.112	-0.269	0.237	0.066	-0.251	-0.147	-0.175	0.147	0.178
× Rank	(0.181)	(0.163)	(0.149)	(0.210)	(0.376)	(0.287)	(0.218)	(0.168)	(0.123)	(0.159)
Treatment × Female	0.049	0.078	0.168	-0.013	0.040	0.041	0.057	0.105	-0.014	-0.005
	(0.072)	(0.090)	(0.117)	(0.069)	(0.066)	(0.092)	(0.092)	(0.101)	(0.105)	(0.098)
N	652	652	652	652	652	661	661	661	661	661
Adj. R^2	0.61	0.62	0.62	0.61	0.62	0.53	0.53	0.53	0.53	0.53

Note: Each column in each panel presents regression coefficients of the treatment effect (standard error) from separate regressions using ordinary least squares. In the case of each column, all coefficients stem from the same regression, where the slope is allowed to differ between boys and girls. Rank refers to a dummy equal to one if the child scored below/above a certain percentile threshold for their gender for that skill at baseline (e.g. below the 25th percentile for boys). The column headers denote the threshold used in each case. The interpretation of the coefficient $Treatment \times Gender \times Rank$ is thus the effect of being a treated boy or girl starting from a particular segment of the test score distribution at baseline, relative to the gender-specific treatment effect on the rest of the distribution. For both assessment periods the models regresses the outcome on the triple interaction and the gender-treatment interaction, while controlling for baseline test scores, the rank dummy, gender, birth month, parental characteristics (mother and father's education level, earnings, an indicator for non-Western country of birth), and indicators for late consent and not having participated in the T1 assessment. All regressions are clustered on, and control for, randomization block.

Table B1.B—Math

	Post-Intervention (T2)					*Follow-Up (T3)*				
	Bottom 10%	Bottom 25%	Bottom 50%	Top 25%	Top 10%	Bottom 10%	Bottom 25%	Bottom 50%	Top 25%	Top 10%
Treatment × Male	0.174	0.165	0.071	-0.087	-0.518	0.351	0.356	0.087	-0.023	-0.126
× Rank	(0.198)	(0.199)	(0.202)	(0.276)	(0.398)	(0.187)	(0.187)	(0.175)	(0.202)	(0.255)
Treatment × Male	0.150	0.153	0.152	0.212	0.230	0.239	0.237	0.275	0.334	0.336
	(0.129)	(0.131)	(0.129)	(0.151)	(0.129)	(0.090)	(0.090)	(0.125)	(0.101)	(0.090)
Treatment × Female	-0.162	-0.160	-0.020	0.037	-0.287	-0.239	-0.196	-0.100	0.002	-0.253
× Rank	(0.119)	(0.163)	(0.149)	(0.210)	(0.376)	(0.253)	(0.185)	(0.173)	(0.198)	(0.255)
Treatment × Female	0.136	0.158	0.119	0.107	0.126	0.161	0.177	0.167	0.117	0.132
	(0.098)	(0.105)	(0.126)	(0.093)	(0.087)	(0.081)	(0.089)	(0.100)	(0.084)	(0.078)
N	650	650	650	650	650	661	661	661	661	661
Adj. R^2	0.44	0.44	0.44	0.44	0.44	0.37	0.37	0.36	0.36	0.37

Note: Each column in each panel presents regression coefficients of the treatment effect (standard error) from separate regressions using ordinary least squares. In the case of each column, all coefficients stem from the same regression, where the slope is allowed to differ between boys and girls. Rank refers to a dummy equal to one if the child scored below/above a certain percentile threshold for their gender for that skill at baseline (e.g. below the 25th percentile for boys). The column headers denote the threshold used in each case. The interpretation of the coefficient $Treatment \times Gender \times Rank$ is thus the effect of being a treated boy or girl starting from a particular segment of the test score distribution at baseline, relative to the gender-specific treatment effect on the rest of the distribution. For both assessment periods the models regresses the outcome on the triple interaction and the gender-treatment interaction, while controlling for baseline test scores, the rank dummy, gender, birth month, parental characteristics (mother and father's education level, earnings, an indicator for non-Western country of birth), and indicators for late consent and not having participated in the T1 assessment. All regressions are clustered on, and control for, randomization block.

Table B1.C—Executive Function

	Post-Intervention (T2)					*Follow-Up (T3)*				
	Bottom 10%	Bottom 25%	Bottom 50%	Top 25%	Top 10%	Bottom 10%	Bottom 25%	Bottom 50%	Top 25%	Top 10%
Treatment × Male	0.174	0.138	0.010	-0.017	-0.098	0.084	0.354	0.170	-0.004	0.011
× Rank	(0.237)	(0.153)	(0.108)	(0.191)	(0.266)	(0.415)	(0.230)	(0.165)	(0.159)	(0.209)
Treatment × Male	0.184	0.172	0.196	0.210	0.208	0.128	0.051	0.060	0.142	0.139
	(0.063)	(0.058)	(0.077)	(0.085)	(0.066)	(0.098)	(0.099)	(0.108)	(0.106)	(0.095)
Treatment × Female	0.158	0.074	-0.040	-0.049	-0.197	-0.148	-0.133	-0.329	0.057	0.142
× Rank	(0.331)	(0.282)	(0.179)	(0.166)	(0.151)	(0.446)	(0.246)	(0.163)	(0.153)	(0.190)
Treatment × Female	0.048	0.014	0.051	0.057	0.055	-0.001	-0.002	-0.128	-0.037	-0.052
	(0.084)	(0.094)	(0.106)	(0.095)	(0.080)	(0.089)	(0.091)	(0.103)	(0.102)	(0.092)
N	652	652	652	652	652	660	660	660	660	660
Adj. R^2	0.50	0.50	0.49	0.49	0.50	0.38	0.39	0.39	0.38	0.39

Note: Each column in each panel presents regression coefficients of the treatment effect (standard error) from separate regressions using ordinary least squares. In the case of each column, all coefficients stem from the same regression, where the slope is allowed to differ between boys and girls. Rank refers to a dummy equal to one if the child scored below/above a certain percentile threshold for their gender for that skill at baseline (e.g. below the 25th percentile for boys). The column headers denote the threshold used in each case. The interpretation of the coefficient $Treatment \times Gender \times Rank$ is thus the effect of being a treated boy or girl starting from a particular segment of the test score distribution at baseline, relative to the gender-specific treatment effect on the rest of the distribution. For both assessment periods the models regresses the outcome on the triple interaction and the gender-treatment interaction, while controlling for baseline test scores, the rank dummy, gender, birth month, parental characteristics (mother and father's education level, earnings, an indicator for non-Western country of birth), and indicators for late consent and not having participated in the T1 assessment. All regressions are clustered on, and control for, randomization block.

Table B1.D—Language

	Post-Intervention (T2)					*Follow-Up (T3)*				
	Bottom 10%	Bottom 25%	Bottom 50%	Top 25%	Top 10%	Bottom 10%	Bottom 25%	Bottom 50%	Top 25%	Top 10%
Treatment × Male	-0.169	0.041	-0.080	-0.163	-0.144	0.067	0.049	-0.137	-0.202	0.055
× Rank	(0.244)	(0.115)	(0.167)	(0.206)	(0.243)	(0.364)	(0.198)	(0.146)	(0.162)	(0.205)
Treatment × Male	0.075	0.051	0.099	0.098	0.076	0.102	0.090	0.180	0.158	0.102
	(0.087)	(0.086)	(0.082)	(0.112)	(0.096)	(0.084)	(0.095)	(0.113)	(0.091)	(0.087)
Treatment × Female	-0.201	-0.340	-0.173	0.022	0.133	0.074	-0.112	0.030	-0.067	-0.190
× Rank	(0.243)	(0.229)	(0.215)	(0.196)	(0.274)	(0.285)	(0.188)	(0.180)	(0.180)	(0.202)
Treatment × Female	-0.021	0.040	0.056	-0.045	-0.050	-0.038	-0.004	-0.050	-0.017	-0.010
	(0.097)	(0.121)	(0.155)	(0.083)	(0.085)	(0.094)	(0.103)	(0.122)	(0.104)	(0.095)
N	648	648	648	648	648	659	659	659	659	659
Adj. R^2	0.53	0.53	0.53	0.53	0.53	0.49	0.49	0.49	0.49	0.49

Note: Each column in each panel presents regression coefficients of the treatment effect (standard error) from separate regressions using ordinary least squares. In the case of each column, all coefficients stem from the same regression, where the slope is allowed to differ between boys and girls. Rank refers to a dummy equal to one if the child scored below/above a certain percentile threshold for their gender for that skill at baseline (e.g. below the 25th percentile for boys). The column headers denote the threshold used in each case. The interpretation of the coefficient $Treatment \times Gender \times Rank$ is thus the effect of being a treated boy or girl starting from a particular segment of the test score distribution at baseline, relative to the gender-specific treatment effect on the rest of the distribution. For both assessment periods the models regresses the outcome on the triple interaction and the gender-treatment interaction, while controlling for baseline test scores, the rank dummy, gender, birth month, parental characteristics (mother and father's education level, earnings, an indicator for non-Western country of birth), and indicators for late consent and not having participated in the T1 assessment. All regressions are clustered on, and control for, randomization block.

Appendix C:

Heterogeneity by Socioeconomic Background

The policy interest in the vulnerable boys partly stems from the fact that lower achievement tend to overlap with other forms for disadvantage such as having parents with low education or income. To assess whether our skill heterogeneity results might be driven by low skills proxying for socioeconomic background we perform a similar estimation as that carried out in 6.3 by considering the model

$$\begin{aligned} y_{i,c} = &\alpha + \phi_1(Boy_i \times T_c \times Low_i) + \theta_1(Boy_i \times T_c \times High_i) + \\ &\gamma_1(Boy_i \times Low_i) + \phi_2(Girl_i \times T_c \times Low_i) + \\ &\theta_2(Girl_i \times T_c \times High_i) + \gamma_2(Girl_i \times Low_i) + \\ &\delta Girl_i + \boldsymbol{\beta X_i} + \epsilon_{i,c} \end{aligned} \tag{3}$$

In (3) we follow the approach used in (1) to estimate the average treatment effect on the treated boys and girls separately, but further decomposing the estimate by splitting by socioeconomic status. We report the results from this estimation in tables 3 and 4, using education and income respectively to denote the socioeconomic status (SES) of the household. For the former we take the average of the mother and father's years of education, and define households as high or low SES using a median split. Similarly for the latter we construct household income by averaging across parents, and then splitting the resulting income distribution at the median. As such, *Low* and *High* are indicators taking the value 1 if the household is below (above) the median in terms of education or income.

We find limited evidence that the gender-specific treatment effects vary across socioeconomic background. While there are some discrepancies in the specific skill domains, notably the *executive function* measure, there are no indication overall that the benefits of our intervention vary by parental background.[12] The results are similar using both education and

[12]This result is consistent with a similar analysis in Rege et al. (2021) where the authors investigate whether the treatment effect varies by parental background for the sample as whole, and find no evidence of such heterogeneity.

income to denote the socioeconomic status of the households, and imply that the heterogeneity in treatment effects by baseline skills is not merely picking up differences across children of different backgrounds. Rather, it suggests that low-achieving boys in particular are benefiting more from our intervention regardless of social class. This result is consistent with a similar analysis in Rege et al. (2020) where the authors investigate whether the treatment effect varies by parental background for the sample as whole, and find no evidence of such heterogeneity.

Table C.1—SES Heterogeneity in Treatment Effects – Education

	Sum Score	Math	EF	Language
Males				
Low Education x Treat	0.207	0.352	0.042	0.106
	(0.146)	(0.109)	(0.157)	(0.175)
High Education x Treat	0.256	0.320	0.210	0.102
	(0.089)	(0.129)	(0.173)	(0.103)
Females				
Low Education x Treat	-0.010	0.081	-0.085	-0.021
	(0.165)	(0.211)	(0.128)	(0.146)
High Education x Treat	0.050	0.146	0.028	-0.047
	(0.069)	(0.095)	(0.055)	(0.115)
N	661	661	660	659
Adj. R^2	0.55	0.40	0.41	0.52

Note: Each column in each panel presents the regression coefficient of treated (standard error) interacted with gender and an indicator for high or low SES using ordinary least squares. The coefficient for the interaction gives the total treatment effect for that gender-SES pairing, so that the difference between them represents the marginal effect with regards to SES. We denote SES status by computing the average years of education for the parents of each child, and then split the sample at the median. Each column represents one regression. In each of the models we follow the specification presented in (4) and regress the outcome on the treatment-gender-SES interactions, controlling for gender×SES, baseline test scores, gender, birth month, parental characteristics (mother and father's years of schooling, earnings, an indicator for non-Western country of birth), and indicators for late consent and not having participated in the T1 assessment. The analysis is based on T3 data. All regressions are clustered on, and control for, randomization block. $^{+}$ p<0.1, * p<0.05, ** p<0.01.

Table C.2—SES Heterogeneity in Treatment Effects – Income

	Sum Score	Math	EF	Language
Males				
Low Income x Treat	0.240	0.368	0.071	0.138
	(0.138)	(0.086)	(0.154)	(0.134)
High Income x Treat	0.233	0.304	0.189	0.183
	(0.149)	(0.150)	(0.163)	(0.127)
Females				
Low Income x Treat	0.021	0.116	-0.083	0.018
	(0.122)	(0.135)	(0.128)	(0.114)
High Income x Treat	0.041	0.1131	0.043	-0.072
	(0.102)	(0.148)	(0.080)	(0.124)
N	661	661	660	659
Adj. R^2	0.55	0.40	0.41	0.52

Note: Each column in each panel presents the regression coefficient of treated (standard error) interacted with gender and an indicator for high or low SES using ordinary least squares. The coefficient for the interaction gives the total treatment effect for that gender-SES pairing, so that the difference between them represents the marginal effect with regards to SES. In this table we denote SES status by computing household earnings as the average of mother and fathers income, and then split the sample at the median. Each column represents one regression. In each of the models we follow the specification presented in (4) and regress the outcome on the treatment-gender-SES interactions, controlling for gender×SES, baseline test scores, gender, birth month, parental characteristics (mother and father's years of schooling, earnings, an indicator for non-Western country of birth), and indicators for late consent and not having participated in the T1 assessment. The analysis is based on T3 data. All regressions are clustered on, and control for, randomization block. $^{+}$ $p<0.1$, * $p<0.05$, ** $p<0.01$.

Appendix D:

Additional tables and specifications

Table D.1—Balance Test for T2 Analytical Sample

	Boys			Girls		
	Control	Treat	Difference	Control	Treat	Difference
Child characteristics						
Birth Month	6.277	6.296	-0.070	5.987	6.081	0.128
	(3.187)	(3.290)	(0.355)	(3.185)	(3.135)	(0.265)
Immigrant	0.108	0.161	0.049	0.133	0.222	0.080
	(0.311)	(0.369)	(0.040)	(0.341)	(0.417)	(0.062)
Mother Education	14.314	14.150	-0.107	14.436	14.175	-0.150
	(2.566)	(2.575)	(0.265)	(2.621)	(2.610)	(0.328)
Father Education	13.910	13.645	-0.276	13.728	13.756	-0.053
	(2.509)	(2.520)	(0.318)	(2.619)	(2.492)	(0.295)
Mother Earnings	347,142	329,301	-19,694	336,093	310,234	-27,654
	(225,887)	(200,712)	(30,109)	(218,945)	(212,404)	(33,959)
Father Earnings	562,409	556,487	-4,658	528,667	560,976	37,772
	(266,498)	(257,256)	(28,644)	(252,128)	(278,473)	(29,722)
Baseline Scores						
T1 Sum Score	-0.131	-0.015	0.016	0.166	0.033	-0.103
	(1.058)	(0.979)	(0.1377)	(0.911)	(1.034)	(0.073)
T1 Math	-0.121	-0.084	0.003	0.133	0.063	-0.060
	(1.012)	(1.042)	(0.119)	(0.898)	(1.045)	(0.110)
T1 EF	-0.103	-0.055	0.026	0.069	0.079	0.039
	(1.058)	(0.976)	(0.139)	(0.958)	(1.020)	(0.094)
T1 Language	-0.093	-0.032	0.008	0.200	-0.062	-0.226
	(0.963)	(0.933)	(0.118)	(1.027)	(1.025)	(0.140)
Missing T1 Scores	0.064	0.038	-0.028	0.040	0.029	-0.014
	(0.245)	(0.191)	(0.025)	(0.196)	(0.168)	(0.024)
N	141	186	327	151	172	323

Note: The columns provide mean (standard deviation) for child characteristics and T1 test scores separately by gender and treatment status for the T3 analytic sample. The column labeled *Difference* is the estimated coefficient (standard error) from regressing each covariate against treatment status, while controlling for randomization block. Regressions are also clustered on the block level.

Table D.2—Attrition

	With Late Consenters			*Without Late Consent*		
	T1	T2	T3	T1	T2	T3
Treated × Boy	0.023	-0.006	0.027	0.021	-0.000	0.026
	(0.031)	(0.018)	(0.027)	(0.033)	(0.024)	(0.029)
Treated × Girl	0.013	0.001	0.012	0.000	0.011	0.008
	(0.025)	(0.015)	(0.022)	(0.029)	(0.016)	(0.023)
Female	0.034	0.007	0.002	0.042	0.013	-0.005
	(0.030)	(0.018)	(0.025)	(0.034)	(0.016)	(0.026)
Difference	-0.010	0.007	-0.014	0.021	-0.000	0.026
	(0.031)	(0.021)	(0.030)	(0.033)	(0.024)	(0.029)
N	691	691	691	561	561	561
Adj. R^2	0.03	0.02	0.05	0.04	0.02	0.05

Note: In this table we present results from regressing indicators for participating in assessments at T1, T2 and T3 respectively on treatment status, gender and their interaction. We use a specification akin to (2) in section 5.3. In the row labeled *Difference* we test for significant differences in the estimates for the treatment-gender interactions. In the right-most panel we repeat the estimation, but exclude all late consenters from the sample. All regressions are clustered on, and control for, randomization block. $^{+}$ $p<0.1$, * $p<0.05$, ** $p<0.01$.

Table D.3—Main Results: Excluding Baseline Scores

	Post-Intervention (T2)				*Follow-Up (T3)*			
	Sum Score	Math	EF	Language	Sum Score	Math	EF	Language
Treatment Effect	0.240*	0.222	0.238*	0.120	0.281+	0.350*	0.170	0.165
Boys	(0.105)	(0.142)	(0.091)	(0.103)	(0.133)	(0.119)	(0.138)	(0.124)
Treatment Effect	0.001	0.068	0.042	-0.110	-0.037	0.059	-0.035	-0.114
Girls	(0.074)	(0.111)	(0.086)	(0.096)	(0.097)	(0.132)	(0.074)	(0.105)
Difference	-0.240*	-0.087	-0.197	-0.230	-0.319*	-0.291	-0.205	-0.279*
	(0.105)	(0.141)	(0.102)	(0.131)	(0.137)	(0.177)	(0.160)	(0.115)
p-value	0.039	0.143	0.130	0.101	0.036	0.123	0.220	0.030
Wild Cluster	0.066	0.291	0.145	0.106	0.052	0.141	0.234	0.043
RI	0.094	0.282	0.179	0.158	0.032	0.059	0.170	0.067
N	652	650	652	648	661	661	660	659
Adj. R^2	0.19	0.16	0.15	0.14	0.17	0.13	0.13	0.22

Note: In this table we replicate the results from Table 2 in the main text using a specification where we exclude baseline test scores from the controls. Each column in each panel presents the regression coefficient of treated (standard error) interacted with gender using ordinary least squares. The coefficient for the interaction gives the total treatment effect on that gender, so that the difference between them represents the marginal effect. The *Difference* panel reports coefficients and errors for tests on significant differences between the gender specific estimates. Below we report three sets of p-values computed using clustering on block, the Wild T bootstrap procedure, and randomization inference respectively. For both assessment periods we regress outcome on the treatment-gender interaction, controlling for gender, birth month, parental characteristics (mother and father's education level, earnings, an indicator for non-Western country of birth), and indicators for late consent and not having participated in the T1 assessment. All regressions are clustered on, and control for, randomization block. + p<0.1, * p<0.05, ** p<0.01.

Table D.4—Main Results: No covariates

	Post-Intervention (T2)				*Follow-Up (T3)*			
	Sum Score	Math	EF	Language	Sum Score	Math	EF	Language
Treatment Effect	0.230^{+}	0.242	0.223^{+}	0.087	0.276^{+}	0.357*	0.183	0.128
Boys	(0.124)	(0.147)	(0.108)	(0.116)	(0.138)	(0.124)	(0.137)	(0.123)
Treatment Effect	-0.042	0.091	-0.005	-0.191	-0.086	0.061	-0.041	-0.230
Girls	(0.096)	(0.093)	(0.112)	(0.122)	(0.102)	(0.123)	(0.077)	(0.136)
Difference	-0.272*	-0.151	-0.229	-0.277^{+}	-0.362*	-0.296	-0.224	-0.358**
	(0.104)	(0.120)	(0.138)	(0.134)	(0.134)	(0.173)	(0.170)	(0.111)
p-value	0.021	0.230	0.119	0.058	0.017	0.110	0.209	0.006
Wild Cluster	0.029	0.226	0.149	0.066	0.024	0.130	0.219	0.008
RI	0.056	0.309	0.121	0.085	0.021	0.052	0.152	0.027
N	652	650	652	648	661	661	660	659
Adj. R^2	0.05	0.03	0.05	0.05	0.06	0.06	0.06	0.08

Note: In this table we replicate the results from Table 2 in the main text using a specification where we include no additional controls outside of randomization block fixed effects. Each column in each panel presents the regression coefficient of treated (standard error) interacted with gender using ordinary least squares. The coefficient for the interaction gives the total treatment effect on that gender, so that the difference between them represents the marginal effect. The *Difference* panel reports coefficients and errors for tests on significant differences between the gender specific estimates. Below we report three sets of p-values computed using clustering on block, the Wild T bootstrap procedure, and randomization inference respectively. All regressions are clustered on, and control for, randomization block. $^{+}$ p<0.1, * p<0.05, ** p<0.01.

Table D.5—Main Results: Excluding Late Consenters

	Post-Intervention (T2)				*Follow-Up (T3)*			
	Sum Score	Math	EF	Language	Sum Score	Math	EF	Language
Treatment Effect	0.132	0.093	0.161*	0.057	0.195+	0.291*	0.097	0.091
Boys	(0.084)	(0.120)	(0.075)	(0.092)	(0.094)	(0.099)	(0.123)	(0.084)
Treatment Effect	0.001	0.094	-0.029	-0.066	0.040	0.133	-0.010	-0.025
Girls	(0.078)	(0.096)	(0.083)	(0.092)	(0.092)	(0.135)	(0.053)	(0.095)
Difference	-0.131	0.001	-0.190	-0.123	-0.156	-0.158	-0.106	-0.115
	(0.090)	(0.119)	(0.109)	(0.113)	(0.099)	(0.177)	(0.135)	(0.072)
p-value	0.169	0.994	0.103	0.294	0.139	0.386	0.444	0.130
Wild Cluster	0.154	0.994	0.119	0.256	0.157	0.421	0.440	0.115
RI	0.137	0.994	0.144	0.367	0.220	0.246	0.431	0.397
N	532	530	532	528	534	534	533	532
Adj. R^2	0.63	0.44	0.54	0.52	0.55	0.40	0.42	0.50

Note: In this table we replicate the results from Table 2 in the main text using a specification where we exclude all observations of children whose parents submitted their consent sheet after the deadline (130 observations, 18.8 percent of the gross sample). Each column in each panel presents the regression coefficient of treated (standard error) interacted with gender using ordinary least squares. The coefficient for the interaction gives the total treatment effect on that gender, so that the difference between them represents the marginal effect. The *Difference* panel reports coefficients and errors for tests on significant differences between the gender specific estimates. Below we report three sets of p-values computed using clustering on block, the Wild T bootstrap procedure, and randomization inference respectively. All regressions are clustered on, and control for, randomization block. $^{+}$ $p<0.1$, * $p<0.05$, ** $p<0.01$.

Table D.6—Main Results:
Excluding Items 3, 11 and 17 From The *ABMT*

	Post-Intervention (T2)		*Follow-Up (T3)*	
	Sum Score	Math	Sum Score	Math
Treatment Effect	0.190*	0.194	0.208+	0.263*
Boys	(0.084)	(0.118)	(0.103)	(0.105)
Treatment Effect	0.035	0.102	0.010	0.101
Girls	(0.065)	(0.102)	(0.095)	(0.109)
Difference	-0.155+	-0.092	-0.198	-0.162
	(0.080)	(0.119)	(0.126)	(0.164)
p-value	0.072	0.451	0.137	0.342
Wild Cluster RI	0.089	0.453	0.164	0.377
N	652	650	661	661
Adj. R^2	0.61	0.39	0.52	0.28

Note: In this table we replicate the results from Table 2 in the main text using a specification where we exclude the three items of the Ani Banani Math Test that ten Braak and Størksen (2021) reported showed signs of gender bias. We only report results for those measure where the ABMT is included. Each column in each panel presents the regression coefficient of treated (standard error) interacted with gender using ordinary least squares. The coefficient for the interaction gives the total treatment effect on that gender, so that the difference between them represents the marginal effect. The *Difference* panel reports coefficients and errors for tests on significant differences between the gender specific estimates. Below we report three sets of p-values computed using clustering on block, the Wild T bootstrap procedure, and randomization inference respectively. All regressions are clustered on, and control for, randomization block. + p<0.1, * p<0.05, ** p<0.01.

Table D.7—Main Results: Excluding Imputed Pre-scores

	Post-Intervention (T2)				*Follow-Up (T3)*			
	Sum Score	Math	EF	Language	Sum Score	Math	EF	Language
Treatment Effect	0.197*	0.210	0.212**	0.050	0.247*	0.320**	0.173	0.112
Boys	(0.077)	(0.125)	(0.068)	(0.085)	(0.095)	(0.083)	(0.116)	(0.099)
Treatment Effect	0.032	0.091	0.013	-0.055	0.019	0.093	-0.035	-0.011
Girls	(0.072)	(0.091)	(0.082)	(0.097)	(0.091)	(0.112)	(0.053)	(0.112)
Difference	-0.175*	-0.119	-0.199+	-0.105	-0.228+	-0.227	-0.208	-0.123
	(0.080)	(0.109)	(0.102)	(0.106)	(0.115)	(0.153)	(0.134)	(0.0096)
p-value	0.046	0.293	0.071	0.341	0.067	0.160	0.143	0.219
Wild Cluster	0.069	0.310	0.084	0.340	0.076	0.190	0.148	0.222
RI	0.058	0.271	0.074	0.401	0.048	0.073	0.096	0.338
N	625	623	625	621	629	629	628	627
Adj. R^2	0.65	0.48	0.53	0.56	0.57	0.41	0.43	0.53

Note: In this table we replicate the results from Table 2 in the main text using a specification where we exclude all observations not assessed at T1, for which we imputed the pre-scores (33 observations, 4.8 percent of the gross sample). Each column in each panel presents the regression coefficient of treated (standard error) interacted with gender using ordinary least squares. The coefficient for the interaction gives the total treatment effect on that gender, so that the difference between them represents the marginal effect. The *Difference* panel reports coefficients and errors for tests on significant differences between the gender specific estimates. Below we report three sets of p-values computed using clustering on block, the Wild T bootstrap procedure, and randomization inference respectively. All regressions are clustered on, and control for, randomization block. $^{+}$ p<0.1, * p<0.05, ** p<0.01.

Chapter 4 – Essay III

Alumni Satisfaction, Rankings, and College Recommendations

Eric Bettinger[1*] and Andreas Fidjeland[2]

Abstract

While college-access organizations as well as state and federal governments invest millions of dollars in informational campaigns, rankings, and college websites, students often rely more on parents and other trusted adults for information on college options. We use new data from a nationally representative survey of adults to identify the extent to which those who attended college are willing to recommend their own university experience. We demonstrate that alumni place more value on their personal experiences than on prominent rankings in formulating recommendations. Interestingly, individuals with low economic outcomes also express high levels of satisfaction and willingness to recommend their own college experience to others. We discuss the implications of such strong preferences and the potential of satisfaction as an additional metric for college evaluation.

JEL Codes: I2, I23, I26, I28, D83, L15

Keywords: Student satisfaction, college choice, returns to education

[*]Corresponding author. We are thankful to the Strada Education Network for graciously sharing their data. Edwin Leuven, Hans H. Sievertsen, and participants at the University of Stavanger Quantitative Forum and the EEA 2020 annual congress have provided valuable feedback. Fidjeland acknowledges funding from the Norwegian Research Council, grant numbers 270703 and 290675. The views expressed in this paper are those of the authors and do not necessarily represent the views of the Strada Education Network.

[1] Stanford University, ebetting@stanford.edu
[2] University of Stavanger, andreas.fidjeland@uis.no

1 Introduction

Colleges are the primary vehicle for intergenerational changes in inequality. Not only do college graduates earn 67 percent more than high school graduates, they are also only half as likely to be unemployed (Ma et al., 2016). Moreover, low-income students who make it into highly selective colleges have similar economic outcomes to students from more affluent backgrounds in the same schools (Chetty et al., 2017, 2020).

Traditional human capital models suggest that students might rationally choose to attend college based on their expected costs and benefits. However, economists have repeatedly demonstrated that students' expectations differ dramatically from the truth, both in that they overestimate the cost of tuition (Avery and Kane, 2004; Horn et al., 2003) and the barriers to securing financial aid (Bettinger et al., 2012) and in that they underestimate the market returns to education (Jensen, 2010; Wiswall and Zafar, 2015). This lack of accurate expectations has motivated many large-scale experiments focusing on providing additional information and application assistance to students, but the results of such interventions have been disappointing at best (Barone et al., 2017; Bergman et al., 2019; Bird et al., 2021; Carrell and Sacerdote, 2017; Cunha et al., 2018; Gurantz et al., 2021; Hyman, 2020; Kerr et al., 2020; McGuigan et al., 2016).

Most of the interventions to date focus on institutions, governments, and/or high school officials providing accurate information to students. However, survey data suggest that students are more likely to talk to their parents than to seek out other sources (Otto, 2000; Oymak, 2018). While high school counselors play a part, students generally turn to parents and peers before professionals (Carrell and Sacerdote, 2017; Mulhern, 2020). Indeed, information and nudge-style interventions seem to be more effective at improving educational outcomes when targeting parents rather than children (Oreopoulos, 2020). The importance of private information held by family members about college options is also underscored by the fact that several studies find striking similarities in the enrollment pat-

terns of siblings (Aguirre and Matta, 2021; Altmejd et al., 2021; Goodman et al., 2015). Given parents' importance as advisors, private information held by those of them who themselves attended college might be particularly salient for student decision-making. In this paper, we attempt to understand the information and perceptions that adults have about colleges. We introduce a new data set which focuses on how alumni view their college experiences. To our knowledge, only one other paper has used these data (Rothwell, 2019).

Our data come from a nationally representative survey of the US working population initiated by the Strada Education Network in collaboration with Gallup. Their newly established Education Consumer Survey provides a sample that includes rich information about individual labor market outcomes and experiences with college from about 323,000 respondents. Importantly, our data include information about individuals' satisfaction with college and measures of their willingness to recommend their college to others.

We find that the majority of alumni are very satisfied with the education they received and report a high willingness to recommend others to follow the same educational path that they took. This is true across a wide spectrum of individual and institutional characteristics. For example, we find little evidence that satisfaction is correlated with subsequent labor market success. In fact, individuals with seemingly "bad" economic outcomes from a particular institution also express high levels of satisfaction and willingness to recommend their college to others. Further, we demonstrate that satisfaction-based measures of college quality are correlated with existing college ratings but that they predict the willingness to recommend to others with greater power than any other traditional ranking, while many other quality metrics are found to have little predictive power at all.

Traditional interventions often rely on high school counselors transmitting accurate information on college options to students. However, if parents are their primary advisors, such information might have limited effect on college choices. And if more objective data about college quality is less salient for parents, it is valuable to know on what information they

base their recommendations instead. While our survey data cannot observe the recommendations alumni might give directly — and while they are subject to the usual concerns regarding accuracy — our results do provide evidence consistent with the notion that parents place greater emphasis on their subjective experience from college, which is cognitively more readily available, than on more objective measures and rankings (see, e.g., the availability heuristic [Tversky and Kahneman, 1973]). In particular, satisfaction proves to be the strongest predictor by far of alumni's willingness to recommend their alma mater to others. Our results also underscore the salience of private information in the college decision-making process (Altmejd et al., 2021). To that effect, the satisfaction construct might represent a broader array of variables that are important to potential students, such as the consumption value of college attendance, affordability, social fit, and safety — to name only a few. While satisfaction may not be a good proxy for all of policymakers' goals, it might provide some insight into students' attendance patterns. For college administrators, our results provide insights that might be relevant for improving the recruitment of future students, enhancing the retention of current students, and boosting donations from former students.

2 Satisfaction: A Basis for Recommendation

A conventional human capital model in the tradition of Becker (1962) posits that individuals rationally choose if and where to go to college based on the expected costs and returns of obtaining higher education. It follows logically from such models that people choose to enroll in college only if they perceive that the net benefits exceed those of other options, From the point of view of the economic scholar, this is an investment decision problem. However, the extant literature indicates that the information on the basis of which students make their college decisions is at best incomplete or insufficient, which often result in mismatches between student and college quality (Campbell et al., 2021; Dillon and Smith, 2017; Hoxby

and Turner, 2015; Mabel et al., 2020). For example, compared with more affluent students with a similar academic profile, high-achieving students in low-income and rural areas are more likely to apply to less selective colleges or to fail to apply to college at all, despite the fact that in many cases students would actually pay less if they enrolled in a selective college, because of generous financial aid (Hoxby and Avery, 2013).

A common reaction by both scholars and policymakers to these empirical findings is to highlight the need for more and better information. While academics have tried to create more rigorous measures of college quality (e.g., Chetty et al., 2017; Hoxby, 2019; Smith et al., 2017), popular rankings such as the US News and World Report Ranking, Forbes Top Colleges, London Times Educational Supplement, and the Shanghai Rankings have captivated both universities and students. Further, government agencies in the United States have increasingly published data on outcomes for colleges, with perhaps the most notable effort coming from the White House College Scorecards. Each of these indices is built from a comprehensive set of outcomes ranging from school quality (e.g., faculty and facility characteristics) over graduate outcomes (e.g., earnings and job placement) to admissions practices (e.g., selectivity and yield). As an example, the US Government claimed at the launch of its College Scorecard that this would "provide [counselors, parents and students] with the information they found *most valuable* in making decisions about where to enroll" (U.S. Department of Education, 2013, italics added)

Despite these efforts, there is limited evidence that the information provided by such rankings and databases influences the college decision-making process (Hurwitz and Smith, 2018; Mabel et al., 2020). While some experimental evidence suggests that students may be receptive to expert information (e.g., Hoxby and Turner, 2015), they frequently rely on information and advice from trusted adults, such as parents, who may lack expertise (Oymak, 2018). In turn, parents who themselves attended college might hold private information and preferences with regard to college options that prospective students use to inform their decisions. However, if these trusted adults rely on different, more subjective, or imperfect information compared with that available from professional sources in for-

mulating their advice, this channel may provide a plausible explanation for why informational interventions aimed at students often do not succeed, and for why mismatches continue to occur despite the many efforts made to reduce information asymmetries and barriers.

If students mostly ignore the ostensibly objective measures of college quality in choosing between institutions, what then are they using to formulate expectations? If indeed they are relying on advice from parents and other adults not professionally associated with schools, then what do those adults draw upon when formulating their perceptions about the benefits and costs of colleges?

To answer this question, we use new data on alumni satisfaction. These satisfaction indices may be more suited for understanding how adults think about their willingness to recommend their educational path than the more objective external metrics. While we do not have a specific definition of satisfaction, it likely encapsulates a broader set of outcomes and experiences, including not only the college's perceived impact on outcomes, but also the consumptive value of attending. For policymakers, satisfaction metrics may not only provide additional information about colleges beyond traditional rankings, but also provide greater insight into the experiences and perceptions of adults who might be giving important advice. However, depending on the outcomes that policymakers might be hoping to maximize, it is unclear whether they would prefer students to place more or less weight on such subjective advice than they do at present. We discuss the policy implications in the conclusion of the paper.

To demonstrate that alumni base their college recommendations on subjective satisfaction, we rely on several established measures of college quality, as well as on new survey data that include self-reported evaluations of the college experience. Those new data come from the recently established Strada–Gallup Education Consumer Survey (ECS), an annual, nationally representative survey of the US labor force aimed at tracking consumer satisfaction with postsecondary education. In the ECS, respondents rate their college experience across a variety of dimensions. Respondents also detail their subsequent labor market outcomes and describe to what extent they attribute these outcomes to the schooling they received.

The survey and its features were first described in Rothwell (2019), in which the author assessed its validity for measuring subjective satisfaction and wellbeing.

2.1 Measures of satisfaction

We base our general satisfaction index on six survey items pertaining to students' assessment of the quality of the education they received at the institution where they obtained their highest level degree. These items, listed in Table 1, ask respondents to indicate on a five-point scale the degree to which they agree with a statement about their college experience (from "Strongly disagree" to "Strongly agree"). To construct our satisfaction index, we standardize the responses for each item to a mean of zero and a standard deviation of one, before averaging across all items for each individual and standardizing again. Hence the index is scaled in a manner that gives it a natural interpretation in comparison with other measures.[1]

To mitigate response bias and potential endogeneity in outcomes of interest, we further construct a leave-out mean of the average satisfaction among the peers of a particular respondent. Respondents' peers are primarily identified as those graduating from the same institution in the same year. However, since the existence and number of peers vary greatly between institutions and cohorts we include all students graduating from the same institution as a given respondent within $+/-$ 5 years of his or her graduation year in our peer satisfaction measure. Even so, we anchor the measure by restricting our analysis to cases where we can observe at least one peer in the same graduation year as the respondent. For any given individual, this leave-out mean gives the average satisfaction of all the other alumni present in the 10-year window, which we then standardize. [2]

[1] Cronbach's α for the items in the satisfaction index is 0.86.

[2] Because the respondents are distributed over a large time window with regard to graduation year, the number of observations per institution per year is in many cases too small to yield reliable measures of a pure institution-by-cohort leave-out mean. We therefore resort to using the 10-year moving leave-out mean described above to approximate peer satisfaction. We construct the leave-out-mean peer measure by summing the satisfaction scores for all peers, subtracting the satisfaction of the individual in

Table 1—Survey Items Used in the Paper

Satisfaction Index

1. You received a high-quality education
2. You would not be where you are today without your education
3. You learned important skills in your college courses that you use in day-to-day life
4. The coursework you took is directly relevant to what you do at work
5. Your education experience make you attractive to potential employers
6. Your education was worth the cost

Willingness to Recommend

1. You would recommend the educational path you took to others like you

Attribution to College

How helpful have each of the following been to you so far in your career?

1. The highest level of education you received at your terminal institution
2. The field you studied at your terminal institution
3. The people you met through your college during your studies
4. The reputation of the institution where you got your highest level of education
5. The courses you took during your studies

Note: The table lists all the items used in the analysis in this paper. The satisfaction and willingness-to-recommend items are measured on a five-point scale ranging from "Strongly disagree" to "Strongly agree". The *attribution-to-college* items are measured on a four-point scale.

We consider a variety of outcomes and how they relate to alumni satisfaction. For our main research questions we explore the likelihood that respondents will recommend their educational path to others, such as their children. In this context, we pay particular attention to the survey item involving the statement "You would recommend the educational path you took to other people like you," which is answered using on a five-point scale from "Very unlikely" to "Very likely."

question, and then divide on the n-1 observations obtained for that institution in that graduation window. In models not included in the paper, we also tried constructing a similar peer measure using all available observations from a given institution, regardless of graduation year. Using this approach did not yield substantively different results, but it strikes us as an intuitively less reasonable definition of "peers," given the wide

2.2 Additional measures

We compile institution-level characteristics from three sources. First, we use the unique college identifiers collected by Gallup to link the respondents' colleges to the Integrated Postsecondary Education Data System (IPEDS) database and the College Scorecard. From these, we measure an array of covariates such as the composition of the student body, enrollment and completion rates, school profile and choices of majors, and proxies for structural quality such as instructional expenditures, faculty salary, and more. However, we note that many of the covariates in the IPEDS and College Scorecard are based on undergraduate cohorts while our satisfaction measure tracks the last institution attended, which might be graduate school. Still, to the extent that these measures serve as proxies for overall characteristics of a college, we use the same measures for respondents who have a postgraduate education. We collect the data at an institution-year level for all the years for which a given variable is available. All monetary values are adjusted to 2016 levels, before we average over all years by institution. The resulting averages are then standardized and matched to the survey respondents using the unique college identifier. We are able to match 99.6% of the sample to their college's covariates. For certain analyses we combine a large subset of these characteristics in an index we refer to as *Structural quality* by standardizing each variable, taking a simple average, and re-standardizing the resulting composite.[3]

Additionally, we make use of the public data provided by Chetty et al. (2017) on college productivity. Following its publication, their index for social mobility has been advocated as a candidate for measuring college quality (e.g., Barr and Castleman, 2018). In particular, we use their

span in graduation year.

[3]We include the following covariates: admission rate, completion rate, retention rate for undergraduates, average SAT score, size of the undergraduate cohort, fraction of full -time faculty, average faculty salary, fraction of first-generation students, median family income, fraction of students receiving Pell grants, fraction of students receiving student loans, tuition for in- and out-of-state students, net cost for low-income students, average total cost of attendance, instructional expenditure per full-time student equivalent, median graduate debt, and indicators for whether a school is a public institution and a four-year institution.

measure of college-specific mobility rates — measured as the fraction of students entering the college from the lowest quintile of the income distribution who subsequently end up in the highest quintile — as an alternative approach to our satisfaction index. In addition, we consider their "1% mobility," which measures the fraction of low-income students who end up in the top 1 percent of the income distribution. We use their rating as a contrasting predictor of alumni satisfaction and willingness to recommend one's college. From the Chetty et al. (2017) database we also collect information about the selectivity of each college, measured using Barron's Selectivity index, which uses a six-point scale ranging from "Most selective" to "Nonselective". In models where we make use of Barron's index we also include institutions that are unranked, in a category marked "Not ranked".

Finally, we consider colleges' position on the Forbes Top Colleges ranking. We use the 2018 edition, which ranks 650 colleges and universities according to a broad set of metrics including postgraduation salaries and debt, retention and graduation rates as well as "signs of individual success including academic and career accolades"(Coudriet, 2018).[4] As with the Chetty et al. mobility measures, we make use of the Forbes ranking as an example of established measures of college quality to contrast with our satisfaction index when it comes to predicting alumni's willingness to recommend their educational path and alma mater to others.

3 Data and Analytic Framework

Gallup collected the data on a rolling basis between June 2016 and January 2019, producing a representative sample of the US labor force consisting of 323,218 surveyed individuals aged 18–65 years. Each individual was asked to answer a comprehensive set of questions about their college experiences and outcomes and their labor market outcomes as well as

[4]The full list is available at https://www.forbes.com/top-colleges. The list we use in our analysis was retrieved from that address on June 14, 2019.

to provide information about their demographic and socioeconomic status and background. Throughout our analysis, we use this background information to characterize subsamples. For labor market outcomes we consider two separate indicators. First, we consider the respondents' employment status measured with a categorical variable indicating whether they are employed full-time, self-employed full-time, part-time employed, unemployed, or not in the workforce. Second, we use a five-point categorical variable for level of income to designate where respondents rank in the income distribution. Unfortunately, since the ECS asks their interviewees only about current income, we do not have access to the respondents' wage profiles. For socioeconomic background, we define *low parental education* as having a mother who did not complete high school.

3.1 Sample

We base our analysis sample on the 183,049 respondents who enrolled in college at some point and can be linked to their institution through the unique college identifier. At its core, the ECS asks how the respondents would evaluate their college experience. In line with that, the inaugural survey in 2016 contains information only about college graduates. However, subsequent years also include noncompleters and individuals who never attended college. Because of the focus of the survey, however, respondents who never enrolled in college were asked only a limited number of items. For this reason, we focus our analysis — with the results having to be interpreted through that lens — on the reasoning behind, and experiences of, going to college, conditional upon having enrolled (but not necessarily completed) in the first place.

The College Subsample

Starting from the sample of college enrollees, we impose some necessary sample restrictions. To begin with, among the initial 183,049 observations, 10,308 are still in school and thus excluded. We define individuals as still being in school if their projected year of graduation is later than the year in which they were interviewed or if they report that they are still at-

tending college courses full-time. Further, we exclude five observations that are inconsistent in that the individuals concerned claim to have attended a college but do not acknowledge in other survey questions that they completed even "some college." Finally, another 364 observations are either missing all the satisfaction items, or the crucial item asking them about their willingness to recommend their educational path to others. We exclude these, as well as those who are the sole observation from their institution and for whom it is thus not possible to construct a leave-out-mean satisfaction measure. Imposing these restrictions leaves an analysis sample of 171,317 observations.

As we noted above, an important limitation of the data is that we observe only students' terminal institution. For most (73.4%), this is their undergraduate institution, but for others it represents graduate school.

Match With Other Data Sources

Of our analysis sample of 171,317 observations, 155,619 (90.8%) can be matched to the mobility-rate measures retrieved from the Chetty et al. (2017) open database. We observe the Barron's Selectivity rank for 155,487 observations. Further, we successfully match the set of covariates obtained from the College Scorecard and IPEDS data bases to 170,619 (99.6%) observations. A total of 88,936 individuals (51.9%) attended a school ranked in the 2018 Forbes Top Colleges list. We code the rank in increasing order, with the value 1 representing the 650th placed college and 650 representing the top-ranked one. Unranked institutions are given a rank of zero, and we include a dummy indicating whether or not the institution was ranked at all.

3.2 Summary Statistics

The ECS is a nationally representative survey, designed to mirror the US working population. In Table 2 we present summary statistics on key variables for both the full sample retrieved from Strada and our analysis sample. As the latter is conditioned upon the respondent being linked to a postsecondary institution, it is not representative in the above sense

but rather constitutes a subsample of individuals who at least attended college for some time. Although only two-thirds of all ECS respondents attended some college, in demographic terms the analysis sample is not markedly different from the full sample, albeit slightly more likely to be White, and coming from higher socioeconomic backgrounds (measured by parental education). However, the restricted sample is obviously more educated and hence has higher incomes and a higher likelihood of being employed. Still, only 65 percent of college attendees report that they are employed full-time. In terms of education, the average respondent in the analysis sample has a four-year college degree and attended a selective school ranked in the lower third of the Forbes college rankings, with a Chetty et al. mobility rate of 2%. Individuals and their peers are very satisfied with their education and report a high willingness to recommend their educational path to others. They attribute a substantial part of their labor market outcomes to the education they received at their terminal institution.

3.3 Analytic Framework

Our goal is to demonstrate how adults might recommend colleges. In particular, we demonstrate that college rankings and other government-backed scorecards provide less of the basis for recommendations than individuals' subjective satisfaction with their own college experience. Loosely speaking, we hope to conduct a "horse race" between existing metrics of school quality and the ECS's more subjective measure of satisfaction.

To do this, we conduct two complementary activities. As a first step, we focus on individuals' willingness to recommend their own experience for our main analysis. Here we run simple predictive models to show what measures predict individuals' willingness to recommend. We argue that the relative effect sizes of the metrics in explaining the willingness to recommend provide a good indication of what the respondents in the samples base such recommendations on. The results suggest that satisfaction has the most predictive power, far beyond any alternative metric. In fact, most other metrics seem to have little predictive power at all. We show

the robustness of these results by conducting similar analyses on a battery of subpopulations where school satisfaction could and should conceivably be lower. The superiority of our satisfaction indices as a predictor of the willingness to recommend proves to be remarkably consistent across all our models.

As a second step, we investigate what might cause alumni to be satisfied. First we explore satisfaction levels across a variety of labor market outcomes and school characteristics. In doing so, we demonstrate that alumni satisfaction is remarkably high and stable across the covariates that we consider. Second we discuss the respondents' self-reported reasoning for pursuing and choosing a college education. Third, we discuss the overall correlation between our satisfaction measure and existing measures of college quality. We demonstrate that satisfaction metrics are in most cases positively correlated with other rankings, and that rankings are stronger predictors of peer satisfaction than of individual satisfaction.

In a supplemental analysis in the appendix, we also assess the validity of satisfaction as an alternative instrument for college quality. Following Rothwell (2019), we use earnings as an example "ruler" to demonstrate that satisfaction indices in many cases have greater predictive power than models based on rankings. The results indicate that, at worst, our indices perform as well as established quality measures in explaining alumni income levels. The fact that our satisfaction metrics are predictive of real-life outcomes adds to their salience as indicators of students' academic experiences and perceptions of the value of college.

It should be emphasized that our goal is not to provide causal evidence on satisfaction. We do not manipulate either satisfaction or rankings in any experimental or quasi-experimental way. Rather, our goal is to demonstrate a positive relationship between the recommendations adults give about college and their subjective satisfaction with their own college decisions. If indeed parents' recommendations matter, and they are based on criteria that are not consistent with policymakers' preferences, then we could see inefficiencies in the clearing of the college marketplace. These satisfaction measures might give some hint as to the source of such inefficiency, but our results should solely be viewed as descriptive.

Table 2—Summary Statistics

	Full Sample		Analysis Sample	
	Mean	SD	Mean	SD
Female	0.46	0.50	0.48	0.50
White	0.76	0.43	0.81	0.39
Black	0.18	0.31	0.10	0.30
Hispanic	0.12	0.32	0.07	0.26
Asian	0.03	0.17	0.03	0.18
Age	45.3	14.1	47.0	12.9
Has children	0.68	0.47	0.69	0.46
Some college	0.28	0.45	0.34	0.47
College graduate	0.24	0.43	0.40	0.49
Postgrad	0.17	0.37	0.27	0.44
Mother's education	3.27	2.37	3.77	2.38
Father's education	3.27	2.60	3.90	2.62
Income (2016 base)	69,101	190,088	82,987	199,423
Employed full time	0.59	0.49	0.65	0.48
Barron's Selectivity Index			2.56	1.96
Forbes Top Schools rank			191	231
Mobility rate			0.02	0.01
Satisfaction			3.92	0.98
Peer satisfaction			3.92	0.42
Willingness to recommend			3.97	1.26
Degree of Attribution to college			2.78	0.83
Observations	323,218		171,317	

Note: In this table we present means and standard deviations of key covariates for the full and analysis samples, respectively. Demographics, education level and employment status represent dummies equal to one if that characteristic is true for the respondent. Age is a running variable measured in years. Mother's and father's education is measured on an eight-point categorical scale, ranging from "Less than high school" to "Postgraduate degree". Income is based on self-reported levels, converted to 2016 values.*Barron's Selectivity Index* takes values from 0 to 6, with 0 being "Not ranked" and 6 being "Elite". The Forbes Top Colleges ranking is coded so that the value increases with rank, and nonranked colleges are set to 0. The mobility rate is the share of students entering from the lowest quintile ending up in the top quintile of the income distribution, as constructed by Chetty et al. (2017). The *Satisfaction*, *Peer satisfaction*, *Willingness to recommend*, and *Degree of attribution to college* measures are calculated using the nonstandardized answers from the items included in the satisfaction indices described above. The latter refers to the extent to which the respondents attribute their current labor market situation to the education they obtained. In all cases, respondents who refused to answer are excluded.

4 Analysis

4.1 College Quality and Willingness to Recommend

Our goal is to understand what are the characteristics of those individuals who are eager to advocate on behalf of their schools as well as the characteristics of the schools that those individuals attended. We first report results in Table 3 from regressing the willingness to recommend (WtR) on the aforementioned college quality metrics. In the "horse race" model in column 1, we find a large and statistically meaningful relationship between subjective satisfaction and willingness to recommend. By contrast, the alternative metrics all return small and substantively insignificant estimates.

In columns 2–8 we regress WtR on each quality metric separately. We see that the coefficient for the satisfaction index does not change in either model. However, unlike in column 1, we find moderately positive associations between the other metrics and WtR in the single-metric models, although the magnitude is at most one-fifth of the association with satisfaction. What is more, the fact that the point estimate for the satisfaction index is substantial in both cases, even when we control for all the other college quality measures, implies that our satisfaction index captures variation that the other metrics do not. This is also evident in the explanatory power exhibited by the different metrics. In the bivariate regressions reported in columns 2–8 we find that the metric explaining on its own the most variation in the willingness to recommend is — by far — the satisfaction index. Indeed, variation in the satisfaction index alone accounts for 30 percent of the variation in the WtR measure. By contrast, the other college-quality metrics hardly explain any variation at all. We consider it interesting that the established metrics appear to be substantially less suited as predictors of WtR, both in terms of the magnitudes of the point estimates and in terms of R-squared, particularly in the horse-race model. For example, we find precise zeros for the Forbes ranking and the main Chetty et al. mobility measures. These results are surprising, given the prominence and emphasis given to some of these other measures in dis-

Table 3—Predicting Willingness to Recommend

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Satisfaction	0.547**	0.547**						
	(0.003)	(0.002)						
Mobility Rate	-0.002		0.013**					
	(0.005)		(0.005)					
1% Mobility Rate	0.008**			0.061**				
	(0.003)			(0.007)				
Mobility Rank	0.006				0.019**			
	(0.005)				(0.005)			
Barron's Selectivity	0.013*					0.113**		
	(0.005)					(0.004)		
Forbes Ranking	0.006						0.101**	
	(0.004)						(0.004)	
Structural Quality	-0.020**							0.114**
	(0.004)							(0.004)
Mean Satisfaction	3.92							
Mean WtR	3.97							
Observations	155,619	171,317	155,619	155,619	155,619	171,317	171,317	171,317
R^2	0.297	0.299	0.000	0.004	0.000	0.013	0.010	0.013

Note: The table reports results from estimating a set of models where willingness to recommend is regressed separately on each college-quality metric. Displayed in columns 2–8 are the resulting standardized point estimates for each metric. Column 1 reports the estimates from a multivariate model including all the metrics. "Mean Satisfaction" and "Mean WtR" refer to average scores for a subsample on the satisfaction index and the willingness-to-recommend item, respectively, both measured in absolute terms on a 1–5 scale (5 highest). Cluster-robust standard errors clustered at the college level in parenthesis. $^{+}$ $p<0.1$, * $p<0.05$, ** $p<0.01$.

cussions about the college decision-making process. In fact, they suggest that a college's ability to produce subjective satisfaction among its alumni by far outperforms more objective rankings when it comes to the likelihood that people will recommend that college to others.

Next, we run the same horse-race model for certain subsamples, some of which — such as low-income and unemployed individuals — could be expected to be less satisfied with their educational experience. In fact, if labor market success is something college attendees value, seemingly poor returns should translate into low satisfaction. We display the results from these estimations in Table 4. In column 1 we consider whether there

are gender differences in the predictive power of the satisfaction index. We find little indication of such differences, as the point estimates in a model where we condition on gender are practically identical to those in the main model presented in Table 3. Further, in column 2 we consider whether there is heterogeneity with respect to respondents' socioeconomic background. We proxy SES status with maternal education level and define low SES as having a mother who did not complete high school. As with gender, we find no indication that the association between subjective satisfaction and willingness to recommend varies with SES status. In columns 3 and 4 we consider respondents' labor market outcomes and condition the horse-race model on having had either labor market success (defined as being in the highest income category, earning more than USD 100,000 annually) or labor market struggles (defined as being unemployed or outside the workforce). Again we find few indications that the association in question is different across these subgroups. If anything, we note that the point estimate for the high-income group is somewhat lower than in other models; overall, however, the results in Table 4 suggest that the relationship between satisfaction and willingness to recommend uncovered in Table 3 is stable across various background characteristics and outcomes.[5]

What Might Cause Satisfaction?

If subjective satisfaction is a strong predictor of alumni's willingness to recommend, what might explain their satisfaction? We investigate a variety of potential characteristics broadly categorized as either labor market outcomes or college characteristics. While these obviously overlap to some extent, we make this distinction to separate between individual-level outcomes and institution-level characteristics. To investigate how satisfaction levels relate to these covariates, we chart the average, unstandardized

[5]In Figure D.1 in the appendix we provide plots of additional subsample estimations, where we consider a broader spectrum of labor market and educational outcomes. We find no indication that the results presented in Tables 3 and 4 are isolated to particular segments of the sample. In Table D.1 we also show that this result is not driven by particular age groups, as we estimate similar associations between subjective satisfaction and willingness to recommend across all cohorts of alumni.

Table 4—Willingness to Recommend in Subsamples

	Subsamples			
	Women	Low parent ed.	Currently high income	Currently not working
Satisfaction	0.559**	0.523**	0.500**	0.549**
	(0.004)	(0.009)	(0.007)	(0.006)
Mobility Rate	0.007	0.014	0.000	-0.008
	(0.006)	(0.010)	(0.007)	(0.009)
1% Mobility Rate	-0.013**	-0.018	-0.005	-0.004
	(0.004)	(0.012)	(0.006)	(0.007)
Mobility Rank	0.000	0.009	-0.002	0.010
	(0.006)	(0.012)	(0.008)	(0.009)
Barron's Selectivity	0.007	0.016	0.018+	0.009
	(0.007)	(0.015)	(0.010)	(0.011)
Forbes Ranking	0.010*	-0.006	0.003	-0.012
	(0.005)	(0.013)	(0.008)	(0.008)
Structural Quality	-0.005	-0.009	-0.025**	-0.003
	(0.006)	(0.013)	(0.009)	(0.009)
Mean Satisfaction	4.02	3.93	4.11	3.99
Mean WtR	4.02	4.01	4.19	3.96
Observations	74,491	13,346	29,363	30,980
Adjusted R^2	0.287	0.269	0.266	0.285

Note: The table reports results from estimating horse-race models where willingness to recommend is regressed on the set of college-quality metrics, separately for various subsamples. "Mean Satisfaction" and "Mean WtR" refer to average scores for that subsample on the satisfaction index and the willingness-to-recommend item, respectively, both measured in absolute terms on a 1–5 scale (5 highest). Cluster-robust standard errors clustered at the college level in parenthesis. $^{+}$ $p<0.1$, * $p<0.05$, ** $p<0.01$.

score of our satisfaction index across various subsamples in Figures 1 and 2. To do so, we simply average the numerical values related to a respondent's answers on the six items listed in Table 1 to obtain an individual score and then average across individuals for the subsamples of interest. The indices are then averaged for any given subsample. Note that the satisfaction indices can take values from 1 to 5, where a 3 indicates that a respondent neither agrees nor disagrees with a given statement. The overall average satisfaction in the sample is 3.9.

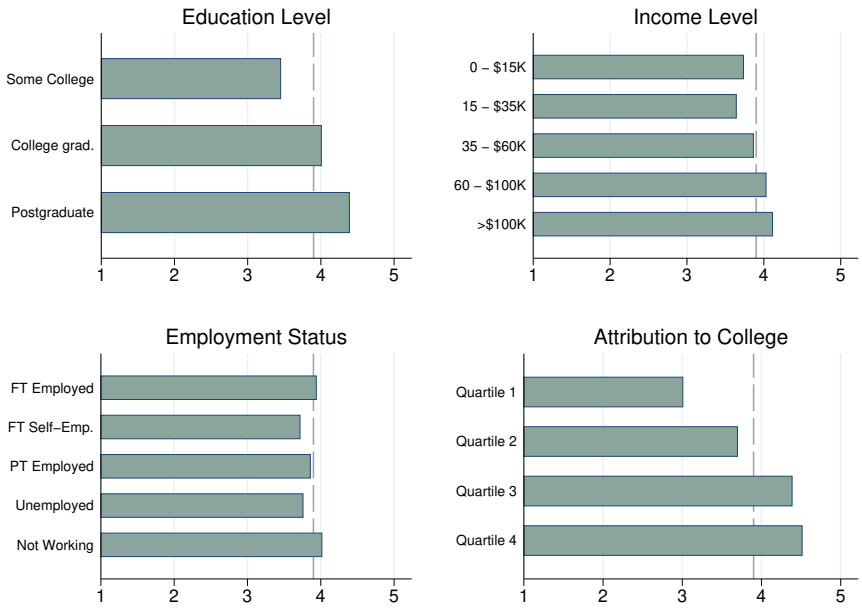


Figure 1: Satisfaction Levels by Labor Market Outcomes

Note: Displayed are raw average scores on the subjective satisfaction index across various indicators of the respondents' labor market outcomes. In each case, the minimum score possible is 1 and the maximum is 5. The dashed line indicates the sample average (3.9). *Education level* is the respondents' highest completed degree, where *some college* includes graduates from two-year programs. Thus *college graduate* refers to those having completed a four-year program.*Income level* and *employment status* are categorical variables indicating a respondent's self-reported current labor market status. For the latter, *Not working* means that the respondent does not consider themselves part of the workforce. *Degree of attribution to college* is an index indicating to what extent a respondent feels that their college and education have been helpful in their career.

Figure 2: Satisfaction Levels by College Characteristics

Note: Displayed are raw average scores on the subjective satisfaction index across various characteristics of the respondents' colleges. In each case, the minimum score possible is 1 and the maximum is 5. The dashed line indicates the sample average (3.9). *Barron's* refers to the selectivity index, ranking colleges from nonselective to most selective. *Forbes ranking* is the Forbes Top Colleges list comprising 650 institutions; Not ranked are respondents from those institutions not included in the ranking, while the ranked institutions are split into three categories. The *Chetty mobility rate* measures the fraction of students entering a college from the lowest quintile of the income distribution who subsequently end up in the highest quintile. For the purpose of this figure, the mobility-rate distribution is split into four quartiles. A similar split into quartiles is performed for the index for structural quality, which is based on institution-level observable characteristics.

Figure 1 charts average satisfaction across labor market outcomes. As is evident from the figure, we find no subsample of respondents who are not leaning toward "satisfied." Even those who are unemployed or report a very low income exhibit high levels of satisfaction with their college experience. We find both of these groups to have an average satisfaction of almost 4 out of 5. For both employment status and income — two of the main outcomes often used to evaluate the benefits from attending college — satisfaction is relatively stable across outcomes, although there are statistically significant differences between them. The lowest average satisfaction found in any subsample is that among those in the lowest quartile of the attribution distribution, that is, among those who feel that their education has been the least helpful in advancing their career. However, it should be noted that even for this group, average satisfaction is not in the lower half of the scale. In fact, even if we look specifically at those who (i) are either unemployed or working but earning a low income and (ii) find themselves in the lower half of the attribution-to-college distribution, we still find an average satisfaction slightly above 3.

We find a similar stability across the distribution of college characteristics, as displayed in Figure 2. While there appears to be a positive relationship between some of the characteristics and satisfaction, such as increasing satisfaction with increasing selectivity, the differences are in most cases trivial, and small. We expected larger differences for subsamples where it intuitively seems reasonable that college satisfaction should be lower, such as those individuals who today earn a low income despite having gone to college, who are unemployed, or who attended colleges that score low on structural measures of institutional quality, but we find few such differences. In fact, of all the variables we considered, we find reasonably large variation in the satisfaction level only for terminal academic degree — particularly the difference between those with a postgraduate education and those with "some college" — and also for the extent to which respondents report that their education has been helpful in their career (*Degree of attribution to college*).

Still, we urge caution in interpreting these results. As we do not ob-

serve counterfactual outcomes, we cannot conclude that the seemingly high satisfaction among individuals with poor labor market outcomes is unjustified. Their outcomes might have been even worse without the education they pursued. Similarly, we cannot rule out the influence of other psychological mechanisms that might explain such high levels of satisfaction — although such mechanisms would probably also operate when ex-college students are asked to give recommendations. For example, cognitive dissonance theory could explain why some report being very satisfied despite poor outcomes, if admitting that their choices were bad or wrong would be painful (Festinger, 1962). It could also be that respondents suffer from egocentric bias and therefore define educational satisfaction (and success) in a manner that is self-aggrandizing and reflects favorably upon themselves (Dunning and Cohen, 1992; Dunning et al., 1995). On a similar note, sociologists and philosophers argue that preferences can be adaptive to one's current circumstances and context, implying that we might find high levels of satisfaction because the respondents have adjusted their expectations and preferences in accordance with their realized life outcomes (Bruckner, 2009).

Determinants for College Choices

Next, we supplement our finding that alumni are overall very satisfied with their college experience with an overview of the enrollment pattern of the college attendees in our sample. The main takeaway from these survey results is that, conditional on deciding to pursue higher education, most claim to have chosen their college for reasons other than labor market prospects or institutional prestige. In Table 5 we list the five most common answers given by respondents to the question, "What is the main reason you decided to enroll in your school?" ("school" in this case being their terminal institution). As is evident in the first row, proximity to home is by far the most common reason for which individuals choose their college. In total, one in five respondents reported proximity as the main reason for choosing the college they attended. Women reported this reason more often than men, and there also appears to be a socioeconomic gradient, as proximity to home is less likely to be reported

Table 5—Top 5 Reasons Respondents Chose Their College

	Total	**Subsamples**			
		Women	Low parent ed.	Currently high income	Currently not working
Reason					
College was close to home	20.48	22.47	24.14	15.95	24.72
Reputation of the school/program	12.95	12.24	9.27	18.41	11.01
Wanted a specific program	11.23	11.29	10.40	10.99	11.30
Location of college in general	7.97	8.34	7.06	8.31	6.96
It was affordable	6.70	6.48	6.48	6.01	5.91
Observations	151,236	72,446	13,678	28,497	30,536

Note: The table reports the shares of respondents who cited each reason as the main reason for why they chose to attend their school. The reasons listed in the table are the top five responses in all the subsamples included there. Low parental education is defined as the mother having dropped out of high school. High income is defined as earning more than USD 100,000 today.*Not working* refers to those who answered "Unemployed" or "Not in the workforce" on their employment status.

as the reason for choosing one's college by high-income individuals than by individuals who are currently unemployed or whose parents have little education. In fact, high-income individuals (those currently earning more than USD 100,000 per year), together with those who attended a school ranked "Most Selective" by the Barron's Index, are alone among the subsamples we looked at in emphasizing school reputation over location. The latter group is actually the one most distinct from the rest of the sample, with roughly 40 percent responding that they chose their college for its prestige and reputation. For all the other subsamples, practical concerns and individual fit seem to be of more importance for college decisions.[6]

In addition to the reasons listed in Table 5, receiving a scholarship and general "convenience" were other popular responses (ranks 6 and 7 overall), which reinforces the impression that prospective students value

[6]In Table D.2 in the appendix we estimate the horse-race model from Table 3, but conditioned on the respondents' main reason for their college choice. We find no substantive variation in results across reasons.

availability and affordability. More career-focused reasons like "Get a good job/make money" and "Advance my career" received less support: only 3.29 and 2.21 percent, respectively, of the sample (rank 10 and 14 overall) put forward these options as their primary motivator for college selection.[7] Note that these responses do not allow us to conclude, for example, that students who emphasize the prestige of a college over it being close to home will end up earning a higher income after graduation as a result of their choice. Still, the relatively low importance placed on characteristics pertaining to the return on investment from attending their preferred college suggests that once the decision to enroll has been made, prospective students choose specific colleges for a host of reasons other than just improving their labor market prospects. In turn, this pattern could explain why satisfaction with your college experience seems rather detached from the labor market returns associated with it.

4.2 Peer Satisfaction and Willingness to Recommend

One concern with our preceding analysis is that, to some extent, subjective satisfaction and subjective willingness to recommend might be two measures of the same underlying construct. To mitigate this endogeneity concern, we replicate our preceding analysis using the leave-out-mean peer satisfaction measure. For any given individual, this measure calculates the average satisfaction of all other alumni graduating from the same institution within +/− 5 years of the respondent in question. Thus, it gives us an indication of whether an institution generally tends to produce alumni who are subjectively satisfied. This means that the peer satisfaction measure is perhaps a college-quality metric of a type that is more comparable to the alternatives employed in our analysis and may thus be of greater relevance to college administrators as well as researchers.

In this peer satisfaction analysis we follow the same approach as at the beginning of this section but substitute the individual satisfaction index

[7] Note also that we find that those who report having enrolled in their college specifically to obtain a good job or earn money, but who report being unemployed or in the lowest income category today, still report a fairly high level of satisfaction with their education (3.88 and 3.59, respectively).

Table 6—Predicting Willingness to Recommend With Peer Satisfaction

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Peer Satisfaction	0.112**	0.163**						
	(0.006)	(0.004)						
Mobility Rate	0.001		0.012*					
	(0.006)		(0.005)					
1% Mobility Rate	-0.013**			0.057**				
	(0.005)			(0.007)				
Mobility Rank	0.011+				0.020**			
	(0.006)				(0.006)			
Barron's Selectivity	0.037**					0.118**		
	(0.007)					(0.005)		
Forbes Ranking	0.012*						0.099**	
	(0.005)						(0.005)	
Structural Quality	0.029**							0.120**
	(0.007)							(0.005)
Mean Peer Satisfaction	3.95							
Mean WtR	3.99							
Observations	126,740	136,079	126,740	126,740	126,740	136,079	136,079	136,079
R^2	0.021	0.019	0.000	0.004	0.000	0.013	0.011	0.014

Note: The table reports results from estimating a set of models where willingness to recommend is regressed separately on each college-quality metrics. Displayed in columns 2–8 are the resulting standardized point estimates for each metric. Column 1 reports the estimates from a multivariate model including all the metrics. *Peer satisfaction* is calculated as a leave-out-mean measure of the satisfaction of a respondent's peers, defined as those graduating from the same institution within 5 years of the respondent. "Mean peer satisfaction" and "Mean WtR" refer to the sample-average scores on the peer satisfaction index and the willingness-to-recommend item, respectively, both measured in absolute terms on a 1–5 scale (5 highest). Cluster-robust standard errors clustered at the college level in parenthesis. $^{+}$ $p<0.1$, * $p<0.05$, ** $p<0.01$.

with our leave-out-mean peer measures. In Table 6, we report the results from our horse-race models estimating the relationship between college-quality metrics and alumni's willingness to recommend. Although levels of predictive power are markedly lower, as with the individual satisfaction measure we find that peer satisfaction is the only meaningful predictor of WtR, as is evident from column 1 in the table. In this full model, we find that an increase by one standard deviation in average satisfaction among your college peers is associated with an increase by 11.2 percent of a standard deviation in the likelihood that you would recommend others to attend the same school. We also find that the peer satisfaction estimate

is robust to whether or not we include the other college-quality metrics in the same model. Keep in mind that the peer satisfaction measure is based on a leave-out-mean procedure where a respondent's own subjective satisfaction does not contribute to the score. Interestingly, other quality measures appear to be substantially poorer predictors. For example, increasing the selectivity of the school by one category is associated with an increase by only 3.7 percent of a standard deviation in the willingness to recommend, while we find no significant associations for the main Chetty et al. mobility-rate measures.

In Table 7 we report results from estimating the peer satisfaction horse-race models for the same subsamples as those considered in Table 4. As before, we find the estimates to be consistent across all the subsamples considered. However, this is not surprising given the remarkable stability in satisfaction scores across different subsamples seen in Figures 1 and 2. Similar patterns are observed for willingness to recommend. As with satisfaction, the most striking result is the remarkable stability, even for groups with poor labor market outcomes that we expected would have lower satisfaction and willingness to recommend. This is further illustrated in Figures B.1 and B.2 in the appendix, where we chart average peer satisfaction levels across labor market outcomes and school characteristics. As with individual satisfaction, the most apparent feature is the stability of the peer measure, even in subsamples with poor labor market returns.

4.3 Correlating Satisfaction With Alternative Measures

We conclude our analysis by investigating the correlation between our satisfaction indices and other proposed measures of college quality. Strong correlations might shed light on the characteristics associated with colleges that alumni value and that they might thus draw particularly upon when formulating their advice. Pairwise correlations between all measures considered are presented in Table 8. In particular, focus should be placed on columns 1 and 2, which display the correlations between individual satisfaction and peer satisfaction, respectively, and the alternative mea-

Table 7—Peer Satisfaction Subsamples

	Subsamples			
	Women	Low parent ed.	Currently high income	Currently not working
Peer Satisfaction	0.101**	0.092**	0.116**	0.089**
	(0.007)	(0.016)	(0.011)	(0.011)
Mobility Rate	0.010	0.004	0.011	-0.008
	(0.008)	(0.013)	(0.009)	(0.012)
1% Mobility Rate	-0.022**	-0.020	-0.015*	-0.016+
	(0.006)	(0.017)	(0.007)	(0.008)
Mobility Rank	0.008	0.034*	-0.002	0.019
	(0.008)	(0.016)	(0.010)	(0.012)
Barron's Selectivity	0.041**	0.040+	0.038**	0.035*
	(0.008)	(0.022)	(0.013)	(0.016)
Forbes Ranking	0.005	0.004	0.009	0.005
	(0.007)	(0.017)	(0.010)	(0.011)
Structural Quality	0.042**	0.031	0.028*	0.029+
	(0.009)	(0.021)	(0.011)	(0.016)
Mean Peer Satisfaction	3.94	3.88	4.04	3.91
Mean WtR	4.04	4.02	4.21	3.96
Observations	59,557	10,327	25,485	25,082
Adjusted R^2	0.020	0.017	0.019	0.014

Note: The table reports results from estimating horse-race models where willingness to recommend is regressed on the set of college-quality metrics, separately for various subsamples. *Peer satisfaction* is calculated as a leave-out-mean measure of the satisfaction of a respondent's peers, defined as those graduating from the same institution within 5 years of the respondent. "Mean Peer Satisfaction" and "Mean WtR" refer to the sample average scores on the peer satisfaction index and the willingness-to-recommend item, respectively, both measured in absolute terms on a 1–5 scale (5 highest). Cluster-robust standard errors clustered at the college level in parenthesis. + $p<0.1$, * $p<0.05$, ** $p<0.01$.

sures. First, column 1 indicates that a respondent's own satisfaction has a correlation of 0.21 with their peers' satisfaction. To the extent that a person's own satisfaction should be considered an accurate reflection of the college's overall ability to generate satisfaction among its students (which giving advice based on one's own experience might presuppose), this correlation is low. In fact, such a low correlation between an individ-

ual's satisfaction and that of the remaining alumni peer group suggests that students evaluate their experience subjectively and that experiences vary between students within a college.

Next, rows 3–5 report correlations between our satisfaction measures and the Chetty et al. mobility metrics. In row 3, we see that the general mobility rates are weakly correlated with student satisfaction. That is, we cannot say, for example, that satisfied students are generally those who attended colleges with high social mobility. We find a comparably weak association between the mobility rank and satisfaction. On the other hand, there is a stronger relationship between satisfaction and the 1% mobility rate. This is particularly the case for peer satisfaction, with which the 1% mobility has a correlation of 0.29. Considering the substantial upward mobility represented by a jump from the lowest quintile to the top 1% of the income distribution, it is reasonable for colleges that excel on this metric to have satisfied students on average. In particular, we would expect certain subsamples of students at these high-mobility schools (e.g., those with a low socioeconomic background, in our case defined as those whose mothers have a low educational level) to have especially appreciated the potential for upward mobility offered by their schools. However, for these samples as well, we find that the correlation is, if anything, lower between the satisfaction and mobility measures.

In the final three rows we consider examples of "objective" measures of college quality. All three show weak to modest correlations with the individual satisfaction measure, but strong correlations with the peer satisfaction measure. Both the Barron's Selectivity Index, the structural-quality index, and the Forbes ranking have a correlation with peer satisfaction which is close to 0.5. This is to be expected if colleges ranked high on the Forbes list are also highly selective and score high on observable characteristics such as admission rates, tuition, faculty salaries, etc. Judging from these results, they also produce satisfied alumni.

Finally, in Table 9 we fit a similar horse-race model as in the analysis above using the set of college-quality metrics. However, now we use our satisfaction indices as the outcome variable, to see which metrics are stronger predictors of alumni satisfaction. For both the individual and the

Table 8—Correlations Between Satisfaction and Quality Metrics

	Satisfaction	Peer Satisfaction	Mobility Rate	1% Mobility Rate	Mobility Rank	Barron's Selectivity	Forbes Ranking	Structural Quality
Satisfaction	1.000							
Peer satisfaction	0.214	1.000						
Mobility rate	0.022	0.063	1.000					
1% Mobility rate	0.125	0.292	0.432	1.000				
Mobility rank	0.034	0.085	0.819	0.423	1.000			
Barron's	0.213	0.473	0.055	0.506	0.068	1.000		
Forbes ranking	0.177	0.420	0.099	0.346	0.066	0.703	1.000	
Structural quality	0.226	0.514	0.007	0.471	0.063	0.772	0.629	1.000

Note: The table reports the pairwise correlations between the college quality indicators listed. All correlations are significant at the 1 percent level.

Table 9—Predicting Satisfaction

	Individual Satisfaction	Peer Satisfaction
Mobility Rate	0.003	0.016
	(0.009)	(0.021)
1% Mobility Rate	-0.006	-0.022
	(0.008)	(0.020)
Mobility Rank	0.019*	0.033
	(0.009)	(0.022)
Barron's Selectivity	0.092**	0.225**
	(0.011)	(0.032)
Forbes Ranking	0.022**	0.049*
	(0.009)	(0.023)
Structural Quality	0.141**	0.320**
	(0.011)	(0.032)
Observations	155,619	126,740
Adjusted R^2	0.056	0.422

Note: The table reports results from the regression of individual and peer satisfaction, respectively, on the set of college-quality metrics. $^{+}$ $p<0.1$, * $p<0.05$, ** $p<0.01$.

peer satisfaction measure, we find that school selectivity and structural quality are the only metrics with meaningful predictive power. However, we note that, in the case of the individual index, the association is much weaker. All else being equal, these estimates suggest that graduating from a more selective school or from a school with higher structural quality is predictive both of your own and of your peers' satisfaction with the educational experience in that school. More generally, the results in Tables 8 and 9 indicate that the alternative measures of college quality are positively correlated with a college's ability to generate satisfaction among its students. However, those measures do not seem to be strong predictors of satisfaction at the individual level, suggesting that the individual experience and evaluation of a particular college are highly subjective and only partly correlated with the general satisfaction of the alumni population.

5 The Role of Rankings in Decision-Making

Before discussing the implications of our findings, we would like to stress at the onset that there are limitations to our analysis. We have no exogenous variation in satisfaction that might facilitate causal estimates. The satisfaction measures are retrospective, and we do not know the counterfactual satisfaction levels that would have been obtained if the respondents had chosen to attend different schools, or had experienced different labor market outcomes. Moreover, our results shed little light on how individuals with no college experience view college or on how alumni view colleges other than their own. Our results focus on alumni and on their satisfaction with their own alma mater. While our analysis provides some clues, it does not fully reveal the underlying reasons for individual satisfaction. This could come from satisfaction with academic quality, from satisfaction with the link between academic studies and subsequent careers, from satisfaction with school amenities, from satisfaction with professional and social networks, from satisfaction with affordability, from satisfaction with location, from satisfaction with friends and partners met while in school, from satisfaction with any other dimension of the types of offerings made to students, and of course from any combination of the above.

Yet despite these limitations, our analysis provides some novel clues to how college alumni view their educational experiences. We document that the vast majority of college attendees are very satisfied with the educational path they took, even those who seemingly had poor labor market returns. While the indices exhibit positive correlations with other metrics of college quality, we find only limited evidence that subjective satisfaction is predicted by outcomes in adulthood or by more established measures of school quality. The satisfaction indices also have substantial predictive power over alumni's willingness to recommend others to follow a similar path in terms of college choices — much more than other "objective" rankings, measures of social mobility, and college-quality metrics.

The finding that alumni hold such strong preferences for their alma mater, that the satisfaction metrics have such a broad predictability rel-

ative to future outcomes, and that trusted adults are so important as college-choice advisors also have policy implications. If the criteria that lead to satisfaction align with the goals of the central planner, then satisfaction metrics might inform and reinforce strategic policy proposals. However, if the criteria do not align with those goals, then the realities underpinning these satisfaction metrics might instead counteract them. As an example, suppose a central planner is trying to improve intergenerational mobility and that to do so, she is planning to promote colleges that facilitate such mobility. The rankings based on Chetty et al. (2017) may provide guidance in her choice of colleges to promote or model after. However, given the low correlation between the Chetty et al rankings and the satisfaction metrics, advice from adults who rely on satisfaction-based metrics to inform recommendations will not promote the central planner's goals. If, by contrast, the central planner's goal is to maximize utility for individuals who might desire to attend college, even for pure consumption purposes, satisfaction indices might help her attain that goal.

Our findings may also have implications for the design of informational programs designed to affect college choice. If students rely on information from trusted adults and trusted adults rely on satisfaction to shape recommendations, then informational programs that are not in harmony with satisfaction may ultimately fail. By contrast, informational programs aimed at trusted adults might strengthen the recommendations that they give for institutions that they did not themselves attend.

While satisfaction indices merit the caveats previously discussed, their relevance even among diverse subsamples and their ability to predict outcomes are striking. As those satisfaction metrics are further honed, researchers might develop a richer sense of the underlying preferences that individuals have across a wide swath of collegiate offerings. Such information can inform both policymakers and college administrators about how students and alumni view colleges, and about how those views might influence future cohorts of students in their decision-making. As our results imply that satisfaction indices have relevance both for the college decision-making process and for college evaluations, they should warrant interest from other researchers going forward.

References

Aguirre, J., & Matta, J. (2021). Walking in Your Footsteps: Sibling Spillovers in Higher Education Choices. *Economics of Education Review*, *80*, 102062.

Altmejd, A., Barrios-Fernández, A., Drlje, M., Goodman, J., Hurwitz, M., Kovac, D., ... Smith, J. (2021). O Brother, Where Start Thou? Sibling Spillovers on College and Major Choice in Four Countries. *The Quarterly Journal of Economics*, *Forthcoming*.

Avery, C., & Kane, T. J. (2004). Student Perceptions of College Opportunities. The Boston COACH program. In C. M. Hoxby (Ed.), *College Choices: The Economics of Where to Go, When to Go, and How to Pay For It* (Chap. 8). University of Chicago Press.

Barone, C., Schizzerotto, A., Abbiati, G., & Argentin, G. (2017). Information Barriers, Social Inequality, and Plans for Higher Education: Evidence From a Field Experiment. *European Sociological Review*, *33*(1), 84–96.

Barr, A., & Castleman, B. (2018). An Engine of Economic Opportunity: Intensive Advising, College Success, and Social Mobility. *Working Paper*.

Becker, G. S. (1962). Investment in Human Capital: A Theoretical Analysis. *Journal of Political Economy*, *70*(5, Part 2), 9–49.

Bergman, P., Denning, J. T., & Manoli, D. (2019). Is Information Enough? The Effect of Information about Education Tax Benefits on Student Outcomes. *Journal of Policy Analysis and Management*, *38*(3), 706–731.

Bettinger, E. P., Long, B. T., Oreopoulos, P., & Sanbonmatsu, L. (2012). The Role of Application Assistance and Information in College Decisions: Results from the H&R Block FAFSA Experiment. *The Quarterly Journal of Economics*, *127*(3), 1205–1242.

Bird, K. A., Castleman, B. L., Denning, J. T., Goodman, J., Lamberton, C., & Rosinger, K. O. (2021). Nudging at Scale: Experimental Evidence from FAFSA Completion Campaigns. *Journal of Economic Behavior & Organization*, *183*, 105–128.

Bruckner, D. W. (2009). In Defense of Adaptive Preferences. *Philosophical Studies*, *142*(3), 307–324.

Campbell, S., Macmillan, L., Murphy, R., & Wyness, G. (2021). Matching in the Dark? Inequalities in Student to Degree Match. *NBER Working Paper Series*, (29215).

Carrell, S., & Sacerdote, B. (2017). Why Do College-Going Interventions Work? *American Economic Journal: Applied Economics*, *9*(3), 124–151.

Chetty, R., Friedman, J. N., Saez, E., Turner, N., & Yagan, D. (2017). Mobility Report Cards: The Role of Colleges in Intergenerational Mobility. *NBER Working paper*, (23618).

Chetty, R., Friedman, J. N., Saez, E., Turner, N., & Yagan, D. (2020). Income Segregation and Intergenerational Mobility Across Colleges in the United States. *The Quarterly Journal of Economics*, *135*(3), 1567–1633.

Coudriet, C. (2018). Top Colleges 2018: The Methodology. *Forbes*.

Cunha, J. M., Miller, T., & Weisburst, E. (2018). Information and College Decisions: Evidence From the Texas GO Center Project. *Educational Evaluation and Policy Analysis*, *40*(1), 151–170.

Dillon, E. W., & Smith, J. A. (2017). Determinants of the Match Between Student Ability and College Quality. *Journal of Labor Economics*, *35*(1), 45–66.

Dunning, D., & Cohen, G. L. (1992). Egocentric Definitions of Traits and Abilities in Social Judgment. *Journal of Personality and Social Psychology*, *63*(3), 341.

Dunning, D., Leuenberger, A., & Sherman, D. A. (1995). A New Look at Motivated Inference: Are Self-Serving Theories of Success a Product of Motivational Forces? *Journal of Personality and Social Psychology*, *69*(1), 58.

Festinger, L. (1962). *A Theory of Cognitive Dissonance*. Stanford, CA: Stanford University Press.

Goodman, J., Hurwitz, M., Smith, J., & Fox, J. (2015). The Relationship Between Siblings' College Choices: Evidence from One Million SAT-Taking Families. *Economics of Education Review*, *48*, 75 –85.

Gurantz, O., Howell, J., Hurwitz, M., Larson, C., Pender, M., & White, B. (2021). A National-Level Informational Experiment to Promote Enrollment in Selective Colleges. *Journal of Policy Analysis and Management*, *40*(2), 453–479.

Horn, L. J., Chen, X., & Chapman, C. (2003). Getting Ready To Pay for College: What Students and Their Parents Know about the Cost of College Tuition and What They Are Doing To Find Out. *National Center for Education Statistics Report No. 2003030*.

Hoxby, C. M. (2019). The Productivity of U.S. Postsecondary Institutions. In C. M. Hoxby & K. Stange (Eds.), *Productivity in Higher Education* (Chap. 2, pp. 31–66).

Hoxby, C. M., & Avery, C. (2013). The Missing "One-Offs": The Hidden Supply of High-Achieving, Low Income Students. *Brookings Paper on Economic Activity*, (Spring), 1–66.

Hoxby, C. M., & Turner, S. (2015). What High-Achieving Low-Income Students Know About College. *American Economic Review: Papers & Proceedings*, *105*(5), 514–517.

Hurwitz, M., & Smith, J. (2018). Student Responsiveness to Earnings Data in the College Scorecard. *Economic Inquiry*, *56*(2), 480–492.

Hyman, J. (2020). Can Light-Touch College-Going Interventions Make a Difference? Evidence from a Statewide Experiment in Michigan. *Journal of Policy Analysis and Management*, *39*(1), 159–190.

Jensen, R. (2010). The (Perceived) Returns to Education and the Demand for Schooling. *The Quarterly Journal of Economics*, *125*(2), 515–548.

Kerr, S. P., Pekkarinen, T., Sarvimäki, M., & Uusitalo, R. (2020). Post-Secondary Education and Information on Labor Market Prospects: A Randomized Field Experiment. *Labour Economics*, *66*, 101888.

Ma, J., Pender, M., & Welch, M. (2016). *Education Pays 2016: The Benefits of Higher Education for Individuals and Society. Trends in Higher Education Series*. Washington, D.C.: The College Board.

Mabel, Z., Libassi, C., & Hurwitz, M. (2020). The Value of Using Early-Career Earnings Data in the College Scorecard to Guide College Choices. *Economics of Education Review*, *75*, 101958.

McGuigan, M., McNally, S., & Wyness, G. (2016). Student Awareness of Costs and Benefits of Educational Decisions: Effects of an Information Campaign. *Journal of Human Capital*, *10*(4), 482–519.

Mulhern, C. (2020). Beyond Teachers: Estimating Individual Guidance Counselors' Effects on Educational Attainment. *Working Paper*.

Oreopoulos, P. (2020). Promises and Limitations of Nudging in Education. *IZA Discussion Paper*, (13718).

Otto, L. B. (2000). Youth Perspectives on Parental Career Influence. *Journal of Career Development*, *27*(2), 111–118.

Oymak, C. (2018). *High School Students' Views on Who Influences Their Thinking About Education and Careers*. Washington, D.C.: U.S. Department of Education, National Center for Education Statistics.

Rothwell, J. (2019). Assessing the Validity of Consumer Ratings for Higher Education: Evidence from a New Survey. *Journal of Consumer Affairs*, *53*(1), 167–200.

Smith, J., Hurwitz, M., & Avery, C. (2017). Giving College Credit Where It Is Due: Advanced Placement Exam Scores and College Outcomes. *Journal of Labor Economics*, *35*(1), 67–147.

Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology*, *5*(2), 207 –232.

U.S. Department of Education. (2013). Education Department Releases College Scorecard to Help Students Choose Best College for Them. *[Press Release]*.

Wiswall, M., & Zafar, B. (2015). How Do College Students Respond to Public Information about Earnings? *Journal of Human Capital*, *9*(2), 117–169.

Appendix

Appendix A:

More Determinants for College Choices

Table A.1—Top 15 Reasons Respondents Chose Their College

	Total	**Subsamples**			
		Women	Low parent ed.	Currently high income	Currently not working
Reason					
College was close to home	20.48	22.47	24.14	15.95	24.72
Reputation of the school/program	12.95	12.24	9.27	18.41	11.01
Wanted a specific program	11.23	11.29	10.40	10.99	11.30
Location of college in general	7.97	8.34	7.06	8.31	6.96
It was affordable	6.70	6.48	6.48	6.01	5.91
Received scholarship/aid	5.18	5.04	4.64	5.77	4.65
Convenience	4.18	4.45	5.32	4.21	3.92
Advance knowledge/like to learn	3.81	3.39	5.57	3.16	4.98
School was a good fit	3.41	3.59	2.49	2.97	2.91
Get a good job/make more money	3.29	2.82	4.87	2.76	4.00
Other	3.28	3.14	3.11	3.74	3.03
School offered online/night classes	2.86	3.49	3.19	2.97	2.31
Friends/family go there	2.22	2.23	1.51	2.19	2.11
Advance career	2.21	1.75	2.92	2.59	2.44
Got accepted/recruited	2.02	1.63	1.44	2.41	1.75
Observations	151,236	72,446	13,678	28,497	30,536

Note: Table reports share of respondents who cited this reason as the main reason for why they chose to attend their school. The reasons listed, and the ordering, is based on the top 15 reasons in the full sample. Low parental education is defined as the mother having dropped out of high school. High income is defined as earning more than $100,00 today. Not working is those who answered "Unemployed" or "Not in workforce" on their employment status. Unranked selectivity refers to schools not ranked by the Barron's Selectivity Index, while most selective is the highest rank.

Table A.2—Top 15 Reasons Respondents Pursued Higher Education

	Total	**Subsamples**			
		Women	Low parent ed.	Currently high income	Currently not working
Reason					
Get good job/better pay	27.09	27.60	30.25	26.86	26.73
Advance knowledge/learn	17.21	17.38	20.00	14.61	19.77
Wanted to attend a specific program	13.63	14.98	13.1	13.46	15.19
Advance career	13.23	12.88	12.04	16.89	11.34
It is expected	7.59	7.26	4.03	7.02	6.09
Family influence/first to graduate	4.45	4.60	3.34	4.77	3.86
Bored/Something to do	4.05	3.87	3.73	3.83	4.26
Other reason	3.42	3.37	3.41	3.86	3.19
It was affordable	1.96	1.89	2.30	1.99	2.05
Change careers	1.74	1.71	2.06	1.43	1.85
Received scholarship/financial aid	1.55	1.00	1.67	1.63	1.65
School offered online/night classes	0.55	0.33	0.70	0.33	0.54
Recommendation from friend	0.51	0.38	0.50	0.47	0.39
Don't know	0.44	0.41	0.42	0.37	0.45
School was a good fit	0.34	0.33	0.30	0.41	0.29
Observations	151,236	72,446	13,678	28,497	30,536

Note: Table reports share of respondents who cited this reason as the main reason for why they chose to pursue higher education. The reasons listed, and the ordering, is based on the top 15 reasons in the full sample. Low parental education is defined as the mother having dropped out of high school. High income is defined as earning more than $100,000 today. Not working is those who answered "Unemployed" or "Not in workforce" on their employment status. Unranked selectivity refers to schools not ranked by the Barron's Selectivity Index, while most selective is the highest rank.

Appendix B:

Average Individual and Peer Satisfaction Across Subsamples

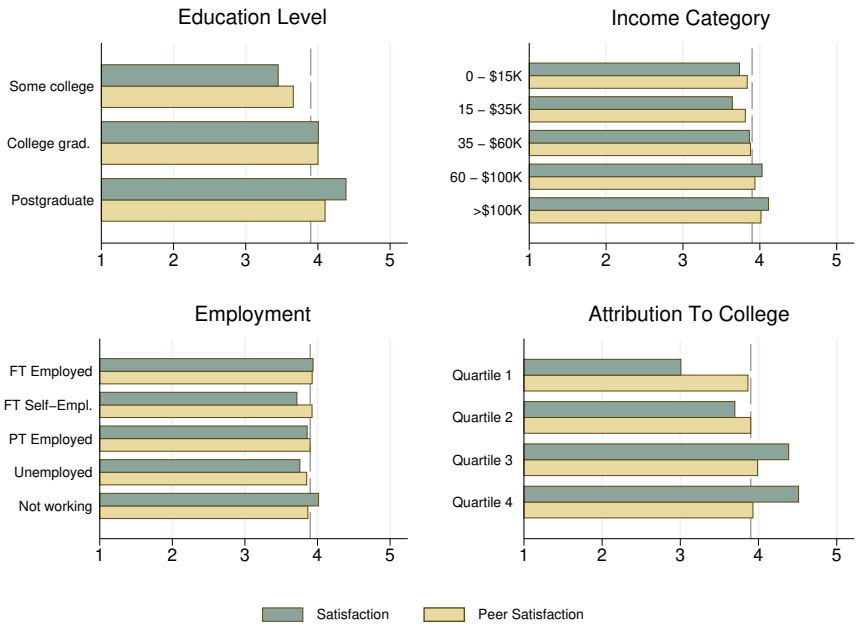


Figure B.1: Satisfaction Levels by Labor Market Outcomes

Note: Displayed are raw average scores on the individual satisfaction and the leave-out-mean peer satisfaction index across various indicators of the respondents' labor market outcomes. To construct the latter we sum all satisfaction index score for alumni of a specific institution, subtract the respondents own score, and then divide on the number of observations for that institution.In each case, the minimum score possible is 1, and the maximum 5. The dashed line indicates the sample average (3.9). *Education level* is the respondents highest completed degree, where *some college* includes graduates from 2-year programs. Thus *college graduate* refers to those completing a 4-year program.*Income category* and *employment status* are categorical variables indicating the respondent's self-reported, current labor market status. For the latter, *Not working* means that the respondent does not consider themselves part of the work force. *Degree of Attribution to college* is an index indicating to what degree the respondent feels their college and education has been helpful in their career.

Figure B.2: Satisfaction Levels by College Characteristics

Note: Displayed are raw average scores on the individual satisfaction and the leave-out-mean peer satisfaction index across various characteristics of the respondents' colleges. To construct the latter we sum all satisfaction index score for alumni of a specific institution, subtract the respondents own score, and then divide on the number of observations for that institution.In each case, the minimum score possible is 1, and the maximum 5. The dashed line indicates the sample average (3.9). Barron's refers to the selectivity index, ranking colleges from nonselective to most selective. Forbes ranking is the Forbes Top Colleges list comprising 650 institutions. Not ranked are respondents from those institutions not included in the ranking, while the ranked institutions are split in three.The Chetty mobility rate measures the fraction of students entering the college from the lowest quintile of the income distribution who subsequently end up in the highest quintile. For the purpose of this graph, the mobility rate distribution is split in four quartiles. Similarly, we split the index for structural quality, based on institution level observable characteristics, in quartiles, and consider means separately within each of them.

Appendix C:

Predicting Income (Replicating Rothwell, 2019)

In the following analysis we replicate and build on Rothwell (2019) who showed that satisfaction is positively correlated with income, to investigate to what extent satisfaction explains variation in income levels, and more so than the alternative measures. We do so by running a series of regression models, paying particular attention to the explanatory power (R-squared) of the model. We consider two outcomes, the ECS 5-tier income category variable, as well as a continuous measure of the log of real income. The latter is only observed for individuals whose income is positive at the point of the survey (111,920 observations, 65.3%). We therefore include results from models using the income category as the outcome to include those that reportedly have no income. We find that alumni satisfaction seems to have substantial predictive power for subsequent earnings.

As a first stage, we run univariate models to simply assess the predictive power of each of the quality measures for subsequent income. To be able to compare effect sizes across metrics, all dependent and independent variables are standardized to reflect standard deviation units. We then regress the two income measures on the nine separate measures of college quality to assess to what extent income is explained by variation in these metrics. The results from this exercise are displayed in Figure C.2, where we normalize all the effect sizes to values in the range 0 to 1, with the largest effect size taking the value 1. All correlations displayed here are significant at the 1% level. We regress the two income measures on nine separate measures of college quality to assess to what extent income is explained by variation in these metrics. For the satisfaction indices we find that a one standard deviation increase in both your own satisfaction, and the average satisfaction of your peers appears to be predictive of higher earnings. In fact, a one standard deviation increase in the peer satisfaction measure predicts an increase in income among positive income earners of almost 23 percent of a standard deviation, an effect size four times larger than that of the predicted increase in income associated with attending a school with a one standard deviation higher mobility rate. We find that for both income measures, the point estimate for peer satisfaction measure is larger than any other quality metric that we tested.

If we turn to the explanatory power of these simple models, we find that all of the bivariate regressions have an R-squared in the 0.1–5.5 percent range. The

metric that alone explains the most variation in income is our structural quality index with an R-squared of 5.0–5.6 percent. Peer satisfaction alone explains 3.8–4.3 percent of income variation, on par with that of for example school selectivity and the Forbes college ranking. This implies that for our data, peer satisfaction as a measure of college quality is equally adept at explaining subsequent alumni income as more established metrics and rankings. In Figure C.1 we illustrate the relative amount of explained variance across our bivariate models, with the largest R-squared normalized to take the value 1. While it is evident that several of these metrics are equally adept at explaining the variation in alumni income levels, it is worth noticing the relatively poor performance of the Chetty et al. mobility measure, despite being based on subsequent income distributions. To that point, we see that the 1 percent mobility rate, which indicates a substantial increase in the number of alumni who end up in the top 1 percent of the income distribution, has a much higher R-squared.

To further explore the relative predictive power of these metrics, we also run a "horse race" model where all quality metrics are included in the same regression. These results are charted in Figure C.3. For both the categorical and the continuous income measure we find that peer satisfaction has the second highest estimated predicted correlation with income, only exceeded by the comprehensive structural quality index, with the Barron's selectivity index coming in third. The remaining metrics return smaller to insignificant effect sizes. The fact that the point estimate for peer satisfaction remains relatively large and significant, even when we control for all the other college quality measures, implies that our satisfaction index captures variation that the other metrics don't. In the opposite case—if, say, the estimate for peer satisfaction dropped to zero once we controlled for college selectivity—we would be concerned about the overlap with existing measures contradicting our claim that satisfaction has potential as a separate, additional metric for college quality. Overall, we argue that the results from the simple exercises in this section demonstrate that peer satisfaction at worst performs as well as established quality measures in explaining alumni income levels. The fact that these satisfaction indices are predictive of real life outcomes to the extent that they are, adds to their saliency as indicators of students' academic experiences.

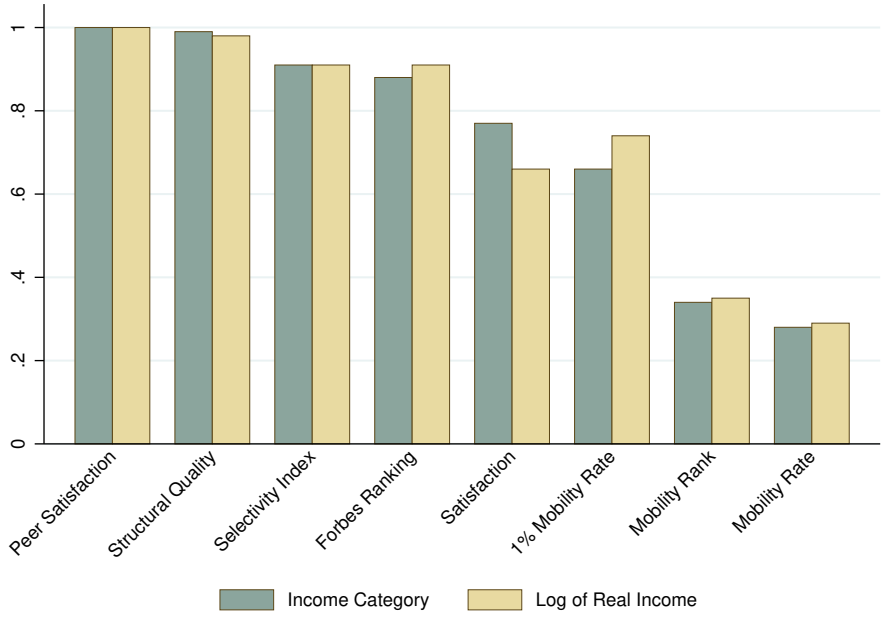


Figure C.1: Point Estimates From Bivariate Predictions of Income

Note: The figure charts beta coefficients from bivariate regression of the respective income measures on the various proposed metrics for college quality. All coefficients are first standardized to have a mean of 0 and a standard deviation of 1, then results are normalized, with the largest beta coefficient set to 1. The bars are in descending order according to the beta coefficient obtained from regression the respondents income category on the metric in question

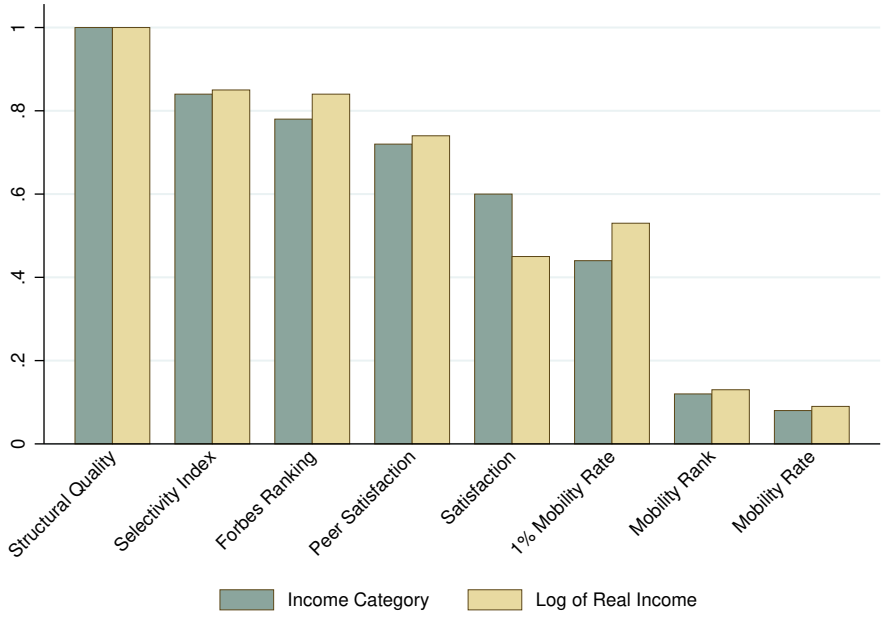


Figure C.2: Explained Variation by Quality Metric

Note: The figure charts the explained variation (R^2) in the respective income measures, when performing a bivariate regression of that measure on the various proposed metrics for college quality. The results are normalized, with the highest resulting R^2 set to 1. The bars are in descending order according to the R^2 obtained from regression the respondents income category on the metric in question.

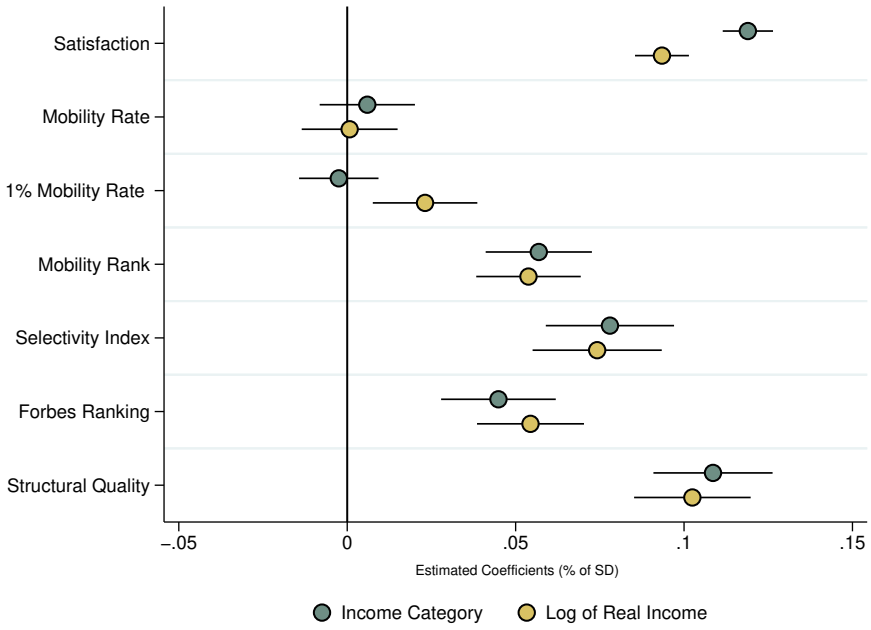


Figure C.3: Horse Race Model – Predicting Income

Note: The figure displays standardized coefficients and cluster-robust standard errors from a multivariate regression using all the proposed measures of college quality as predictors of income. The dots represent separate regressions, with (standardized) income category (top) and the log of real income (bottom) as the dependent variables, respectively.

Appendix D:

Horse Race Models Across More Subsamples

Table D.1—By Reason the Respondent Chose to Enroll in Their College

	Reason				
	College was close to home	Reputation of school/program	To attend specific program	Location in general	It was affordable
Satisfaction	0.542**	0.591**	0.529**	0.523**	0.542**
	(0.005)	(0.009)	(0.009)	(0.009)	(0.010)
Mobility Rate	-0.008	-0.001	0.001	0.012	0.000
	(0.010)	(0.010)	(0.009)	(0.015)	(0.015)
1% Mobility Rate	-0.008	-0.003	-0.021*	-0.010	0.003
	(0.010)	(0.006)	(0.009)	(0.011)	(0.014)
Mobility Rank	-0.000	0.004	0.010	0.011	0.018
	(0.010)	(0.010)	(0.010)	(0.014)	(0.016)
Barron's Selectivity	0.007	0.022+	0.011	0.029+	-0.029
	(0.012)	(0.012)	(0.014)	(0.016)	(0.019)
Forbes Ranking	-0.004	0.002	0.025*	0.005	0.003
	(0.010)	(0.010)	(0.012)	(0.014)	(0.014)
Structural Quality	-0.001	-0.024*	-0.021+	-0.033*	0.006
	(0.011)	(0.010)	(0.012)	(0.015)	(0.009)
Mean Satisfaction	3.75	4.29	4.03	3.90	3.77
Mean WtR	3.84	4.23	4.06	3.95	3.82
Observations	28,782	18,224	15,432	11,277	9,280
Adjusted R^2	0.297	0.279	0.271	0.271	0.297

Note: "WtR" = Willingness to recommend. The table reports results from estimating horse race models where willingness to recommend is regressed on the set of college-quality metrics, separately for the subsamples who gave particular responses to the survey item "What is the main reason you chose to enroll in the college/institution you did?". "Mean Satisfaction" and "Mean WtR" refer to average scores for that subsample on the satisfaction index and the willingness-to-recommend item respectively, both measured in absolute terms on a 1–5 scale (5 highest). Cluster-robust standard errors clustered at the college level in parenthesis. + $p<0.1$, * $p<0.05$, ** $p<0.01$.

Table D.2—By Decade Enrolled in College

	Decade				
	<1980	1980s	1990s	2000s	≥2010
Satisfaction	0.496**	0.533**	0.540**	0.564**	0.582**
	(0.007)	(0.006)	(0.005)	(0.005)	(0.005)
Mobility Rate	-0.002	-0.007	0.002	-0.003	-0.000
	(0.010)	(0.010)	(0.011)	(0.009)	(0.008)
1% Mobility Rate	-0.007	-0.014*	-0.006	-0.004	-0.008
	(0.009)	(0.006)	(0.006)	(0.007)	(0.006)
Mobility Rank	0.012	0.017+	0.005	-0.003	0.002
	(0.011)	(0.010)	(0.010)	(0.009)	(0.008)
Barron's Selectivity	0.006	0.031**	0.021*	-0.016	-0.019*
	(0.013)	(0.010)	(0.010)	(0.010)	(0.009)
Forbes Ranking	0.002	-0.010	0.005	0.013+	0.013+
	(0.010)	(0.008)	(0.007)	(0.008)	(0.008)
Structural Quality	0.012	0.003	-0.015+	-0.021*	-0.060**
	(0.011)	(0.009)	(0.009)	(0.009)	(0.008)
Mean Satisfaction	3.93	3.97	3.96	3.87	3.87
Mean WtR	4.00	4.02	4.01	3.92	3.93
Observations	20,494	34,636	32,791	32,657	35,041
Adjusted R^2	0.254	0.286	0.298	0.316	0.320

Note: "WtR" = Willingness to recommend. The table reports results from estimating horse-race models where willingness to recommend is regressed on the set of college-quality metrics, separately for subsamples based on the decade the respondent attended college. "Mean Satisfaction" and "Mean WtR" refer to average scores for that subsample on the satisfaction index and the willingness-to-recommend item respectively, both measured in absolute terms on a 1–5 scale (5 highest). Cluster-robust standard errors clustered at the college level in parenthesis. + $p<0.1$, * $p<0.05$, ** $p<0.01$.

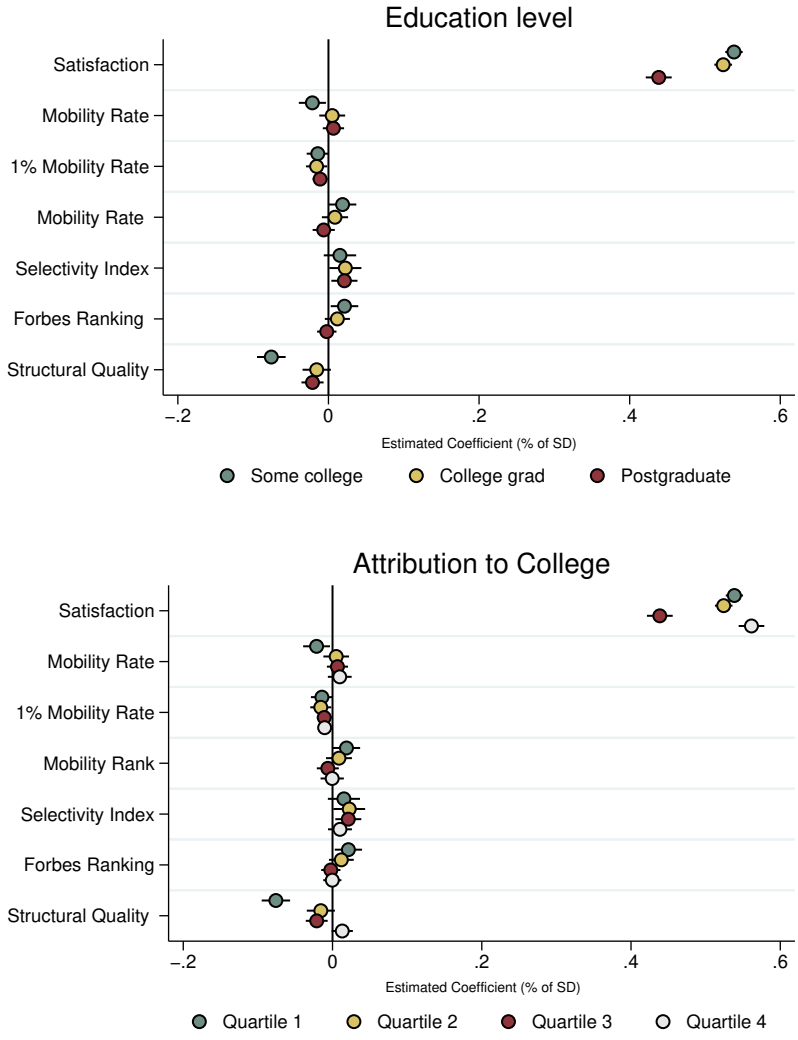


Figure D.1: Predictors of Willingness to Recommend for Different Subsamples

Note: The figures chart results from estimating a set of models where willingness to recommend is regressed on a set of college quality metrics. Displayed are the resulting standardized point estimates for each metric, with 95% confidence intervals. We run separate models for each of subsamples indicated in each figure. *Education Level* is the respondents highest completed degree, where *some college* includes graduates from 2-year programs. Thus *college graduate* refers to those completing a 4-year program. *Degree of Attribution to College* is an index indicating to what degree the respondent feels their college and education has been helpful in their career.

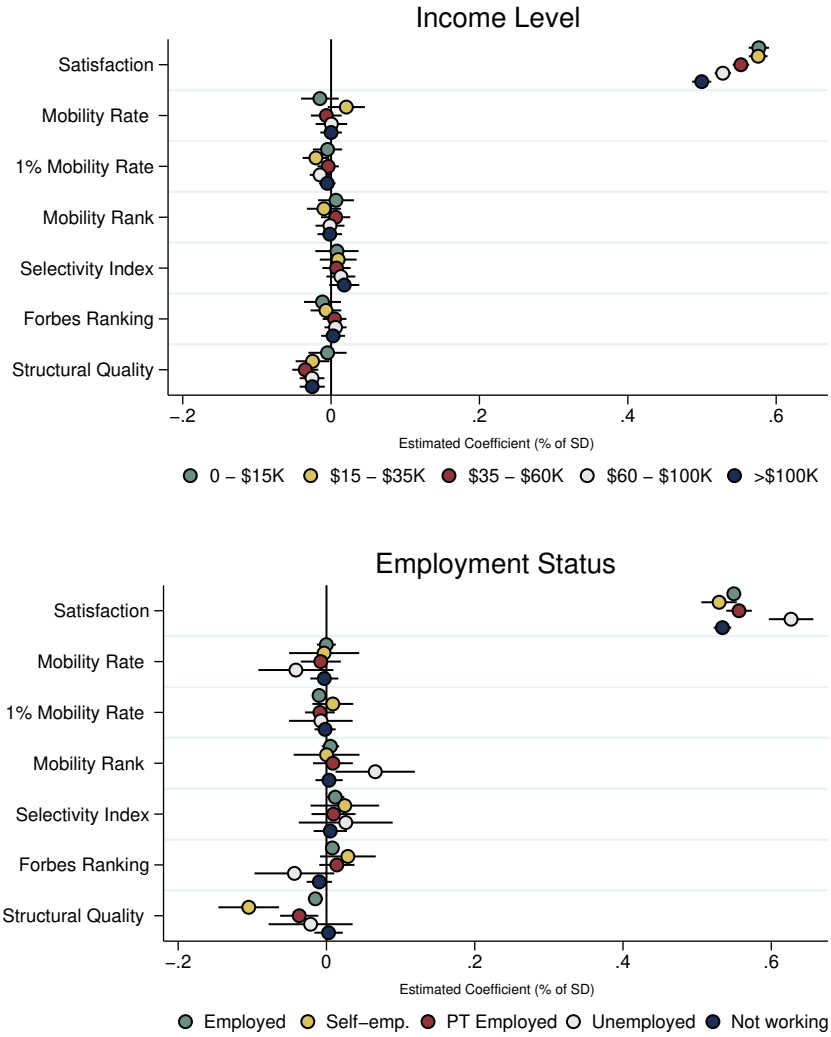


Figure D.1: *Continued.*

Note: The figures chart results from estimating a set of models where willingness to recommend is regressed on a set of college quality metrics. Displayed are the resulting standardized point estimates for each metric, with 95% confidence intervals. We run separate models for each of subsamples indicated in each figure. *Income Category* and *Employment Status* are categorical variables indicating the respondent's self-reported, current labor market status. For the latter, *Not working* means that the respondent does not consider themselves part of the work force.