ifo BEITRÄGE
zur Wirtschaftsforschung

96
2021

Human Capital and Education Policy: Evidence from Survey Data

Elisabeth Grewenig

ifo INSTITUTE
Leibniz Institute for Economic Research
at the University of Munich

ifo BEITRÄGE zur Wirtschaftsforschung

96
2021

Human Capital and Education Policy: Evidence from Survey Data

Elisabeth Grewenig

Herausgeber der Reihe: Clemens Fuest
Schriftleitung: Chang Woon Nam

ifo INSTITUT
Leibniz-Institut für Wirtschaftsforschung
an der Universität München e.V.

Bibliografische Information der Deutschen Nationalbibliothek

Die Deutsche Nationalbibliothek verzeichnet diese Publikation
in der Deutschen Nationalbibliografie; detaillierte bibliografische
Daten sind im Internet über
http://dnb.d-nb.de
abrufbar.

ISBN: 978-3-95942-099-0

Druck: ifo Institut, München

ifo Institut im Internet:
https://www.ifo.de

Preface

Elisabeth Grewenig prepared this study while she was working at the Center for Economics of Education at the ifo Institute. The study was completed in March 2021 and accepted as doctoral thesis by the Department of Economics at the LMU Munich. It consists of five distinct empirical essays that address various aspects of human capital formation and education policy. Chapters 2 and 3 are concerned with the determinants of human capital formation. In particular, chapter 2 investigates the impact of gender norms on labor-supply expectations of adolescents. Chapter 3 analyzes the effects of the Corona-induced school closures on students' time spent with different educational activities. Chapters 4 and 5 are concerned with the implementation and feasibility of educational reforms. Thereby, chapter 4 evaluates the impact of recent reforms on binding teacher recommendations by studying educational outcomes of students in primary and secondary schools. Chapter 5 examines whether support for educational policies is amenable to information provision about party-positions. Finally, chapter 6 contributes to the methodological debate around survey measurement by investigating belief elicitation in large-scale online surveys.

Keywords:	Gender Norms, Female Labor Supply, Survey Experiments, Educational Inequality, COVID-19, Low-Achieving Students, Home Schooling, Distance Teaching, School Tracking, Admission Policies, Student Performance, Political Parties, Partisanship, Information, Endogenous Preferences, Voters, Family Policy, Beliefs, Incentives, Online Search
JEL-No:	C83, C90, C93, D30, D72, D83, H52, I21, I24, J13, J16, J22, J24, J62, P16

Acknowledgements

First and foremost, I would like to thank my advisor Ludger Woessmann for supporting and challenging me, for giving me precious feedback, and for valuing my opinion in countless discussions. It has been a pleasure working together with you. Furthermore, I am very grateful to Joachim Winter and Andreas Peichl for their valuable support and feedback.

I am deeply indebted to Philipp Lergetporer and Katharina Werner. I have enjoyed every meeting, discussion and collaboration with you. In addition to being skilled and cherished colleagues, I also consider you my friends. Thank you for your constant support since my first day at the ifo Institute. I also thank Larissa Zierow for great and reliable co-authorship. I appreciate that your door is always open for guidance and advice.

During the academic year 2018/2019, I had the incredible opportunity to visit Harvard University. I am most grateful to the whole team of PEPG (Program on Education Policy and Government) who took me in as one of their own. I tremendously benefited from numerous discussions and conversations. I am also thankful for the generous financial support from the Bernt Rohrer Foundation for this research stay.

I am grateful to all colleagues at the ifo Center for the Economics of Education for their great company, moral support and many laughs during coffee, lunch or other meetings. In particular, I would like to thank Ulrike Baldi-Cohrs and Franziska Binder for being the kind souls of our Center. Many thanks to Vera Freundl, Sarah Kersten and Franziska Kugler for helping to collect most of the data in this dissertation.

I thank all my amazing peers and friends who did not only accompany my academic journey, but have also been there for me in countless everyday situations. Special thanks to Franziska Hampf who has been the best office neighbor I could have asked for and to Eleonora Guarnieri who made the job market bearable. I am particularly grateful to Stefanie Gäbler, Lea Immel and Feodora Teti for listening to the endless ups and downs of my life as a researcher and making me forget the downsides so easily. I have already planned a next 'Wine Evening' to show my appreciation.

Last, but not least, I would like to express my deepest gratitude to my parents for always loving and supporting me. You are great role models and I am incredibly thankful for your constant faith in me.

Elisabeth Grewenig
March 2021

Human Capital and Education Policy: Evidence from Survey Data

Inaugural-Dissertation
Zur Erlangung des Grades
Doctor oeconomiae publicae (Dr. oec. publ.)

eingereicht an der
Ludwig-Maximilians-Universität München
2021

vorgelegt von

Elisabeth Grewenig

Referent: Prof. Dr. Ludger Wößmann
Korreferent: Prof. Dr. Joachim Winter
Promotionsabschlussberatung: 14.07.2021

Datum der mündlichen Prüfung:	30.06.2021
Namen der Berichterstatter:	Prof. Dr. Ludger Wößmann
	Prof. Dr. Joachim Winter
	Prof. Dr. Andreas Peichl

Contents

List of Figures

List of Tables

1 Introduction

A long-standing literature in economics underscores the relevance of education for individual labor-market success as well as for the prosperity of the economy as a whole (Barro, 2001; Hanushek and Woessmann, 2008; Acemoglu and Autor, 2010; Hanushek and Woessmann, 2015). Starting with the seminal work on human capital formation by Mincer (1958), Schultz (1961) and Becker (1962), education has become a major area of research in the economics field (Hanushek, 2002). The continuously thriving body of research provides ample evidence that education improves a broad range of pecuniary and non-pecuniary outcomes (see, e.g. Card, 1999, Psacharopoulos and Patrinos, 2004 or Oreopoulos and Salvanes, 2011 for an overview of the existing work).

Given the societal relevance of education, it comes as no surprise that the successful design of education policy has become a major area of political concern. Over the last decades, policy-makers are increasingly relying on scientific research to address urgent questions about the effective implementation of education policy (e.g., Burns, 2007). Germany, for instance, has recently experienced a major shift in its political debate. In 2001, the so-called PISA shock—where the public suddenly learned that German students do not perform as well in international achievement tests as previously thought—triggered a lively discourse around evidence-based policy implementation (e.g., Ertl, 2006).

This dissertation contains five essays which revolve around various aspects of human-capital formation and education policy. The remainder of the introduction is structured as follows: Section 1.1 introduces the human-capital theory and highlights individual and societal determinants of educational outcomes. Section 1.2 discusses the importance of education policy and how it relates to public policy preferences. Section 1.3 highlights the merits of using (self-collected) survey data to study central issues in education economics. Finally, section 1.4 provides an outline of each chapter.

1.1 Human Capital Formation

In two seminal contributions, Schultz (1961) and Becker (1962) were the first to formalize the idea that individuals invest in education to accumulate human capital, known as the *Human Capital Theory*. Thereby, human capital is equivalent to the set of skills and characteristics that contribute to individual productivity. In essence, the theory posits that individuals weigh potential costs and benefits when deciding upon their educational investments. On the one hand, potential costs are ’indirect’ opportunity costs in form of forgone wages as well as more ’direct’ costs, such as tuition fees for schools or universities. On the other hand, potential benefits are improved labor market prospects, such as higher wages or lower risk for unemployment (Card, 1999, 2001; Psacharopoulos and Patrinos, 2004; Heckman et al., 2006) as well as other non-pecuniary benefits such as better health or life-expectancy (Currie and Moretti, 2003; Kemptner et al., 2011; Oreopoulos and Salvanes, 2011). Utility-maximizing agents differ in their final investments decisions because they face different costs and benefits associated with education. Similarly, individuals may also discount future returns at different rates.

To better understand determinants of human-capital investments, a recent strand of literature exploits subjective expectations about future labor-market outcomes elicited among students (e.g., Manski, 2004; Arcidiacono et al., 2014; Delavande, 2014; Delavande and Zafar, 2018; Wiswall and Zafar, 2018). The underlying idea is that expectations are a crucial predictor for educational choices and attainment as students with lower expectations have also smaller incentives to invest into education (Beaman et al., 2012; Stinebrickner and Stinebrickner, 2013; Reuben et al., 2017). Findings from this literature suggest that classical 'career concerns' (Wiswall and Zafar, 2015, 2018)—e.g., expected earnings (growth), promotion prospects or job attributes—as well as a number of 'family life' dimensions (Wiswall and Zafar, 2020)—e.g., spouse's earnings or fertility—are major drivers for human-capital investments.

Besides individual factors, cultural components play also an important role for human-capital formation (e.g., Hanushek et al., 2020). Along these lines, several contributions have shown that particularly female migrants—who are living in the same country, but have been socialized under a different culture than natives—show different educational (e.g., Nollenberger et al., 2016; Friedman-Sokuler and Justman, 2020), labor-market (e.g., Fortin, 2005; Fernández and Fogli, 2009), or fertility (e.g., Fernández and Fogli, 2006) outcomes than their native counterparts. While it is often difficult to disentangle which cultural components cause the emerging differences, researchers commonly assume diverging *gender norms* to be one of the main explanatory factors (e.g., Guiso et al., 2008; Bertrand, 2011; Blau and Kahn, 2017). Chapter 2 of this dissertation takes a direct approach to investigate the effects of perceived gender norms on labor-supply expectations among adolescents in Germany. It thereby contributes to a better understanding of the role of social norms for human-capital formation.

Governments with their public school systems are by far the most important provider of education. Many advanced societies have implemented compulsory schooling which aim at endowing all children with some basic amount of skills. As such, schools play a particularly important role for human-capital formation. The literature has estimated that each year of schooling increases earnings by approximately 10 percent (Card, 1999; Harmon et al., 2003; Psacharopoulos and Patrinos, 2004; Heckman et al., 2006; Gunderson and Oreopolous, 2020). It is thus not surprising that longer periods of school closures exert devastating effects on students' skill formation and subsequent labor-market outcomes (see e.g., Belot and Webbink, 2010; Baker, 2013; Jaume and Willén, 2019, on teacher strikes). Further evidence on the German short-schooling years[1] suggests that even a pre-planned shortening can negatively impact students throughout their entire lives (Cygan-Rehm, 2018). In 2020, the Corona pandemic lead to massive school closures, involving more than 1.5 billion school children worldwide (UNESCO, 2020b). Several contemporaneous studies already demonstrate that these closures massively affect learning inputs and outputs, such as online learning (Chetty et al., 2020; Bacher-Hicks et al., 2021) or student performance in standardized tests (Engzell et al., 2020; Maldonado and De Witte, 2020). Chapter 3 of this dissertation scrutinizes how the Corona-induced school closures affected learning time among students in Germany.

1.2 Education Policy

Given the importance of public schools for students' future labor-market success, the question on how to improve school systems is a major concern of policy-makers around the globe. Thereby, one of the main goals is to enhance student achievement, especially since international student assessments have revealed substantial achievement differences across countries (Woessmann, 2016). Some policies target the allocation of school resources, such as class size (Angrist and Lavy, 1999; Krueger, 1999; Hoxby, 2000) instruction time (Lavy, 2015; Andrietti and Su, 2018), or teacher quality (Chetty et al., 2014; Hanushek et al., 2018). Other polices aim at improving the institutional structure of the schooling systems, including autonomy (Hanushek and Woessmann, 2011; Hanushek et al., 2013), accountability (Bergbauer et al., 2018) or tracking (Hanushek and Woessmann, 2006). Particularly, the postponement of school tracking has proven as efficient policy tool to mitigate educational inequality in the long run (see, Meghir and Palme, 2005; Pekkala Kerr et al., 2013; Matthewes, 2020). Chapter 4 of this dissertation contributes to the literature on school tracking by examining an institutional feature, namely binding teacher recommendations, within the tracking procedure in Germany.

Because the allocation of resources to schools and their institutional structure are decided by the political process, the outcome of democratic elections plays a considerable role in shaping education policy. This implies that the feasibility of reforms depends heavily on the electorate's policy preferences (Busemeyer et al., 2018). Consequently, public opinion towards various education policies has moved into the focus of economic research (Bursztyn, 2016; Lergetporer et al., 2018a; Cattaneo et al., 2020; Lergetporer et al., 2020). Findings generally suggest that information campaigns can substantially impact policy preferences (Cruces et al., 2013; Kuziemko et al., 2015; Alesina et al., 2018b; Haaland and Roth, 2020). Chapter 5 extends the literature on public preferences towards education policies by showing that the effects of information provision are not limited to information about facts which underlie the policy itself (e.g., the effect of informing about current educational spending levels). It particularly

[1] To harmonize the start of school years throughout Germany, many German states introduced two short-school years in 1966/1967. Hampf (2019) shows that affected students indeed experienced eight months less schooling.

investigates whether policy preferences are amenable to information on how the different parties stand towards the issue.

1.3 Using Survey Data for Economic Research

Common to all dissertation chapters is their exploitation of large-scale survey data. Over the past years, surveys have become increasingly popular in economics since they offer several advantages: First, surveys provide the unique opportunity to disclose information that can otherwise not be observed in administrative data. By using simple survey questions, researchers can, for instance, elicit knowledge, perceptions, opinions or views. When it comes to peoples' preferences, economists often painstakingly back out revealed preference measures from data on observed choices. But since realized choices also entail the constraints that people face when making decisions (Manski, 2004), they usually do not provide unbiased measures for individual preferences. Surveys, on the contrary, facilitate measuring unconstrained preferences through posing simple survey questions. Chapters 2 and 6 exploit survey data on subjective beliefs and expectations, chapter 5 examines survey data on policy preferences and chapters 3 and 4 analyze survey data on detailed time-use information. All these outcomes can typically be not observed in administrative data.

Second, surveys allow to customize implementation to the researchers' needs. Methodologically, the implementation of so-called *survey experiments* has become particularly popular among economists (Haaland et al., 2020). The idea of those experiments is that randomly chosen subsets of participants are provided with different versions of the same question. This allows researchers, for instance, to study how people form beliefs, preferences or, how they make choices in a controllable environment that cleanly identifies the causal effect of the induced variation. Following this trend, chapters 2, 5 and 6 of this thesis exploit survey experiments for identification.

Third, particularly online surveys can offer access to relatively diverse sets of potential study participants. This can prove useful, for instance, to the experimental literature which traditionally analyzes small university-student samples. Conducting (survey) experiments among representative population samples allows to speak to the external validity of findings. Moreover, if desired, surveys can target specific sub-groups of the population who are most relevant for answering the underlying question of interest. Chapters 5 and 6 draw on samples representative for the general German population. In contrast, chapters 2, 3 and 4 analyze survey information elicited for children and adolescents who constitute a highly relevant study group in the education context.

In sum, (online) surveys—as exploited throughout this dissertation—are a rigorous research tool that enables scientists to collect data difficult to gather otherwise. As such, they allow to address important and novel research questions. At the same time, surveys can only deliver their full potential and benefits if they are carefully designed and calibrated. Therefore, chapter

6 contributes to the methodological debate around survey implementation by investigating belief elicitation in large-scale online surveys.

1.4 Chapter Overview and Relevance

This dissertation consists of five empirical essays investigating various aspects of human-capital formation and education policy. Each essay corresponds to one chapter, is self-contained, and can be read independently. Chapters 2 and 3 investigate the importance of the social environment and public schools for human-capital formation. Chapters 4 and 5 are concerned with the implementation and feasibility of education reforms. Chapter 6 contributes to the methodological debate around survey measures by analyzing the effects of incentivizing accuracy of subjective beliefs. This section shortly summarizes the content of each chapter, followed by a brief discussion on how the respective chapter contributes to the economic and political discourse.

Chapter 2 investigates a causal link between perceived gender norms and labor-supply expectations among German adolescents. A recent literature suggests that these expectations are an important predictor for educational choices and attainment: Although most adolescents are not yet on the labor-market, they already face important labor-market choices—e.g., educational or occupational choices—which may be affected by their expectations even prior to labor-market entry. The starting point of the chapter is the observation that gender gaps in labor-market outcomes often exacerbate with the arrival of the first child. We design and run a large-scale online survey to experimentally study the effects of perceived gender norms on labor-supply expectations. Using a hypothetical scenario, we document that most girls expect to work 20 hours or less per week when having a young child. Conversely, most boys expect to work 30 hours or more. We then randomly administer treatments that inform about the fact that 91 percent of Germans hold the opinion that mothers should reduce their labor supply while only 41 percent hold the opinion that fathers should do so. First, we find that girls largely underestimate the share of Germans who hold the opinion that mothers should reduce their labor supply and, consistently, girls significantly reduce their self-expected labor supply in response to the treatments. Second, we find that boys underestimate the share of Germans who hold the opinion that fathers should reduce their labor supply and, consistently, boys also reduce their labor-supply expectations in response to the treatments. Overall, these findings suggest that (perceived) gender norms can play an important role in shaping gender gaps in outcomes relevant to the labor market.

From a normative policy perspective, the results highlight that changing how adolescents perceive gender norms may be a promising approach to foster gender equality on the labor market. This may be achieved through information campaigns or by changing how men and women or mothers and fathers are portrayed in school books, text books, or advertisements.

Chapter 3 investigates the impact of Corona-induced school closures on students' time spent with studying and other leisure activities. In 2020, governments around the globe shut down schools to mitigate the spread of the novel coronavirus. The chapter argues that low-achieving students may be particularly affected by the lack of educator support during school closures. To test this hypothesis, we collect detailed time-use information on school children before and during the COVID-19 school closures in a survey of over 1,000 parents in Germany. We find that while children reduced learning time by about half on average, the reduction was significantly larger for low-achieving students (4.1 hours per day) than for high-achieving students (3.7 hours). Especially low-achieving students substituted learning time mainly for detrimental activities such as TV or computer games rather than for conducive activities. We also find that the learning gap was not compensated by parents or schools who provided less support for low-achieving students.

From a policy perspective, these results call for universal and binding distance-teaching concepts for school closures. Since it is particularly the low-achieving students who suffer when support of teachers is lacking, any attempt to encourage their learning when schools have to close is likely to reduce future educational inequality.

Chapter 4 evaluates a particular aspect of the German tracking procedure, namely binding teacher recommendations, on students' educational outcomes. Although a large literature in economics investigates the effects of general school tracking on later life outcomes, only little is known about the impact of institutional features within the tracking procedure. Depending on the federal state in Germany, either teachers or parents have the discretion to decide on the highest secondary school track a child may transit to after primary school (grade 4). Applying a differences-in-differences approach, this chapter exploits variation in the implementation and abolition of binding teacher recommendations—which withdraw free choice of secondary school tracks—across states and over time to investigate its effects on students' academic outcomes. Using data from Germany-wide large-scale skill assessments, I show that binding teacher recommendations significantly improve student achievement in fourth grade, prior to track assignment. Effects persist into ninth grade, after consequential track assignment. Further analyses suggest that effects are driven by increased time investments in students' skill development. Overall, binding teacher recommendations thus lead to persistent improvements in students' educational outcomes in the short and medium run.

These findings have important implications for the scientific and political discourse. While the economic literature has mostly focused on the effects of earlier vs. later tracking, this chapter shows direct evidence that institutional features within the tracking procedure are also important for the formation of human capital. Furthermore, the political debate around binding recommendations has mostly revolved around the normative argument that broad populations should be granted access to academic schools. Consequently, the most recent reforms have abolished binding recommendations and guaranteed children and parents free choice of secondary schools. My results, however, suggest the opposite: Free parental choice reduces academic school attendance in the medium run and can potentially harm students' academic performances.

Chapter 5 examines whether policy preferences towards two specific German family policies are amenable to information about the policy positions of political parties. The standard assumption of exogenous policy preferences implies that parties set their positions according to their voters' preferences. Focusing on family policies, this chapter investigates the reverse effect, namely whether the electorates' policy preferences are responsive to party positions. In a representative German survey, we inform randomized treatment groups about the positions of political parties on two disputed policies, child care subsidy and universal student aid. In both experiments, the information treatment aligns policy preferences of specific partisan groups with their preferred party's position, implying endogenous policy preferences. The treatment also affects non-partisan swing voters, suggesting that party positioning can affect the public's preferences beyond their partisans.

These findings bear implications for economic and political theory. The common assumption of the exogeneity of public policy preferences does not hold for the policies studied in this chapter. Therefore, the results highlight the need for a more extensive consideration of potential endogeneities of preferences in the literature. Relatedly, the findings imply the risk of increased polarization among the public if parties take extreme positions. Furthermore, the results also have implications for policy making and politics. Since broad public support is often decisive for successful policy implementation, the mere communication of party positions (even without putting forward any substantive arguments) can be important for the political feasibility of reform proposals.

Chapter 6 investigates the impact of incentivizing belief accuracy on stated beliefs of survey respondents. Measuring people's subjective beliefs about economic facts is essential for understanding economic behavior and choices. To elicit such beliefs, economists often rely on survey questions which do not provide respondents with incentives for accurate answers. This raises concerns of systematic biases in unincentivized belief measures that might stem from lack of cognitive effort invested in truthful reporting or from socially desirable and self-serving answering behavior. In this chapter, we devise randomized experiments in a representative online survey to investigate whether incentivizing belief accuracy affects stated beliefs about average earnings by professional degree and average public school spending. Incentive provision does not impact earnings beliefs, but improves school-spending beliefs. Response spikes suggest that the latter effect likely reflects increased online-search activity. Consistently, an experiment that just encourages search-engine usage produces very similar results. We draw two main conclusions from our analyses: First, unincentivized belief measures do not necessarily suffer from systematic reporting bias. Second, providing monetary incentives in online surveys might increase respondents' use of external resources such as online-search engines to improve the accuracy of their stated beliefs.

These results point to a trade-off that researchers face when deciding upon whether to incentivize beliefs about economic facts. On the one hand, it might be undesirable for many research questions related to subjective beliefs that respondents consult external resources in response to incentive provision. On the other hand, there can be clear advantages when incentivizing beliefs, such as respondents thinking more carefully about their responses. Similarly, incentive provision could be an interesting research approach to induce belief updating in an unobtrusive way.

Taken together, this dissertation highlights once more the importance of education policy for fostering human-capital formation. To ensure a successful education of its citizens, governments need to constantly re-think their education-policy decisions. In particular, changing situations, such as the evolution of social norms or the occurrences of global crises, call for a flexible adjustment of government actions. By taking thought leadership, policy-makers can also directly guide public support which is essential for reform feasibility.

2 Gender Norms and Labor-Supply Expectations: Experimental Evidence from Adolescents*

2.1 Introduction

The birth of the first child has large and persistent negative effects on labor-market outcomes of women, but not of men. This finding holds for different countries and over time (Angelov et al., 2015; Kuziemko et al., 2018; Kleven et al., 2019b). Estimates of so-called *child penalties*—the impact of children on earnings of women relative to men—range from 20–25 percent in Scandinavia to 30–40 percent in Anglo-Saxon countries, and 40–60 percent in German-speaking countries (Kleven and Landais, 2017; Kleven et al., 2019a). In fact, the arrival of children is one of the primary reasons for persistent gender inequalities on the labor market (Kleven et al., 2019b). While the disadvantages in the labor market due to childbirth for women compared to men are well documented, only little is known about the underlying causes. In this chapter, we argue that gender norms concerning parental labor supply can cause gender differences in outcomes relevant to the labor market.

From a theoretical perspective, such norms may encourage women and men to adjust their labor-market choices to what seems socially appropriate for mothers and fathers to do and in this way produce gender gaps on the labor market (e.g., Akerlof and Kranton, 2000, 2010; Bertrand et al., 2015; Cortes and Pan, 2020). Indeed, empirical studies show that existing social norms towards maternal labor supply correlate strongly with child penalties across countries (Steinhauer, 2018; Kleven et al., 2019b). Yet, very little is known about the causal relationship between social norms and labor-market outcomes. We therefore run large-scale online survey experiments that introduce exogenous variation in the salience of, and perceptions about, existing gender norms to study how they affect labor-supply expectations.

Our sample consists of 2,000 German adolescents aged between 14 and 17 years. As in many other developed countries, social norms concerning women in general are relatively gender-equal in Germany, but those directly addressing how mothers should behave on the labor market are still very traditional. We focus on adolescents as they already face important, labor-market relevant decisions such as educational or occupational choices. These choices may be affected by labor-supply expectations even prior to labor-market entry. Moreover, understanding the role of social norms is particularly relevant for adolescents, who are in a key phase of gender-differential socialization (e.g., Hill and Lynch, 1983; Priess et al., 2009) that may lay the foundation for later gender inequalities.

* This chapter is joint work with Philipp Lergetporer and Katharina Werner. It is based on the paper 'Gender Norms and Labor-Supply Expectations: Experimental Evidence from Adolescents', *CESifo Working Paper*, 2020.

We measure labor-supply expectations by presenting adolescents a hypothetical scenario, in which we ask them to imagine themselves at the age of 30 having a child. We then ask respondents about labor-supply expectations for themselves and for their hypothetical partner, allowing us not only to investigate respondents' self-expected labor supply, but also expected labor-supply differences between them and their partner (i.e., the within-family gender gap). The fact that most adolescents are not yet in the labor market allows us to study expected labor-supply before any actual demand-side restrictions on the labor market are likely to become relevant.[1] We follow a long tradition in economics that studies subjective expectations and preferences concerning the labor market and other domains (see e.g., Manski, 2004; Delavande, 2014; Wiswall and Zafar, 2018). To understand the role of gender norms in labor-market decisions, preferences and expectations about future labor supply are an important and relevant outcome to study because realized labor-market choices can be a result of many different combinations of beliefs (e.g., about gender norms), preferences, and labor-market conditions. By obtaining direct measures of expectations, we isolate the effect of social norms on expected future labor supply.

Investigating the causal link between gender norms and labor-supply expectations using observational data is challenging: Cultural traits and (gender) norms usually persist over a long period of time (e.g., Cotter et al., 2011; Alesina et al., 2013), which makes it unlikely to find exogenous variation that would facilitate establishing a causal effect of gender norms.[2] Focusing on the existing social norm prescribing how much mothers and fathers of young children should work on the labor market,[3] we sidestep potential identification challenges by experimentally varying two important aspects of social norms: their salience (e.g., Aloud et al., 2020) and perceptions about their exact content (e.g., Bursztyn et al., 2020). To study the complementarity between both aspects, we consider three treatments that augment the norm's salience, information, and both: The first treatment *salience* asks respondents to guess the share of German adults who hold the opinion that mothers and fathers, respectively, should reduce their labor-market supply while their children are young. These belief-elicitation questions prime respondents to think of the gender roles for mothers and fathers. The second

[1] In the hypothetical scenario, we explicitly ask respondents how many hours per week they would like to work. This gives us a measure for the adolescents' supply-side intentions that abstracts from respondents' assumptions about the equilibrium mechanisms for the allocation of labor or preferences of the employees. We refer to this measure as 'labor-supply expectation' throughout the chapter, but we do not mean to use this term to imply that we have elicited probabilistic expectations or the like.

[2] Some papers scrutinize migration streams to investigate the effects of culture on economic outcomes of both, migrants and natives. While such approaches may be feasible to estimate causal effects of the whole culture in which individuals grew up—including gender norms—on outcomes, they merely disentangle the effects of one specific social norm from the whole set of other characteristics (e.g., preferences, beliefs or values).

[3] We focus on *injunctive*, or *prescriptive*, social norms (i.e., what behavior is commonly approved of by society), but not on *descriptive* social norms (i.e., what most members of society actually do) (e.g., Cialdini and Trost, 1998; Benabou and Tirole, 2011). While injunctive social norms convey directly what an individual ought to do in a certain situation, descriptive social norms also reflect factors outside of an individual's control (e.g., labor-demand constraints), which renders the interpretation of descriptive-norms effects unclear (see Goerges and Nosenzo (2020) for a recent discussion).

treatment *information* informs respondents about a representative study showing that 91 percent of Germans think that mothers should reduce their labor supply while their children are young, and 41 percent think that fathers should do so (Wippermann, 2015).[4] The third treatment, *salience and information*, combines both treatments by first posing the belief-elicitation question from treatment *salience* and afterwards providing correct information about the social norm towards mothers and fathers from treatment *information*. After treatment, all respondents answer the same questions about their labor-supply expectations as the untreated control group.

In the control group, we already find a large gender gap in self-expected labor supply: Most girls (59 percent) expect to work 20 hours or less per week while most boys (70 percent) expect to work 30 hours or more, which translates into a gender difference of 7.3 work hours per week. To study expected within-family gender gaps in parental labor supply, we exploit the fact that we also elicited respondents' labor-supply expectations for their partner. At baseline, girls expect to work 7.7 hours less than their partner, and, conversely, boys expect to work 10.9 hours more than their partner. Moreover, adolescents hold biased beliefs on the content of the existing gender norm (measured in treatments *salience* and *salience and information*): They largely underestimate the share of German adults thinking that mothers should reduce their labor supply (average belief: 66 percent; true value: 91 percent), and they also underestimate the share thinking that fathers should do so (average belief: 35 percent; true value: 41 percent). These misperceptions indicate leeway for correcting inaccurate beliefs through our information treatments.

Among girls, all three treatments reduce labor-supply expectations. Treatment *salience* significantly reduces their self-expected labor supply by 2.0 hours per week. Treatment *information* reduces labor-supply expectations by 2.6, and treatment *salience and information* by 3.4 hours per week. We draw the following conclusions from this treatment-effect pattern: First, the strong effects of treatment *salience* suggest that priming adolescents to think about the existing gender norm already alters their labor-supply expectations. Second, the effect of the combined treatment *salience and information* is significantly larger ($p<0.1$) than the effect of treatment salience, suggesting that providing information about the norm's content has an additional effect beyond the treatment *salience*. Third, the effect of the combined treatment *salience and information* is smaller than the sum of the two separate effects of treatment *salience* and treatment *information*, suggesting that part of the information effect operates through increasing the norm's salience. Overall, reductions in self-expectations translate into expected within-family gender gaps in labor supply that are more gender-unequal: On average, the treatments induce girls to reduce their expectations about their own labor supply by 2.2

[4] Data from the International Social Survey Program (ISSP) 2012 confirms that traditional views concerning maternal labor supply are not unique to Germany. In fact, in many developed countries—including more gender-egalitarian Scandinavian countries—, most residents think that women with children under school age should work at most part-time (see section 2.2.2 for details).

hours compared to their partners' labor supply, thereby significantly increasing expected within-family gender gaps.

Next, we investigate how the gender-norm treatments affects boys, a question that—despite its relevance—received much less attention in the public and scientific discourse than the effects on girls. Again, all three treatments reduce self-expected labor supply by 1.3 hours in treatment *salience*, 1.6 hours in treatment *information* and 2.5 hours in treatment *salience and information*. Contrary to the findings on girls, boys' expectations for their own family become more gender-equal in response to the treatments:[5] On average, the treatments cause boys to reduce their labor-supply expectations by 1.4 hours more than those for their partner. This leads to a reduction of their expected within-family gender gap.

Leveraging our follow-up survey about two weeks after the main survey, we investigate whether information-treatment effects persist beyond the immediate survey horizon.[6] Focusing on treatments *information* and *salience and information*, we find that the information treatments persistently affect both labor-supply expectations and beliefs about the content of the norm. The persistence of treatment effects suggests that they are not due to experimenter-demand effects (Haaland et al., 2020).

We then turn to analyzing the mechanisms underlying our treatment effects. For this purpose, we estimate heterogeneous information-treatment effects by respondents' prior beliefs about the norm's content within the sample of adolescents who received the prior belief elicitation question (treatments *salience* and *salience and information*).[7] The heterogeneity analysis yields two findings: First, treatment effects are already prevalent among respondents with accurate priors, suggesting that information effects are at least partly driven by salience-based information updating (Bleemer and Zafar, 2018). Second, treatment effects are larger in absolute terms (albeit not significant at conventional levels) the more respondents underestimate the content of the social norm, as would be expected for information-based updating.[8]

We subject our main findings to the following robustness checks: First, we address a concern frequently raised against unincentivized expectations measures, namely that respondents do not have any monetary incentives to provide meaningful and honest answers. To test

[5] For both girls and boys, the norm treatments change self-expectations but hardly affect the expectations for their partner.

[6] Since priming effects, like the ones induced by treatment *salience*, are by definition short-term, we focus our persistence analysis on treatments with informational content, which is standard in the information-provision literature (e.g., Haaland et al., 2020).

[7] Information provision may impact individuals because it makes the importance of gender-norm issues for labor-market participation more salient (e.g., Schwarz and Vaughn, 2002; Chetty et al., 2009; DellaVigna, 2009; Bleemer and Zafar, 2018), or because respondents were misinformed about the exact content of the social norm and update their beliefs accordingly (e.g., Rockoff et al., 2012; Bursztyn et al., 2020).

[8] The interpretation of information-based updating is also consistent with the finding that respondents who were informed about the exact content of the norm hold more accurate posterior beliefs about the share agreeing to the respective norm statement in the follow-up survey.

whether treatment effects carry over to outcomes with immediate monetary consequences, we additionally elicited beliefs about the German public's views on a set of gender-related statements posed in the European Values Study, and pay respondents for correct answers. Reassuringly, information provision strongly and significantly affects these incentivized outcomes, which shows that treatment effects are not confined to survey answers without direct monetary consequences. Second, since the traditional gender norm prescribes parental labor supply, the treatments should not necessarily affect labor-supply expectations without child. In fact, treatment effects on labor-supply expectations without child (elicited in an alternative hypothetical scenario) are small and insignificant, which shows that our results are due to a shift in those outcomes directly targeted by the treatments, and not due to a general shift in expectations.

Finally, appreciating the fact that various gender norms prescribing behavior in different domains coexist in society, we investigate whether highlighting a more gender-egalitarian norm can shift outcomes in the opposite direction. We therefore conduct a second experiment in the follow-up survey. The corresponding treatments experimentally highlight a more gender-egalitarian norm towards sharing household responsibilities.[9] In particular, the treatments leverage the fact that 89 percent of Germans think that men should take as much responsibility for the home and children as women (European Values Study 2017). Adolescents hold downward-biased beliefs about this share (average belief: 66 percent), and correcting these false beliefs through randomized information provision significantly reduces expected within-family gender gaps in labor supply after child birth by 1.3 hours per week among girls, and by 1.4 hours per week among boys. Thus, the more egalitarian norm leads to less gender-unequal expectations.

To our knowledge, this chapter is the first to study the causal effects of salience of and information about gender norms on labor-supply expectations (see section 2.2.1 for an in-depth discussion on how we contribute to the existing literature). Our findings suggest that gender norms indeed play an important role in shaping gender gaps in outcomes relevant to the labor market. From a normative policy perspective, our results highlight that changing how adolescents perceive gender norms may be a promising approach to foster gender equality on the labor market. This may be achieved through information campaigns or by changing how men and women or mothers and fathers are portrayed in school books.

The remainder of the chapter is structured as follows: In section 2.2 we discuss how we contribute to the existing literature, and provide background information on female labor-market participation in Germany. Section 2.3 describes our dataset and the experimental design. Section 2.4 provides descriptive evidence on adolescents' labor-supply expectations.

[9] While the traditional gender norm towards mothers' and fathers' labor supply after child birth studied in the first experiment directly prescribes mothers' and fathers' labor supply, it is not the only gender norm that might be relevant for labor-supply decisions. For instance, different norms prescribe relative income within households (Bertrand et al., 2015), within-household division of work at home and on the labor market, or educational decisions (European Values Study 2017).

Section 2.5 shows the results of the first experiment that highlights social norms prescribing parental labor supply, and presents several robustness tests. Section 2.6 presents results of our second experiment on norms regarding shared household responsibilities. Section 2.7 concludes.

2.2 Background

In this section, we first discuss how this chapter relates to and extends different strands of the economic literature. Then, we provide institutional background information on female labor-market participation and gender norms in Germany.

2.2.1 Related Literature

This chapter contributes to several strands of economic research. First, it adds to the growing literature on child penalties that shows that gender gaps in labor-market outcomes often arise with the birth of the first child (e.g., Bertrand et al., 2010; Angelov et al., 2015; Kleven and Landais, 2017; Kuziemko et al., 2018; Kleven et al., 2019a,b). A small subset of this literature studies potential underlying causes of child penalties, and argues that they are not inherent in the biological relationship between mother and child (e.g., Andresen and Nix, 2019; Kleven et al., 2020).[10] Instead, factors related to socialization have been suggested as likely explanations for why the arrival of children has stronger negative labor-market impacts on women than men. The chapter works towards an understanding of whether strong social norms prescribing mothers' and fathers' labor supply–one specific but important societal factor—can explain gender gaps in parental labor supply.

Second, this chapter contributes to a larger strand of literature that studies gender gaps unrelated to parenthood. This literature argues that social norms may promote gender gaps in the labor market (for a survey of this literature see Bertrand et al., 2010; Olivetti and Petrongolo, 2016; Blau and Kahn, 2017; Jayachandran, 2019; Altonji and Blank, n.d.). In particular, several studies suggest that women's labor-market outcomes have a cultural component, which is often interpreted as indirect evidence of the importance of gender norms (e.g., Fernández et al., 2004; Fernández, 2007; Fernández and Fogli, 2009; Alesina et al., 2013; Fernández, 2013; Charles et al., 2018; Giuliano, 2018; Boelmann et al., 2020). In this respect, a few studies (e.g., Fortin, 2005; Giavazzi et al., 2013; Fortin, 2015) demonstrate direct cross-country correlations between labor-market outcomes of women and injunctive gender norms measured in large-

[10] Kleven et al. (2020) compare child penalties among biological and adoptive families and find that in both types of families, men's and women's labor-market trajectories are very similar until the arrival of the first child, and diverge with child arrival due to an abrupt and persistent negative shock on females' labor-market outcomes. Similarly, Andresen and Nix (2019) investigate child penalties among female same-sex couples that include the biological mother of the child. They find no long-term differences in labor-market outcomes between the biological mother and the 'co-mother'.

scale surveys like the World Values Survey.[11] To our knowledge, only two papers investigate such correlations in the context of labor-market outcomes of parents (Steinhauer, 2018; Kleven et al., 2019b). In contrast to these descriptive studies, our survey experiment allows us to identify the direct and causal link between salience and content of gender norms and expected labor-market outcomes.

In this sense, this chapter is closest to the study by Bursztyn et al. (2020) which finds that experimentally shifting perceived norms towards female labor-market participation in Saudi Arabia increases the willingness of married men to let their wives join the labor force. While Saudi Arabia provides an interesting case study for the role of gender norms in an extremely gender-unequal setting, we test the causal link between social norms and labor-market outcomes in Germany, a country with gender equality laws similar to those of most other developed countries.[12] Importantly, we focus on social norms and labor-market outcomes of mothers rather than women in general, which is particularly important in the context of developed countries: In these countries, overall gender gaps in labor-market outcomes have decreased over time and are now relatively small (e.g., Blau and Kahn, 2017), but gender gaps among parents are still large and persistent (Kleven et al., 2019b). Furthermore, our treatment addresses the social norm towards both mothers and fathers, which extends the existing experimental literature that has so far exclusively focused on gender norms relating to females' labor supply. Finally, we not only investigate how different aspects of gender norms affect decisions of girls regarding their own and their partner's labor supply as parents, but also of boys.

Third, this chapter uses elements from the literature that leverages subjective expectation-data to study decision-making under uncertainty (e.g., Manski, 2004; Cavallo et al., 2017; Coibion et al., 2018; Andre et al., 2019; Roth and Wohlfart, 2019). An important strand of this literature focuses on university students, and investigate expectations or preferences about family life, labor-market relevant decisions such as educational choice or investment in children (e.g., Arcidiacono et al., 2012; Cunha et al., 2013; Stinebrickner and Stinebrickner, 2013; Delavande and Zafar, 2019; Wiswall and Zafar, 2020), and sometimes also gender differences in expectations (e.g., Goldin et al., 2006; Zafar, 2013; Reuben et al., 2017; Wiswall and Zafar, 2018) elicited in different hypothetical fertility scenarios (e.g., Gong et al., forthcoming). The rationale for using expectations-data rather than realized outcomes is that observed choices can be consistent with many different combinations of beliefs and preferences (Manski, 2004), which renders the investigation of subjective beliefs and expectations highly relevant.

[11] In this literature, commonly analyzed items usually focus on women's role as caregiver vs. breadwinner (e.g., agreement/disagreement to the statements 'Being a housewife is just as fulfilling as working for pay.' or 'When a wife earns more than her husband, it is almost certain to cause problems.').

[12] Until mid–2019, Saudi Arabia had very strict 'guardianship laws' that would not allow women any actions related to work, leisure, health, finances, and law without the permission or company of a close male relative (The Economist, July 20, 2019).

Finally, the fact that we leverage the norm's salience to measure its effects on labor-market expectations is related to the literature that uses salience-treatments to prime subjects' social identities (e.g., Benjamin et al., 2010; Boschini et al., 2012) or increase salience of topics like immigration (e.g., Alesina et al., 2018a; Bleemer and Zafar, 2018; Aloud et al., 2020). For example, Aloud et al. (2020) focus on female university students in Saudi Arabia and investigate the effects of (i) priming them to think about their parents and family, and (ii) informing them about their peers' aspirations on labor-market expectations. They find that both priming and information increase expected labor-force participation. In contrast to this literature, our treatments directly address the injunctive social norm for parents' work hours.

2.2.2 Institutional Background

Although Germany has a comparatively high female labor-market-participation rate of about 56 percent, it is still around 11 percentage points below the rate of males (OECD, 2017). Large gender differences do not only exist at this extensive margin, but also at the intensive margin: Panel A of Appendix Figure A2.1 shows the share of male and female part-time employees across countries. In Germany, 37 percent (9 percent) of all employed women (men) work part-time, resulting in a gender gap in part-time employment of 28 percentage points, the largest in all observed countries. Recent research suggests that the arrival of children is one of the primary reasons for persistent gender inequalities on the labor market (Kleven et al., 2019a). Indeed, Germany exhibits the largest long-run child penalty of 61 percent among all countries observed (see Panel B of Appendix Figure A2.1).

Appendix Figure A2.2 provides direct evidence for the existence of conservative injunctive gender norms regarding the labor supply of mothers in different countries. The figure shows that 90 percent of Germans think that women with children under school age should work at most part-time (International Social Survey Program (ISSP) 2012). Comparing this share across developed countries, it is striking that (i) the norm prescribes mothers to reduce their working hours in all observed countries—including more gender-egalitarian Nordic countries—and (ii) German gender norms are among the most traditional. Among the large set of existing gender norms (e.g., prescribing the division of responsibility for the home and children, or relative within-household income shares), our main experiment focuses on the norm that directly prescribes parental labor supply, our main outcome.

Germany offers a wide range of family-friendly policies intended to support female labor supply. Since 2013, every child from the age of one has a legal claim for a public childcare place. Childcare is heavily subsidized, which implies comparably low average costs for parents of between 0 Euros and 400 Euros per month (Geis-Thoene, 2018). Parents are entitled to 12 months of paid parental leave after child birth, which can be extended to 14 months if each parent takes at least 2 months of parental leave. Parents are also eligible for unpaid and job-protected parental leave of up to 3 years for each child. Given this policy environment, our main outcomes of interest are labor-market expectations when the child is between 1 year (i.e., when the legal claim for a childcare place becomes effective, and paid parental leave ends) and 6 years old (i.e., when compulsory schooling starts) (see section 2.3.2).[13]

2.3 Data and Experimental Design

In this section, we first describe the data-collection and sampling process, and then present the experimental design, sample characteristics, and the empirical strategy.

2.3.1 Data Collection and Sample

Our online survey was conducted between October and December 2019 and covers a sample of 2,000 German adolescents aged between 14 and 17 years.[14] The main survey comprises 11 questions related to educational, career, and labor-market decisions. In addition, we elicited a rich set of sociodemographic characteristics at the end of the survey. Median completion time was 12 minutes.

Sampling and polling were carried out by the German polling firm *konkret Mafo* (`https://www.konkret-mafo.de/`) who fielded the survey via online access panels.[15] The recruitment of adolescent respondents took place in two ways: First, adolescents who were registered in the online panels were recruited directly (60 percent of our analytic sample). The remaining 40 percent were recruited indirectly via their parents who were registered in the online panels. These parents were first asked for their permission to survey their child. If the parents agreed, they received a survey link to be shared with their child.[16] All respondents answered the questionnaire autonomously on their own digital devices.

To test the persistence of treatment effects, we implemented a follow-up survey about two weeks after the main survey. The follow-up survey re-elicited some outcomes without repeating any treatments from the main survey, and included the second experiment on the effects of a more gender-egalitarian norm towards the end (see section 2.3.2 for details).

[13] Childcare take-up after the child's first birthday is 33 percent for one-year olds, and 66 percent for two-year olds (Alt et al., 2017). Besides factors like childcare-slot shortages, social norms towards maternal labor supply have been discussed as a potential reason for non-take up (e.g., Jessen et al., 2020).

[14] Our experimental setup is based on a short pilot experiment that the we conducted within the scope of the ifo Education Survey 2018 (see Online Appendix for further details).

[15] Throughout the chapter, we present unweighted analyses that assign equal weights to each respondent. It is reassuring that re-weighting observations to match official statistics with respect to gender, age, sate of residence and municipality size does not affect our qualitative results (results available upon request). In the context of adult samples, Grewenig et al. (2018) show that online surveys represent the overall population (online and offline) well.

[16] To ensure that the children and not their parents answered the survey, we incorporated several plausibility checks of age and birth date. In case of failure to provide consistent answers, respondents were exited from the survey. Importantly, treatment effects are prevalent among respondents recruited in both modes (results available upon request).

2.3.2 Experimental Design

Main Survey

The main survey consists of three stages. In stage one, respondents were randomly assigned to a control group or to one of three social-norm treatments (treatment *salience*, *information* or *salience and information*). Stage two elicited labor-supply expectations, our main outcome of interest. In stage three, we asked additional questions, e.g., incentivized outcome questions.

Treatments: Before eliciting outcomes, respondents were randomly assigned to one of four experimental groups with equal probability. One group is the untreated control group. The other three groups receive different norm treatments that emphasize the existing social norm related to parents' labor supply. Unlike previous experimental studies, our treatments emphasize the social prescriptions towards mothers and fathers, not only mothers.[17]

Treatment *salience* used a belief-elicitation question to prime respondents with gender-norm considerations.[18] Before stating the outcomes of interest, treated respondents were asked: '*What do you think, how many adults in Germany hold the opinion that mothers and fathers, respectively, should reduce their labor supply while the children are young? We do not think of the first months after child birth, but the time thereafter.*' Respondents were provided with two open answer fields, one for mothers, and one for fathers (see Panel A of Appendix Figure A2.3 for a screenshot). This treatment was designed to make the social norm salient without providing information about the norm's content.

The second treatment *information* did not elicit beliefs, but instead provided respondents with information about the share of German adults who hold the opinion that mothers and fathers of young children should reduce their labor supply. We drew on results from a representative study by the *Federal Ministry for Family Affairs, Senior Citizens, Women and Youth* fielded in 2015 (Wippermann, 2015): *'Out of 100 adults in Germany, 91 hold the opinion that the mother should reduce her labor supply while the children are young. At the same time, out of 100 adults in Germany, 41 hold the opinion that the father should reduce his labor supply while the children are young.'* Reassuringly, the norm concerning mothers' labor supply in Germany collected by the ISSP in the year 2012 are practically identical to the one of Wippermann (2015) that we use (see section 2.2.2), which indicates the robustness and persistence of the norm.[19] Along

[17] Our gender-bifocal treatment is in contrast to most of previous empirical studies on gender norms in the sense that this literature often exclusively studies norms concerning women (e.g., Fernández and Fogli, 2009; Alesina et al., 2013; Aloud et al., 2020; Bursztyn et al., 2020). To understand the wholistic impact of gender norms on parental labor-market participation and associated gender gaps, we find it crucial to examine norms that also prescribe the behavior of men.

[18] Other examples of studies that use belief-elicitation questions to increase salience are Alesina et al. (2018a) and Aloud et al. (2020).

[19] Since gender norms persist over time (e.g., Cotter et al., 2011), the fact that norms where elicited in 2015, whereas our experiment was conducted four years later in 2019 should not yield major inaccuracies. A necessary condition for treatment information to affect adolescents' expectations is that the adults' opinions highlighted in

with the verbal statement, respondents were shown a graphical illustration of the provided information (see Panel B of Appendix Figure A2.3 for a screenshot).

The treatment *salience and information* combines both treatments: Respondents were first asked the belief-elicitation questions as in treatment *salience*, and were then provided with the factual information about the norm as in treatment *information*. This treatment allows us to study the complementarity between salience and information provision.

Labor-supply expectations: Our main outcomes of interest are the respondents' labor-supply expectations that we elicited after treatment administration. In a hypothetical scenario, respondents were asked to imagine being 30 years old, living with their partner and having a child aged between 1 and 6 years. Our objective was to abstract as much as possible from adolescents' assumptions about the equilibrium mechanisms for the allocation of labor, or preferences of the employees to obtain a meaningful measure for the adolescents' supply-side intentions. After presenting the hypothetical scenario, we therefore elicit labor-supply expectations as follows:[20]: *'What do you think, how many hours per week on average would you like to work in order to earn money?'* and *'And how many hours per week on average would you like your partner to work in order to earn money?'* To minimize the risk of comprehension problems in our diverse sample, we recorded answers to both questions on a 5-point scale ('0 hours, i.e. not at all'; 'about 10 hours'; 'about 20 hours'; 'about 30 hours' ; 'about 40 hours, i.e. full-time').

For our main analyses, we combine responses to these two expectations questions to analyze the following two outcomes: (a) self-expected labor supply with child; and (b) the expected within-family gender gap in labor supply with child, calculated as the difference in (i) expectations regarding the male family member's labor supply (i.e., male respondents' self-expectations, and female respondents' expectations regarding their partner), and (ii) expectations regarding the female family member's labor supply (i.e., female respondents' self-expectations, and male respondents' expectations regarding their partner).[21]

the treatment are relevant for them. Three pieces of evidence suggest that this is in fact the case: First, we do find significant effects of treatment information (see section 2.5). Second, the treatment alters respondents' beliefs about the opinion of their peer groups regarding mothers' and fathers' labor supply in the expected direction (see Online Appendix). Third, the opinions of adolescents in our sample (elicited at the end of our survey) mirror those of the general population: 71 percent (45 percent) think that mothers (fathers) should reduce their labor supply while children are young.

[20] Note that it is common in the literature to study expectation for events that occur several years in the future as we do (e.g., Goldin, 2014; Reuben et al., 2017; Wiswall and Zafar, 2018). Similarly to our study, Gong et al. (forthcoming) also examine labor-supply expectations within a hypothetical scenario where respondents imagine having a spouse and a child. The authors find that those expectations are indeed highly predictive for future labor-market supply, suggesting the validity of our outcome.

[21] In order to avoid implying mixed-sex relationships, we do not refer to the gender of partners when asking about expectations regarding labor supply. Nonetheless, for the sake of simplicity and readability, we refer to partners of female respondents as 'male' or 'father' and partners of male respondents as 'female' or 'mothers' throughout the chapter. If anything, we expect this slight inaccuracy to cause a downward bias in the observed within-family gender gaps.

Since expectations play an important role in any decision under uncertainty, we believe that our main outcomes of interest are well suited to study the impact of gender norms on labor market-relevant behavior. By obtaining direct measures of labor-market expectations, we can isolate the causal effect of gender norms on adolescents' future labor-supply expectations. Reassuringly, several studies show that (labor-market) expectations strongly predict actual (labor-market) realizations several years later (e.g., Wiswall and Zafar, 2018; Kunz and Staub, 2020; Wiswall and Zafar, 2020).

Follow-up Survey

We invited all respondents to participate in the follow-up survey two weeks after the main survey to (i) investigate the persistence of treatment effects, and (ii) implement a second experiment focusing on a more gender-egalitarian norm related to sharing household responsibility. Between 14 and 35 days after the main survey (median time lag: 17 days), we re-surveyed 1,319 respondents (66 percent of the sample).[22] Below, we introduce the individual stages of the follow-up survey:

Persistence of treatment effects: To study persistent treatment effects on labor-supply expectations, we first re-elicited labor-supply expectations as in the main survey. We then asked all respondents the belief-elicitation questions from the treatment *salience* (see section 2.3.2) to assess whether the treatments led to persistent changes in respondents' beliefs about existing norms regarding the labor supply of mothers and fathers with young children.

Second experiment: Thereafter, we implemented our second experiment to test the impact of a different, more gender-egalitarian norm on labor-supply expectations. For this purpose, respondents were randomly assigned to one of three experimental groups—a control and two treatment groups. Randomization in the second experiment was independent from treatment conditions in the first experiment. As pre-specified, we aimed to randomly allocated respondents to the control group with 50 percent probability and to each treatment group with 25 percent probability.[23]

The first treatment *egalitarian information* provides the following information: *'Out of 100 adults in Germany, 89 hold the opinion that men should take as much responsibility for the home and the children as women.'* (European Values Study 2017). As in the main survey experiment,

[22] Re-contact rate and time lag between main survey and follow-up are at the upper end of other recent studies featuring large-scale survey data from adult samples: Kuziemko et al. (2015), Haaland and Roth (2020), Alesina et al. (2018a), and Lergetporer et al. (2020) have re-contact rates of 14 percent, 66 percent, 24 percent, and 64 percent, and time lags of one month, one week, one week, and two weeks respectively.

[23] We chose these probabilities to maximize statistical power in regressions where we pool both treatment groups. Due to a programming error in the follow-up survey, group sizes turned out to be 28, 20 and 52 percent for treatment *egalitarian information*, treatment *salience and egalitarian information*, and the control group, respectively. Reassuringly, sociodemographic characteristics are well balanced across the experimental groups (see Appendix Table A2.4 for details).

we provided a graphical illustration alongside the verbal statement (see Appendix Figure A2.4 for screenshots).

The second treatment group *salience and egalitarian information* received the same information, but had to guess the share of Germans holding that opinion beforehand (similar to treatment *salience and information* in the main experiment).

The control group received no additional questions or information. We re-elicited expectations in the same manner as in the beginning of the follow-up survey and in the main survey.

By focusing on a more gender-egalitarian norm towards sharing household responsibility in the second experiment as compared to the traditional norm concerning parental labor supply used in the first experiment, we take advantage of the coexistence of different gender norms related to labor market behavior. The gender norms that we investigate in both experiments vary along several key dimensions: First, the egalitarian norm focuses on responsibility for tasks performed at the home (that are only indirectly related to our main outcome of interest—labor-supply expectations), while the traditional norm focuses on labor supply. Second, the egalitarian norm targets women and men more generally, while the traditional norm explicitly focuses on parents. Finally, the traditional norm entails specific recommendations for behavior (i.e., reducing labor supply), whereas the egalitarian norm refers to the vaguer concept of equal responsibility.

Debriefing: At the end of the survey, we showed a debriefing screen that (i) informed about the research question of the study (i.e., how gender norms affect labor-supply expectations), (ii) provided the content and data sources of both social-norm information treatments, and (iii) provided correct answers to the additional belief-elicitation questions that we posed during the survey (see section 2.5.3). Furthermore, to counteract that the treatments made gender norms salient, the debriefing also stressed that decisions regarding future work hours depend on many different important factors, and not only on social norms.

2.3.3 Sample Characteristics

Our analytic sample closely resembles the German population aged 14 to 17 years. Appendix Table A2.1 compares the characteristics of our sample to the respective population statistics in the German Microcensus.[24] Reassuringly, in most categories our respondents' characteristics match official statistics well. While there are slight differences between both samples with respect to gender, educational track, and maternal employment, our sample covers a broad and diverse spectrum of adolescents in Germany.

Appendix Table A2.2 presents balancing tests to check whether the randomization successfully balanced respondents' observable characteristics across the experimental groups in the main

[24] Research Data Centres of the Federal Statistical Office and the statistical offices of the Laender, Microcensus, census year 2015.

survey. The first column presents the average characteristics of the control group, and the subsequent columns present characteristics of treatment groups *salience*, *information*, and *salience and information* along with the respective differences to the control group. Reassuringly, only one out of 69 pairwise comparisons is significant at the 5-percent level, and four at the 10-percent level, as would be expected by pure chance. Thus, random assignment worked as intended.

Appendix Table A2.3 investigates whether participation in the follow-up survey is related to treatment assignment in the first experiment that was implemented in the main survey. Regressing a dummy for follow-up-survey participation on treatment indicators and covariates shows insignificant coefficients on treatments *salience* and *information*, and only a marginally significant coefficient on treatment *salience and information*. Furthermore, males, younger respondents, those living in large cities, and those without a degree are more likely to participate in the follow-up survey. Importantly, among those who participated in the follow-up survey, covariates are well-balanced across treatments of the first experiment (see Appendix Table A2.4), implying that treatment-effect estimates of the first experiment on outcomes measured in the follow-up survey are unbiased. Finally, Appendix Table A2.5 confirms that the randomization in the second experiment implemented in the follow-up survey was also successful in balancing respondents' characteristics across experimental groups.

2.3.4 Empirical Strategy

We estimate the causal effects of the social-norm treatments using the following regression model:

$$y_i = \alpha_0 + \sum_j \alpha_{1j} T_i^j + \delta^{'} X_i + \epsilon_i,\ with\ j \in \{S, I, S\&I\} \tag{2.1}$$

where y_i is the outcome variable of interest, and $T_i^S, T_i^I, T_i^{S\&I}$ are treatment indicators equal to 1 if respondent i received treatment *salience*, *information*, or *salience and information*, and 0 otherwise. X_i is a vector of control variables, and ϵ_i is the error term. Average treatment effects α_{1S}, α_{1I}, and $\alpha_{1S\&I}$ are identified because of random assignment. In some selected analyses, we pool treatments to facilitate exposition.

To analyze whether treatment effects are heterogeneous across gender, we extend our basic regression model to:

$$y_i = \beta_0 + \sum_j \beta_{1j} T_i^j + \sum_j \beta_{2j} T_i^j * female_i + \beta_3 female_i + \delta^{'} X_i + \epsilon_i,\ with j\ \in \{S, I, S\&I\} \tag{2.2}$$

The treatment effect for boys is given by β_1, and β_2 gives the additional effect for girls.

Since we elicited labor-supply expectations from each respondent twice in the follow-up survey (once at the very beginning to assess treatment-effect persistence of the first experiment, and again after the second experiment), we can estimate treatment effects on stacked data and include individual fixed effects to increase statistical precision. In the corresponding analysis, we therefore estimate the following regression model:

$$y_{it} = \alpha_0 + \sum_j \alpha_{1j} T_{it}^j + \mu_i + \epsilon_{it},\ with j \in \{EI, S\&EI\} \tag{2.3}$$

where y_{it} is the outcome variable of interest of respondent i at time t (before or after eventual treatment administration), and T_{it}^{EI}, $T_{it}^{S\&EI}$ are indicators for treatment *egalitarian information* and *salience and egalitarian information*, respectively. μ_i are individual fixed effects and ϵ_{it} is the error term. Standard errors are clustered at the individual level.

Finally, to deal with the potential issue of multiple-hypothesis testing, our regression tables further present adjusted p-values following the procedure by List et al. (2019). We adjust for multiple treatments, multiple subgroups (girls and boys) and—where applicable—for multiple outcomes.

2.4 Descriptive Results: Labor-Supply Expectations

We start by describing labor-supply expectations in the control group and discuss how they relate to respondents' characteristics.

Figure 2.1 depicts self-expected labor supply of girls and boys in the control group. The gender difference in expected work hours is large: While the majority of girls (59 percent) expect to work 20 hours or less (Panel A), the majority of boys (80 percent) expect to work 30 hours or more (Panel B). This difference corresponds to a gender gap of 7.3 work hours per week (23.8 versus 31.1 hours). Furthermore, expectations are consistent with gender gaps in actual labor supply: In the German Microcensus (2015), 17 percent of mothers with children aged between 1 and 6 years work full-time, while 46 percent work part-time, and 38 percent do not work at all. On the contrary, most fathers (87 percent) work full-time, and only 7 percent do not work at all.

The fact that expected gender gaps in labor supply are large is also reflected in our second outcome variable of interest: The average expected within-family gender gap in labor supply is 9.1 work hours per week.

Figure 2.2 shows how the two measures of labor-supply expectations vary across different sociodemographic subgroups, and Table A2.6 presents the corresponding bivariate regressions. Respondents' gender matters beyond self-expected labor supply, since girls expect a significantly smaller within-family gender gap than males (see Panel B of Figure 2.1). East German respondents expect to work longer hours than West German respondents (Panel A), and they expect a smaller within-family gender gap (Panel B). These findings reflect the well-documented fact that labor-force participation of women and mothers is traditionally higher in East Germany than in West Germany (e.g., Boelmann et al., 2020). Finally, respondents whose mothers currently work full-time expect smaller within-family gaps, which is in line with the literature on intergenerational transmission of gender norms (e.g., Fernández et al., 2004).

2.5 Effects of the Norm Prescribing Parental Labor Supply

In this section, we first describe baseline beliefs about the existing traditional gender norm concerning parental labor supply in Germany. Next, we present the effects of the norm treatments on labor-supply expectations. Finally, we show treatment effects on additional outcomes that are not directly targeted by the norm.

2.5.1 Beliefs about the Norm

Figure 2.3 depicts respondents' prior beliefs about the share of German adults who hold the opinion that mothers and fathers, respectively, should reduce their labor supply while their children are young—elicited in the treatments *salience* and *salience and information*. On average, respondents believe that 66 percent of Germans think that mothers should reduce their labor supply, and the median belief is 70 percent (true value: 91 percent). Thus, most adolescents underestimate the actual share of Germans holding this opinion (see Panel A). The same pattern applies to the social norm concerning fathers (see Panel B): The mean (median) belief is that 35 (31) percent of Germans think that fathers should reduce their labor supply (true value: 41 percent). Interestingly, prior beliefs do not differ systematically by respondents' gender: Girls' mean (median) belief about the norm towards mothers is 66 percent (70 percent), and it is 35 percent (35 percent) towards fathers. The respective figures for boys are 65 percent (70 percent) and 35 percent (30 percent).

In a nutshell, respondents underestimate the difference between the social norms with respect to mothers and fathers. The stark misperceptions of the prevailing social norm regarding labor supply of mothers and fathers indicate potential leeway for correcting these beliefs through information provision in treatments *information and salience* and *information*. In the next section, we study norm-treatment effects on expected labor supply.

2.5.2 Treatment Effects on Labor-Supply Expectations

Since the gender norm regarding parental labor supply prescribes different labor-market behavior for mothers and fathers, we present treatment-effect estimates separately for girls and for boys.

Girls' Labor-Supply Expectations

The left part of Figure 2.4 displays self-expected labor supply in the control group and in the pooled treatment groups for girls. The treatments significantly reduce self-expected labor supply: The share of girls expecting to work 20 hours or less per week increases significantly from 59 percent to 67 percent (see Panel A), and, conversely, the share expecting to work 30 hours or more significantly decreases from 41 percent to 33 percent (see Panel B).

Turning to regression results, Panel A of Table 2.1 shows that the pooled treatment effects correspond to a significant reduction of 2.6 hours in expected weekly work hours of girls (see column 1).[25] In Table 2.2 we present the effects of the three treatments separately. Each treatment has a highly significant and negative impact on girls' self-expected labor supply. The effect of the combined treatment *salience and information* is significantly ($p<0.1$) larger than the effect of treatment *salience*, and the effect size of treatment *information* is in-between the two. This pattern leads to three important insights: First, the strong effects of treatment *salience* suggests that priming adolescents to think about the existing gender norm already alters their labor-supply expectations. Second, the significant difference between treatment *salience* and the combined treatment *salience and information* suggests that providing information about the norm's content has an additional effect beyond the treatment *salience*. Third, the effect of treatment *salience and information* is smaller than the sum of the effects of treatment *salience* and treatment *information*, suggesting that information provision partly affects adolescents' expectations by making the norm salient. We discuss the role of the norms' salience in more detail in section 5.2.4.

Column 2 of Table 2.1 presents the pooled treatment effects on the expected within-family gender gap in labor supply, i.e., respondents' expected work hours of the male family member minus that of the female family member. The social-norm treatments significantly increase girls' expected within-family gender gap from 7.7 weekly work hours by 2.2 hours. Each treatment has a separate highly significant and positive impact on the expected within-family gender gap (see Table 2.2) which can be explained by the fact the treatments primarily decrease self-expectations, but not their expectations for the partner (see Online Appendix Table O2.1 for separate treatment effects on labor-supply expectations for the partner).[26]

To deal with the potential issue of multiple-hypothesis testing, the main tables additionally display adjusted p-values following the methodology of List et al. (2019). We find that levels of significance do not change substantially when adjusting for multiple subgroups (Table 2.1) or multiple subgroups as well as multiple treatments (Table 2.2).

[25] Online Appendix Table O2.1 presents treatment effects separately on each of the five answer categories.

[26] In line with the fact that we hardly find treatment effects on partners' labor-supply expectations, we do not find any treatment effects on girls' preferences for a set of partner attributes, either (e.g., whether the partner helps with the household or raising children etc.) (results available upon request).

In sum, girls' labor-supply expectations react strongly to treatments that highlight the traditional social norm on of how much mothers of young children should work on the labor market. Girls' expectations become more gender-unequal in the sense that they expect to work fewer hours themselves and expect a larger within-family gender gap in labor-market hours. Thus, our results indicate that gender norms play an important role in explaining gender gaps in labor-market outcomes after child birth.

Boys' Labor-Supply Expectations

Next, we investigate treatment effects on labor-supply expectations of boys. The right part of Figure 2.4 reveals that the pooled social norm treatments also reduce self-expected labor supply for boys: The share of boys expecting to work at most 20 hours per week significantly increases from 20 percent to 28 percent in response to the treatments (see Panel A). At the same time, the share of boys expecting to work 30 hours or more significantly decreases from 80 percent to 72 percent (see Panel B), which is entirely driven by a decrease in the share of boys expecting to work 40 hours.

Panel B of Table 2.1 show that these treatment effects correspond to a significant reduction of 1.8 expected weekly work hours (column 1).[27] Table 2.2 depicts treatment effects separately for each of the three norm treatments. In line with the results for girls, each of the three treatments has a negative impact on boys' self-expected labor supply (see column 1 of Table 2.2). The effect of the combined treatment *salience and information* is the largest one and highly significant ($p<0.01$), the coefficient on treatment-indicator *information* is marginally significant ($p<0.1$), and the coefficient on treatment *salience* does not reach statistical significance.

Columns 2 of Table 2.1 and Table 2.2 present treatment effects on the expected within-family gender gap in labor supply. In contrast to our findings for girls, boys' expectations for their own family tend to become more gender-equal when being confronted with the existing social norm. The treatments decrease (albeit not significantly so) boys' expected within-family gender gap from 11 weekly work hours by 1.4 hours, which is again due to the fact the treatments decrease self-expectations, but do not affect expectations for the partner (see Online Appendix Table O2.2).[28]

Overall, we find that the highlighting the gender norm prescribing mothers' and fathers' labor supply strongly impact adolescents' expectations of their own labor supply. While the

[27] Online Appendix Table O2.2 presents treatment effects separately for each of the five answer categories.

[28] In addition, Appendix Table A2.7 depicts treatment-effect heterogeneities by gender on both labor-supply expectations. Column 1 confirms that the treatments equally affect self-expected labor supply among girls and boys. Column 2 shows treatment effect heterogeneities on the expected within-family gender gap, where we observe differences by gender for all treatments. While girls become more gender-unequal in their expectations for their own family, boys seem to become more gender-equal by expecting a smaller within-family gender gap. In Online Appendix Table O2.3 we also present effect heterogeneities with respect to various other sociodemographic characteristics.

treatments make girls' expectations more gender-unequal, boys' expectations become more gender-equal. This latter result is particularly noteworthy given that the question how gender norms affect boys' labor supply has not yet been studied in the literature.

Persistence of Treatment

Next, we check whether treatment effects persist beyond the immediate survey horizon into the follow-up survey conducted about two weeks later. Given that priming effects, such as the ones induced by treatment *salience*, are by definition temporary and short-term (e.g., Forehand et al., 2002; Benjamin et al., 2010), we only expect persistent effects from treatments that entail information provision. Therefore, our persistence-analysis compares the pooled experimental groups that received and did not receive information on the norm's content (i.e., treatments *information and salience* and *information* versus treatment *salience* and the control group). As the previous section shows that the treatments do not affect respondents' labor supply expectations for their partner, we restrict our analyses of persistence on self-expected labor supply.

Table 2.3 combines data from the main survey and the follow-up survey and regresses self-expected labor supply on an information-treatment dummy, a follow-up-survey dummy, and the interaction of both indicators. For the overall sample, column 1 shows that information-treatment effects persist in the follow-up survey. As expected, the treatment effect in the follow-up survey tends to be somewhat smaller than the one in the main survey (likely due to imperfect recall), although the difference between treatment effects is not statistically significant (see coefficient on the interaction term). Columns 2 and 3 report persistent treatment effects separately for girls and boys. While treatment effects in the follow-up survey for these subsamples are remarkably similar in magnitude to the full sample, they do not reach statistical significance due to limited statistical power.[29]

Turning to belief-updating, Appendix Table A2.9 investigates respondents' stated beliefs relative to accurate values. It shows that information provision persistently improves beliefs about the content of the norm, i.e. the share of Germans who hold the opinion that mothers and fathers, respectively, should reduce their labor supply. Interestingly, respondents seem to internalize in particular the existing norm for their own gender, which is consistent with the fact that the norm treatments mainly affect self-expected labor supply.[30]

[29] In further analyses, we exploit variation in the time lag between main and follow-up survey, and find that treatment-effect persistence does not significantly differ for respondents who participated earlier vs. later in the follow-up (results available upon request). Appendix Table A2.8 analyses persistence for all three treatments separately. While statistical power is again limited, results suggest the effect of treatment *salience* does not persist, while the effects of treatments involving information provision have the expected sign and are partly significant.

[30] Appendix Figure A2.5 presents the entire distribution of beliefs.

In sum, the information treatments in the main survey lead to a persistent updating of self-expected labor supply and beliefs in the follow-up survey, which implies that (i) adolescents indeed understand and remember the provided information, and (ii) treatment effects are unlikely due to experimenter-demand effects (e.g., Cavallo et al., 2017; Bleemer and Zafar, 2018; Haaland et al., 2020).[31]

Heterogeneities by Prior Beliefs

The literature discusses two potential channels through which information provision may affect individuals' responses: salience-based versus information-based updating (Bleemer and Zafar, 2018). In our setting, information treatments may affect labor-supply expectations because information provision increases the salience of the gender norms (e.g., Schwarz and Vaughn, 2002; Chetty et al., 2009; DellaVigna, 2009; Bleemer and Zafar, 2018), or because it corrects respondents' misperceptions about the content of the norm (e.g., Rockoff et al., 2012; Bursztyn et al., 2020).[32] To investigate the relevance of these channels, we estimate heterogeneous information-treatment effects by respondents' prior beliefs (elicited in treatments *salience* and *salience and information*) using the following regression model:

$$y_i = \gamma_0 + \gamma_1 T_i^{S\&I} + \gamma_2 T_i^{S\&I} * Misperception_i + \gamma_3 Misperception_i + \delta^{'} X_i + \epsilon_i \qquad (2.4)$$

where $Misperception_i$ is the difference between the factual share of Germans thinking that respondent i's gender (i.e., mothers or fathers) should reduce his or her labor supply minus respondent i's belief about this share. The coefficient γ_1 captures the average information effect for respondents with correct prior beliefs, and γ_2 captures the additional effect for respondents who initially misperceive the social norm. Information-based updating would imply that γ_2 is significant and negative, whereas salience-based updating would imply that γ_2 is close to zero.

Table 2.4 shows that the coefficients on the treatment *salience and information* are negative and, in the full sample, significantly different from zero. In addition, coefficients on the interaction term are negative as to be expected for information-based updating (but shy of statistical significance).[33]

[31] de Quidt et al. (2018) and Mummolo and Peterson (2019) show that (survey) experiments are largely robust to experimenter-demand effects.

[32] The idea behind salience-based updating is that information about one specific norm increases the salience of that specific norm relative to other aspects that may affect labor-supply expectations. Labor-supply expectations could generally be influenced by a multitude of social-norm considerations, because (i) social identity is multidimensional (for instance, it can refer to gender, race, or social status (Akerlof and Kranton, 2000) and (ii) several social norms may coexist within each domain of social identity. In the context of gender norms, such norms may prescribe labor supply, relative income within households (Bertrand et al., 2015), or shared household responsibility, for instance.

[33] In line with the interpretation of information-based-updating, we also find significant treatment effects of the treatments that provide accurate information on beliefs elicited in the follow-up survey (see section 2.5.2).

Our results on the treatment effects for respondents with accurate priors suggest that providing information about the social norm's content at least partly affects labor-supply expectations by increasing the norm's salience. This interpretation is in line with the previous literature: Bleemer and Zafar (2018) find that the effects of college-returns information on intended college attendance do not vary by respondents' prior beliefs about college returns. Similarly, Alesina et al. (2018a) argue that salience effects drive their negative findings of correcting natives' over-pessimistic beliefs about immigrants on natives' preferences for redistribution.[34]

2.5.3 Treatment Effects on Additional Outcomes

Incentivized Outcomes

A common critique against the expectations-literature is that the main outcomes of interest—survey-based expectations about future events or actions—have no immediate consequences for respondents, which raises concerns about the outcome variables' relevance.

To test whether treatment effects carry over to outcomes with direct monetary consequences, we next present results from a set of incentivized belief-questions that asked respondents to guess the shares of Germans who agree with the following gender-related statements (European Values Study 2017): (i) 'A university education is more important for a boy than for a girl.' (EVS: 16 percent); (ii) 'When the mother works for pay, the children suffer.' (EVS: 33 percent); (iii) 'Being a housewife is just as fulfilling as working for pay.' (EVS: 55 percent); and (iv) 'If a woman earns more than her husband, it's almost certain to cause problems.' (EVS: 20 percent). We introduce immediate monetary consequences for correct answers by paying each respondent two additional Euros for a roughly correct answer (defined as belonging to the better half of guesses) to the randomly chosen question.

Figure 2.5 provides a graphical depiction of the belief distributions for each item across experimental groups with and without information provision, and Table 2.5 reports the corresponding regression results.[35] Information provision significantly affects respondents' incentivized beliefs about the share of Germans agreeing to the different items. This finding, along with the fact that previous research has shown that unincentivized expectations are tightly linked

[34] In the Online Appendix, we study perceived peer pressure as a further potential mechanism driving our treatment effects, and find that the channel seems to be relevant for girls but not for boys.

[35] We expect only the information content of treatments *information* as well as *information and salience* to spill over to the gender-related items as the incentivized outcome questions per se already induces all respondents to think about societal expectations and hence increase salience of the respective issues (similar to the questions posed in treatment *salience*). We therefore pool the two treatments *information and salience* and *information* and compare them to treatments *salience* and the control group. Appendix Table A2.10 reports effects of each treatment separately, and confirms that only those treatments that entail information provision affect the incentivized outcomes.

to real outcomes, gives rise to our interpretation that treatment effects reflect effects of highlighting the gender norm on (labor-market) relevant outcomes.[36]

Labor-Supply Expectations Without Child

The traditional gender norm that we study prescribes labor supply of mothers and fathers. Consequently, the norm treatments should affect labor-supply expectations with child, but not necessarily without child. To perform this additional sanity check, we also elicited respondents' expected labor supply at the age of 30 in a hypothetical scenario without child.

Appendix Table A2.11 shows pooled treatment effects on labor-supply expectations without children. Treatment effects on both self-expected labor supply and the expected within-family gender gap are small and insignificant for both genders. The fact treatment effects are confined to those outcomes that are directly prescribed by the norm further raises confidence that our experimental results reflect genuine effects of highlighting the specific norm, as opposed to general shifts in expectations in response to the treatments.[37]

2.6 Effects of the Norm on Shared Household Responsibility

So far, we have shown that highlighting the prevailing traditional social norm prescribing mothers' and fathers' labor supply decreases labor-supply expectations, and thereby potentially promotes gender gaps in labor-market outcomes. We now investigate whether highlighting a more egalitarian gender norm can have the opposite effect. We therefore conduct a second experiment in the follow-up survey. The corresponding treatments highlight a more gender-egalitarian norm towards sharing household responsibilities. In this section, we first describe baseline beliefs about the norm, and then present treatment effects on labor-supply expectations.

2.6.1 Beliefs about the Norm

Figure 2.6 depicts prior beliefs about the egalitarian gender norm elicited in treatment *salience and egalitarian information*. It shows that respondents underestimate the egalitarianism of the norm: The mean (median) guess is that 59 percent (60 percent) of Germans think that men should take as much responsibility for the household as women, whereas the true share in the German population is 89 percent. While both genders misperceive this norm, girls' beliefs tend to be more accurate than boys' beliefs (60 percent versus 55 percent median guess).

[36] Interestingly, respondents in the treatment group report more conservative beliefs, which undermines the accuracy of beliefs in all items but item (iii). In view of this result, it is particularly important to note that we provided accurate information about the different items in the debriefing stage at the very end of the survey.
[37] In the Online Appendix, we study preferences for job attributes as additional indirect outcome variables, and find little evidence that these preferences are affected by the norm-treatments.

2.6.2 Treatment Effects on Labor-Supply Expectations

Table 2.6 depicts pooled effects of treatments *egalitarian information and salience* and *egalitarian information* on labor-supply expectations. Focusing on the expected within-family gender gap in labor supply as the outcome of interest,[38] we indeed find that the treatments attenuate the expected gender gap in labor supply. In particular, the pooled treatments significantly decrease the expected gap by 1.3 hours per week (column 1). Girls expect a reduction of the gender gap by 1.3 hours (column 2), and boys by 1.4 hours (column 3). Appendix Table A2.12 shows that both treatments have statistically significant negative effects on the expected within-family gender gap. If anything, treatment effects tend to be stronger in the combined treatment *salience and egalitarian information* than in treatment egalitarian information, which resembles the patterns in the first experiment (see section 2.5).

In sum, these results show that highlighting the more egalitarian gender norm towards sharing household responsibility can lead to more gender-equal expectations regarding the within-family gender gap in labor supply after child birth. In the treatment groups both genders expect the mother to reduce her labor supply less relative to the father. Thus, the effects of highlighting gender norms on labor-market expectations can depend on the specific context—and the degree of gender-equality—of the respective norm.

2.7 Conclusion

In many developed countries, gender differences in labor-market outcomes do not emerge until the arrival of the first child. We shed light on the causal relationship between labor-market outcomes and perceived gender norms in large-scale experiments among 2,000 adolescents in Germany, a country with comparatively large child penalties in addition to a very traditional norm on how much mothers should work on the labor market. At baseline, most girls (59 percent) expect to work no more than 20 hours per week with a young child, and most boys (80 percent) expect to work at least 30 hours per week. Administering treatments that highlight the existence of a traditional gender norm in Germany—i.e., that 91 percent (41 percent) of Germans think that mothers (fathers) of young children should reduce labor supply—significantly reduces girls' labor-supply expectations by 2.6 hours per week which increases the expected within-family gender gap in labor supply. While largely neglected by the literature so far, we also study how the gender norm affects boys' labor-supply expectations. Boys also expect to reduce their labor supply in response to the norm treatments, which translates into a reduced expected within-family gender gap. Finally, we show that an alternative treatment highlighting a more gender-egalitarian norm towards sharing household responsibility results in more gender-equal labor-supply expectations among both genders. In sum, our results

[38] We focus on the expected within-family gap outcome of interest because this social norm explicitly addresses the household as a whole. Further analyses indeed reveal that treatment effects on the within-family gender gap are driven by changes in self-expectations as well as partners' expectations (results available upon request).

indicate that gender norms play an important role in shaping outcomes relevant to the labor market.

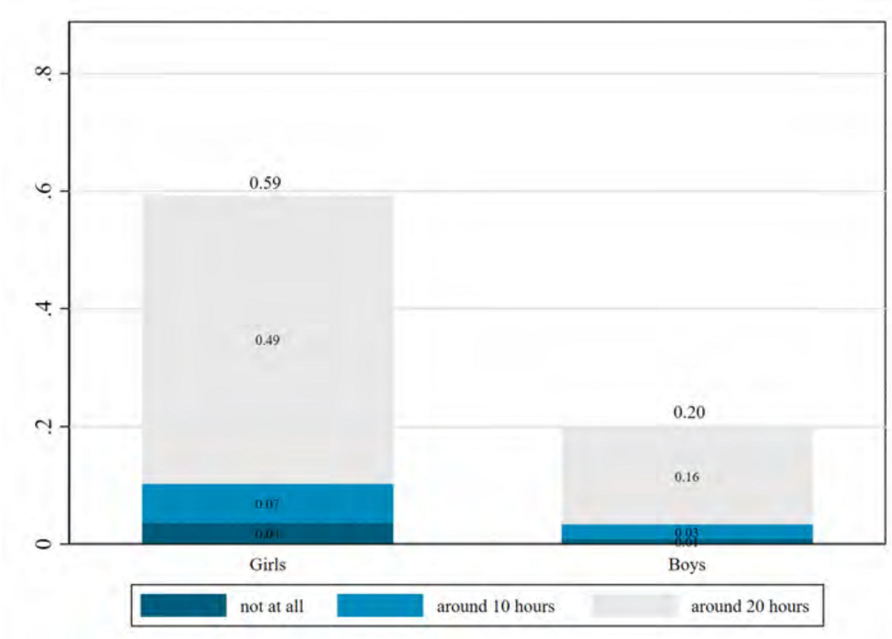
Our research design to investigate the effects of altering the salience of, and perception about existing gender norms comes with both advantages and potential limitations. On the one hand, it allows close control over the exogenous variation induced in the experiment. We can therefore attribute any changes in adolescents' expectations to *the causal effects* of two specific aspects of the social norm—its salience and information about its exact content. Alternative research designs which, for instance, exploit exogenous variation in migration streams, can only identify reduced-form effects of the whole culture (including gender norms, but also other beliefs, values or preferences), but cannot isolate the effect of one specific gender norm. On the other hand, our main outcomes of interest are labor-supply expectations which may not readily translate into future labor-market choices or realizations. In line with the economic literature on (labor-market) expectations, we stress that expectations are a crucial predictor for educational choices and attainment: Expectations can easily become self-fulfilling if adolescents with lower expectations have smaller incentives to invest in academic accomplishments (e.g., Beaman et al., 2012; Stinebrickner and Stinebrickner, 2013; Reuben et al., 2017). Similarly, adolescents may choose different jobs and different occupations that are consistent with their expectations. In addition, labor-market expectations closely relate to actual labor-market realizations, even years later (e.g., Goldin et al., 2006; Wiswall and Zafar, 2018; Kunz and Staub, 2020; Wiswall and Zafar, 2020) and when elicited in hypothetical scenarios with spouse and child (Gong et al., forthcoming). Studying the causal effects of the salience and perceptions of gender norms on actual labor-market outcomes is an interesting avenue for future research.

Our results bear immediate relevance for policy: Policy makers who wish to foster gender-equality may be able to change perceptions about specific gender norms for instance through information campaigns, or by changing how gender roles are presented in school books, advertisements or the media. In addition, some of the short-lived treatment effects presented in this chapter (i.e., the effects of salience) suggest that timing such interventions right before adolescents take crucial educational or occupational decisions may be expedient.

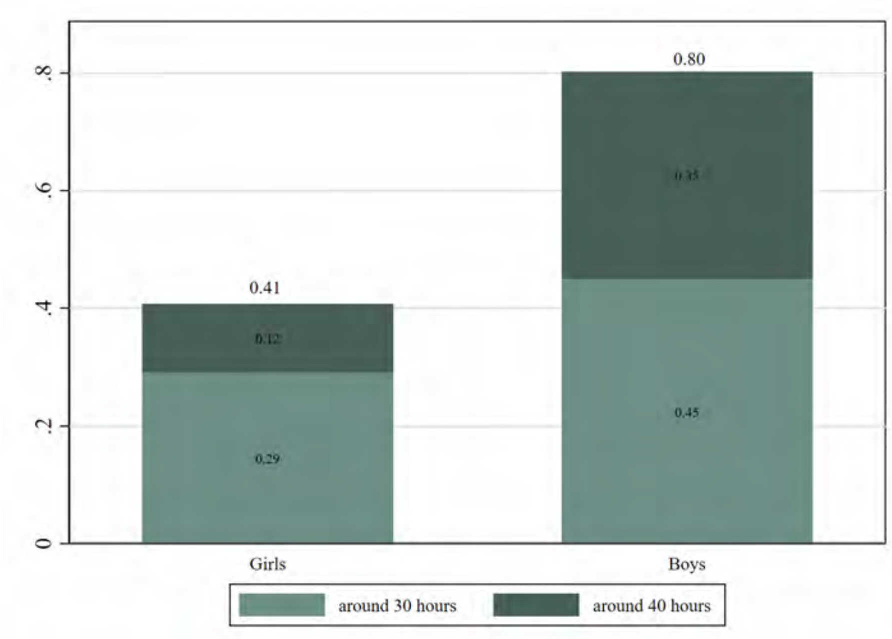
Figures and Tables

Figure 2.1: Labor-Supply Expectations in the Control Group

Panel A: Expectations to Work at Most 20 Hours per Week



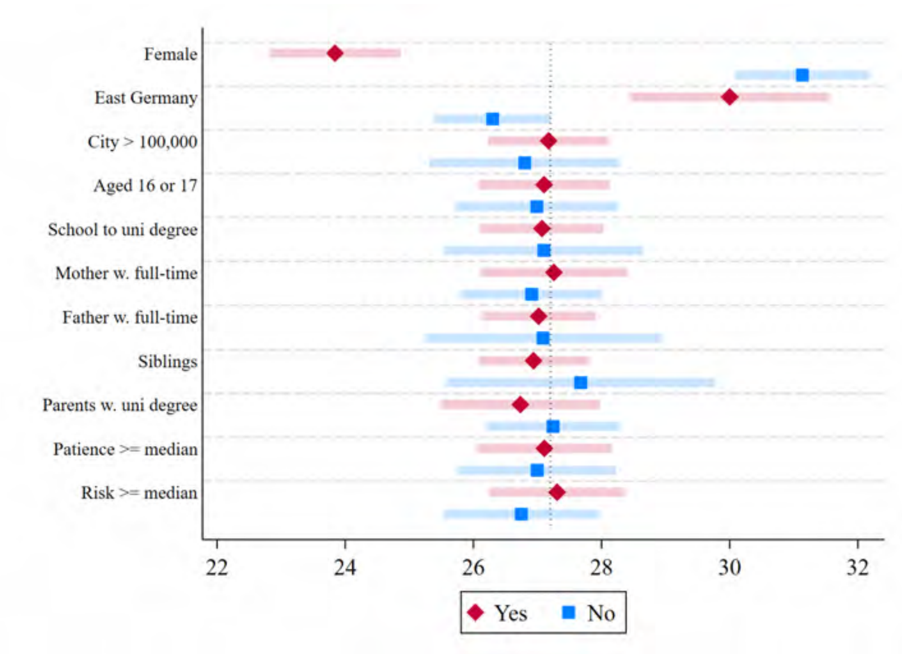
Panel B: Expectations to Work at Least 30 Hours per Week



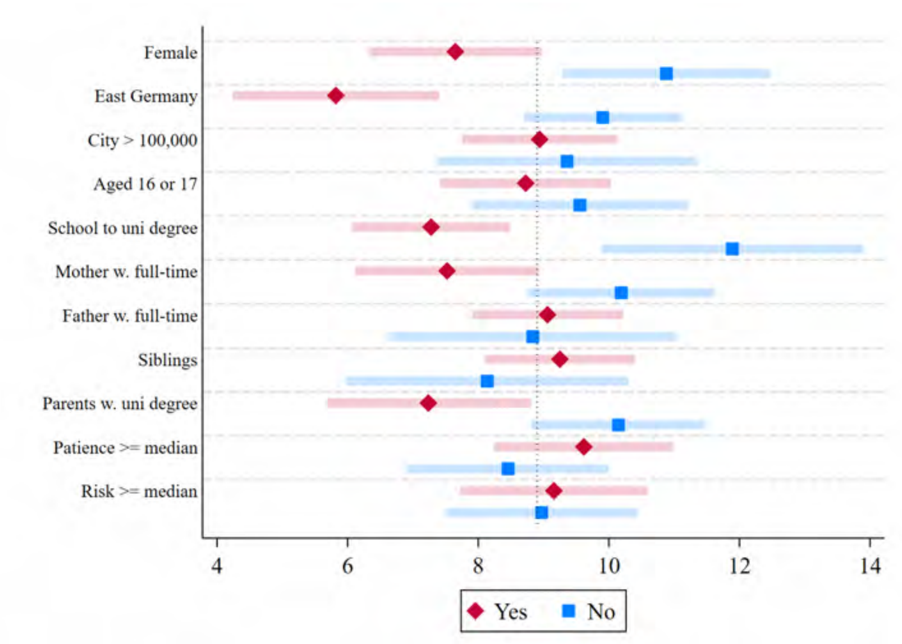
Notes: Responses to the question *'Imagine you are 30 years old and you have a child aged between 1 and 6 years with your partner. What do you think, how many hours per week on average would you like to work in order to earn money?'*. Sample: respondents in the control group.

Figure 2.2: Labor-Supply Expectations across Sociodemographic Characteristics

Panel A: Self-Expected Labor Supply



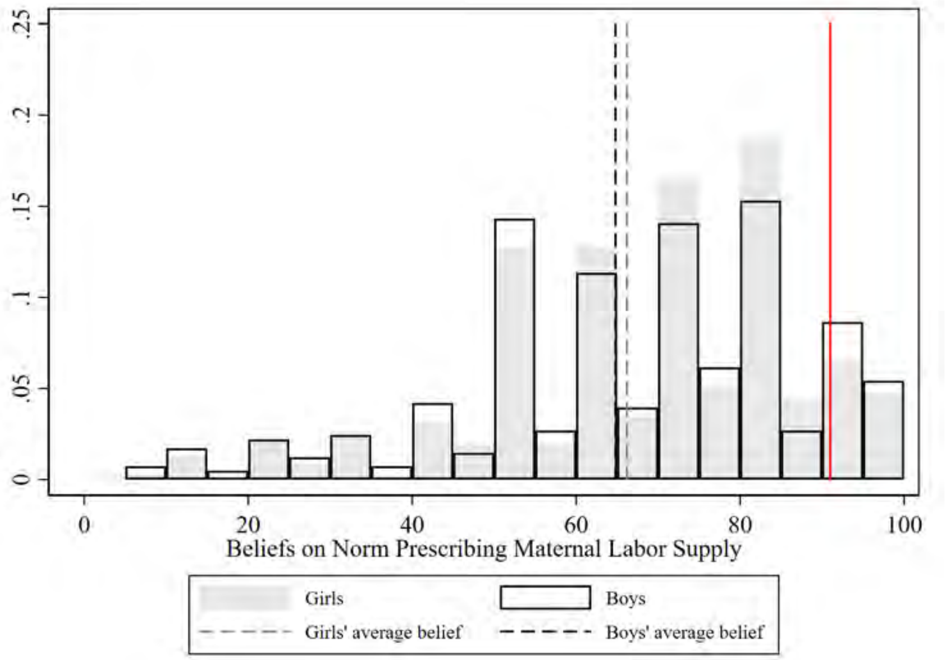
Panel B: Expected Within-Family Gender Gap



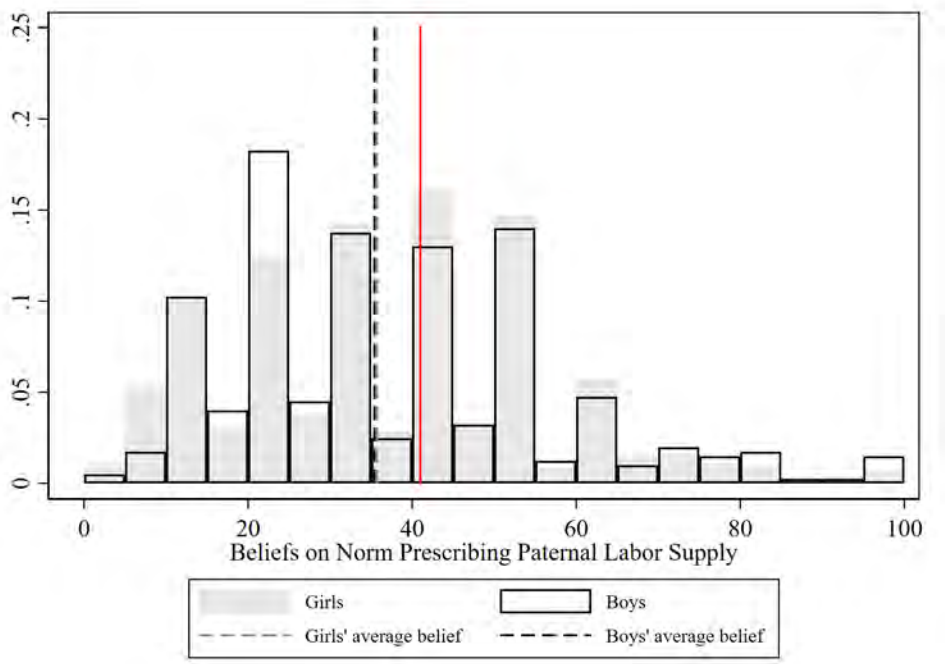
Notes: The figure shows the control group answers to the expected labor supply for different groups of respondents. The shaded areas are 95 percent confidence intervals around the average response. Panel A: hours expected to work per week with child 1–6 (0=0 hours, i.e. not at all; 10= 10 hours; 20= 20 hours, i.e. part-time; 30=30 hours, 40=40 hours, i.e. full-time). Panel B: responses to labor supply for both parents (self and partner) with higher values indicating higher labor market supply of men relative to women. Average hours (full sample) indicated by vertical, dotted line.

Figure 2.3: Distribution of Beliefs about the Norm Prescribing Parental Labor Supply

Panel A: Norm that Mothers Should Reduce Their Labor-Supply when Children Are Young



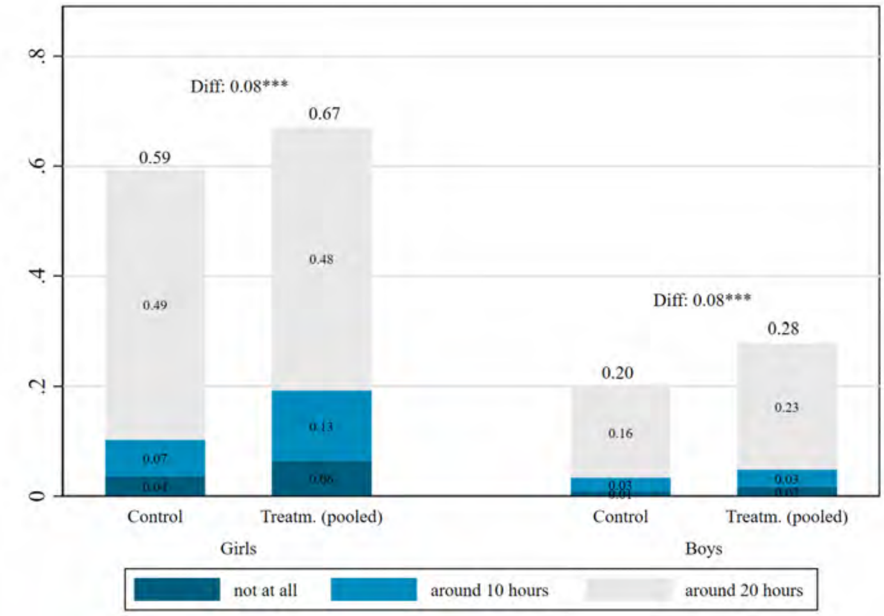
Panel B: Norm that Fathers Should Reduce Their Labor-Supply when Children Are Young



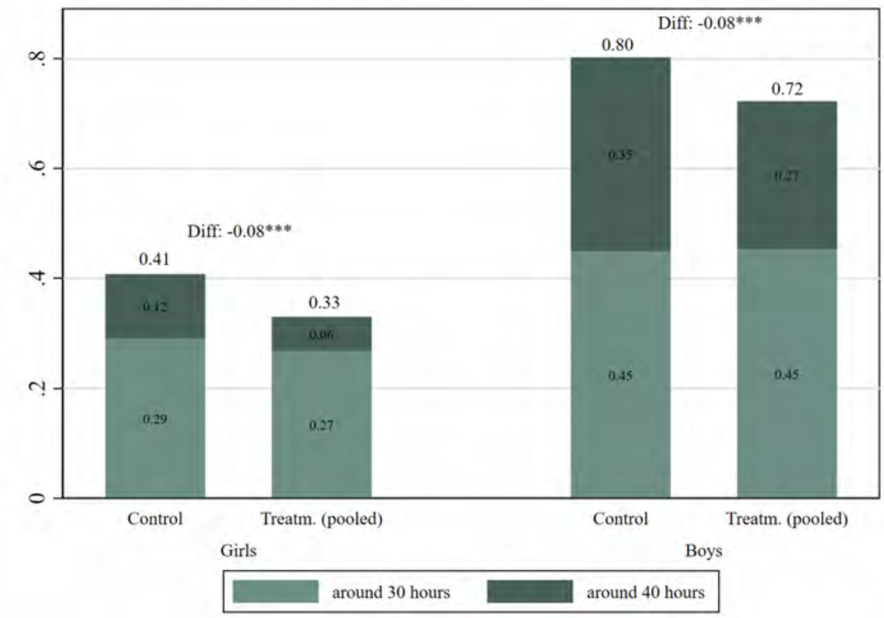
Notes: Beliefs about the extent of the norm that parents should reduce their labor market supply when children are young. Correct values indicated by vertical lines. Sample: Respondents in experimental groups *salience* or *salience and information*.

Figure 2.4: Self-expected Labor Supply across Treatments

Panel A: Expectations to Work at Most 20 Hours per Week

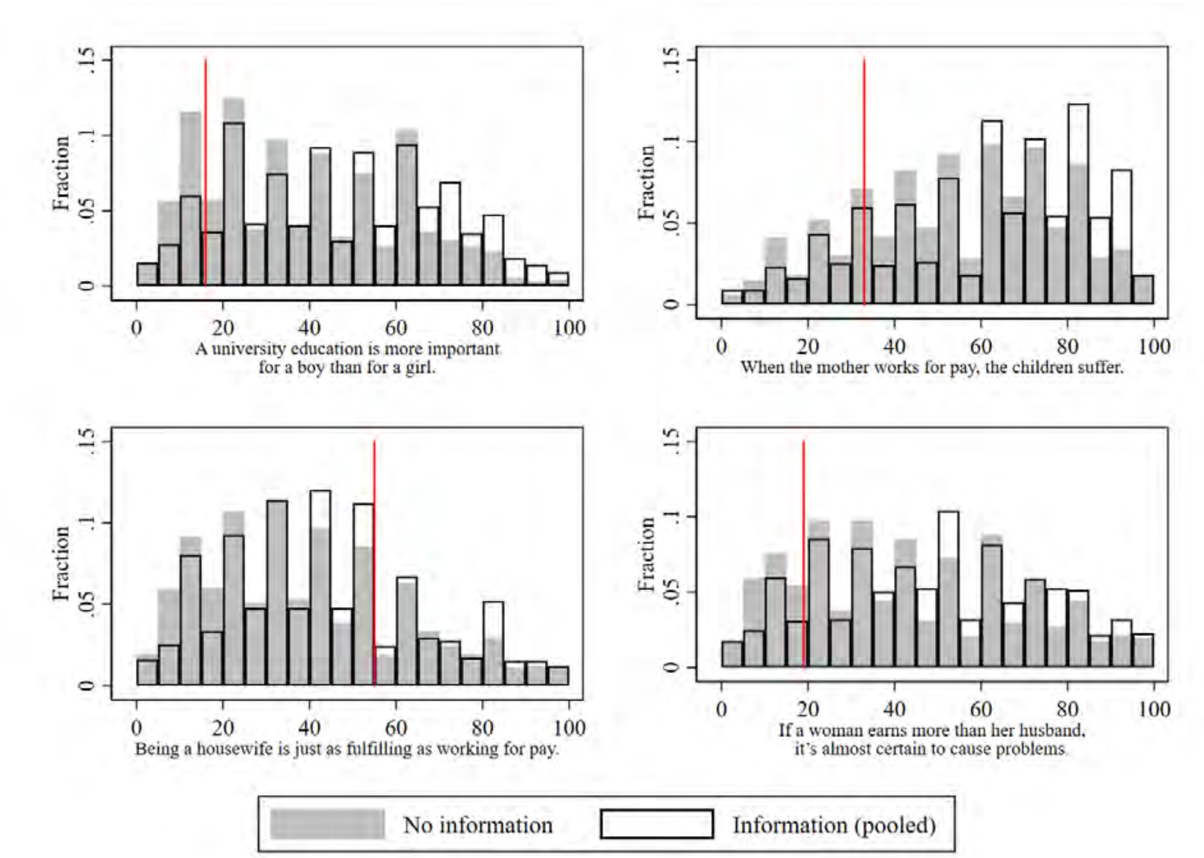


Panel B: Expectations to Work at Least 30 Hours per Week



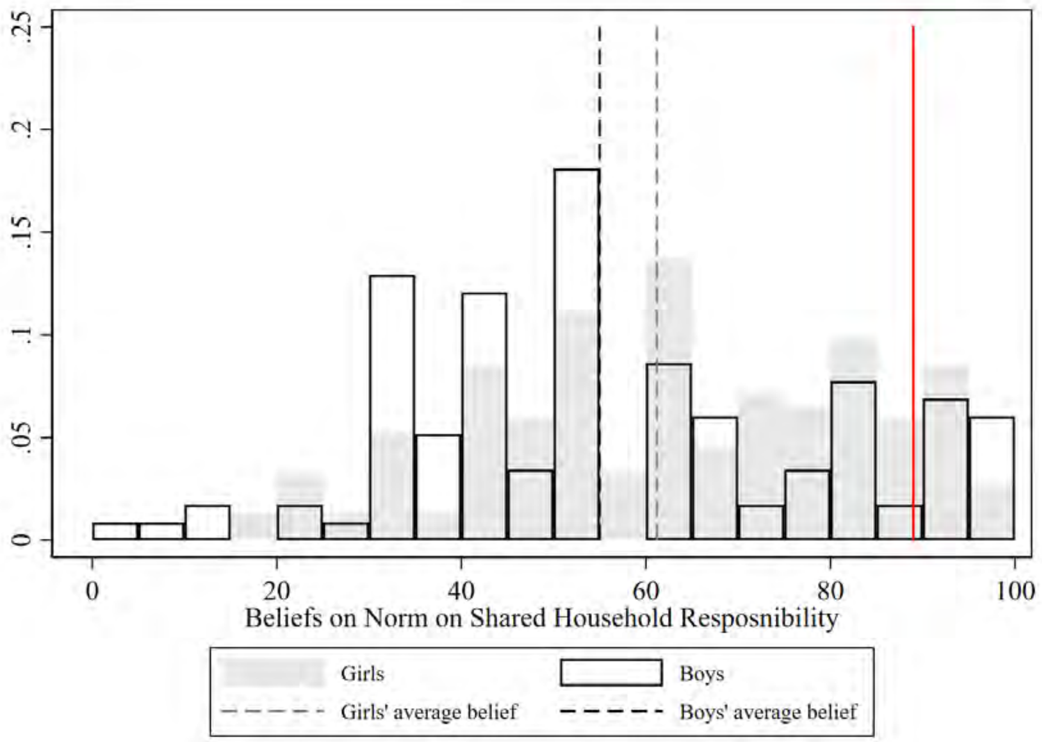
Notes: Responses to the question *'Imagine you are 30 years old and you have a child aged between 1 and 6 years with your partner. What do you think, how many hours per week on average would you like to work in order to earn money?'* Treatm. (pooled): respondent in experimental groups *salience*, *information* or *salience and information*. ***/**/* indicate significance at the 1%/5%/10% level from regression according to equation 2.1.

Figure 2.5: Distribution of Incentivized Outcomes across Treatments



Notes: Responses to incentivized belief questions on share of Germans agreeing with gender-related statements depicted at x-axis. Correct values indicated by vertical lines. No information: respondent in experimental groups *control group* or *salience*. Information (pooled): respondent in experimental groups *information* or *salience and information*.

Figure 2.6: Distribution of Beliefs about the Norm on Shared Household Responsibility



Notes: Beliefs about the extent of the norm that men should take as much responsibility for the home and the children as women. Correct values indicated by vertical lines. Sample: respondents in experimental group *salience and egalitarian information*.

Table 2.1: Treatment Effects on Labor-Supply Expectations

	Self-expected labor supply	Expected within-family gender gap
	(1)	(2)
Panel A: Girls		
Treatments (pooled)	-2.610*** (0.600)	2.240*** (0.766)
p-values: MHT Correction		
Treatments (pooled)	0.000	0.006
Control mean	23.84	7.65
Observations	1164	1164
R-squared	0.099	0.075
Panel B: Boys		
Treatments (pooled)	-1.814*** (0.650)	-1.415 (0.980)
p-values: MHT Correction		
Treatments (pooled)	0.004	0.162
Control mean	31.13	10.88
Observations	836	836
R-squared	0.039	0.050
Panel C: All		
Treatments (pooled)	-2.187*** (0.445)	0.708 (0.608)
Control mean	27.06	9.07
Observations	2000	2000
R-squared	0.191	0.044
Covariates	Yes	Yes

Notes: OLS regressions. Treatments (pooled): respondent in experimental groups *salience*, *information* or *salience and information*. Dependent variables: (1) hours expected to work per week when having a child 1–6 (0=0 hours, i.e. not at all; 10= 10 hours; 20= 20 hours, i.e. part-time; 30=30 hours, 40=40 hours, i.e. full-time); (2) Responses to labor supply for both parents (self and partner) with higher values indicating higher labor market supply of men relative to women. Control mean: mean of the outcome variable in the control group. Covariates include: age, gender, born in Germany, living with parents, currently in school, current school track leading to university entrance degree, mother working full-time, having siblings, West Germany, living in large city, parents with university education, risk, patience, and imputation dummys. MHT Correction refers to the multiple hypothesis testing procedure presented in List et al. (2019) and corrects for two subgroups (girls and boys). Robust standard errors in parentheses. ***/**/* indicate significance at the 1%/5%/10% level.

Table 2.2: Separate Treatment Effects on Labor-Supply Expectations

	Self-expected labor supply	Expected within-family gender gap
	(1)	(2)
Panel A: Girls		
Salience	-1.965*** (0.731)	2.207** (0.924)
Information	-2.582*** (0.774)	1.969** (0.970)
Salience & information	-3.250*** (0.718)	2.552*** (0.941)
p-values: MHT Correction		
Salience	0.022	0.092
Information	0.004	0.164
Salience & information	0.000	0.045
Control mean	23.84	7.65
Observations	1164	1164
R-squared	0.101	0.076
Panel B: Boys		
Salience	-1.328 (0.823)	-1.233 (1.283)
Information	-1.578* (0.872)	-2.145* (1.295)
Salience & information	-2.418*** (0.786)	-0.951 (1.200)
p-value: MHT Correction		
Salience	0.095	0.538
Information	0.126	0.229
Salience & information	0.014	0.407
Control mean	31.13	10.88
Observations	836	836
R-squared	0.041	0.051
Panel C: All		
Salience	-1.534*** (0.555)	0.779 (0.756)
Information	-2.132*** (0.581)	0.372 (0.785)
Salience & information	-2.827*** (0.534)	0.963 (0.748)
p-value: MHT Correction		
Salience	0.007	0.498
Information	0.000	0.639
Salience & information	0.000	0.444
Control mean	27.06	9.07
Observations	2000	2000
R-squared	0.193	0.045
Covariates	Yes	Yes

Notes: OLS regressions. *Salience*, *Information* and *Salience & information* indicate membership of respective treatment groups. Dependent variables: (1) hours expected to work per week when having a child 1–6 (0=0 hours, i.e. not at all; 10= 10 hours; 20= 20 hours, i.e. part-time; 30=30 hours, 40=40 hours, i.e. full-time); (2) responses to labor supply for both parents (self and partner) with higher values indicating higher labor market supply of men relative to women. Control mean: mean of the outcome variable in the control group. See Table 2.1 for included covariates. MHT Correction refers to the multiple hypothesis testing procedure presented in List et al. (2019) and corrects for two subgroups (girls and boys) as well as multiple treatments in Panel B. Results from Wald tests, testing for equal coefficients reject *Salience=Salience & information* in column (1) of Panel A ($p<0.1$) as well as column (1) of Panel C ($p<0.05$). Equal coefficients within all remaining treatment/outcome/subgroup combinations cannot be rejected. Robust standard errors in parentheses. ***/**/* indicate significance at the 1%/5%/10% level.

Table 2.3: Persistence of Information Treatment Effects on Self-Expected Labor Supply (Follow-up Sample)

	Self-expected labor supply		
	(1) All	(2) Girls	(3) Boys
Information provision (pooled)	-1.673***	-1.673**	-1.537**
	(0.493)	(0.676)	(0.702)
Information provision (pooled) x follow-up	0.553	0.641	0.438
	(0.518)	(0.691)	(0.788)
Follow-up	0.447	0.503	0.378
	(0.359)	(0.465)	(0.564)
Info provision in follow-up	-1.120**	-1.032	-1.098
Control mean	26.15	22.63	30.48
Covariates	Yes	Yes	Yes
Observations (respondents)	1319	734	585
R-squared	0.199	0.109	0.028

Notes: OLS regressions. Info provision: respondent in experimental groups *information* or *salience and information*. Dependent variable: Hours expected to work per week when having a child 1–6 (0=0 hours, i.e. not at all; 10= 10 hours; 20= 20 hours, i.e. part-time; 30=30 hours, 40=40 hours, i.e. full-time). *Info provision in follow-up* is the linear combination of the coefficients on *Info provision* plus *Info provision x follow-up*. Control mean: mean of the outcome variable in the omitted group (i.e. experimental groups *control group* or *salience*) reported in the main survey. See Table 2.1 for included covariates. Sample: respondents who participated in the follow-up survey. Robust standard errors, adjusted for clustering at the respondent level, in parentheses. ***/**/* indicate significance at the 1%/5%/10% level.

Table 2.4: Heterogeneous Treatment Effects on Self-expected Labor Supply by Prior Beliefs (Belief Elicitation Sample)

	Self-expected labor supply		
	(1)	(2)	(3)
	All	Girls	Boys
Salience & information	-1.334*	-1.017	-1.297
	(0.687)	(1.241)	(0.845)
Misperception x salience & information	-0.046	-0.145	-0.020
	(0.266)	(0.403)	(0.400)
Misperception	0.292	0.366	0.215
	(0.219)	(0.302)	(0.303)
Covariates	Yes	Yes	Yes
Observations	965	565	400
R-squared	0.197	0.116	0.045

Notes: OLS regressions. Salience & information: Respondents in respective experimental group. Dependent variable: hours expected to work per week when having a child 1–6 (0=0 hours, i.e. not at all; 10= 10 hours; 20= 20 hours, i.e. part-time; 30=30 hours, 40=40 hours, i.e. full-time). Misperception: actual share minus guessed share of German adult population holding the opinion that parents (of respondent's gender) should reduce their labor market supply as long as the children are young, divided by 10. See Table 2.1 for included covariates. Sample: Respondents in experimental groups *salience* and *salience and information*. Robust standard errors in parentheses. ***/**/* indicate significance at the 1%/5%/10% level.

Table 2.5: Information Treatment Effects on Incentivized Outcome

	University education more important for boy.	Children suffer if mother works for pay.	Being a housewife as fulfilling as working for pay.	Causes problems if a woman earns more than her husband.
	(1)	(2)	(3)	(4)
Panel A: Girls				
Information provision (pooled)	0.457***	0.173***	0.095***	0.225***
	(0.085)	(0.042)	(0.024)	(0.075)
p-values: MHT Correction				
Information (pooled)	0.000	0.000	0.000	0.003
Control mean	2.262	1.612	0.668	2.150
Covariates	Yes	Yes	Yes	Yes
Observations	1130	1156	1143	1137
R-squared	0.041	0.053	0.046	0.024
Panel B: Boys				
Information provision (pooled)	0.607***	0.215***	0.069**	0.337***
	(0.102)	(0.049)	(0.029)	(0.087)
p-values: MHT Correction				
Information (pooled)	0.000	0.000	0.014	0.000
Control mean	2.299	1.583	0.678	1.978
Covariates	Yes	Yes	Yes	Yes
Observations	811	829	820	807
R-squared	0.069	0.059	0.057	0.040

Notes: OLS regressions. Info provision: respondent in experimental groups *information* or *salience and information*. Dependent variables: beliefs about share of Germany agreeing with the statements that (1) a university education is more important for a boy than for a girl relative to correct value (=16); (2) the children suffer if the mothers works for pay relative to correct value (=33); (3) being a housewife is just as fulfilling as working for pay relative to correct value (=55); (4) it is almost certain to cause problems if a woman earns more than her husband relative to correct value (=20). Results (not shown) from full interaction model between gender and treatment indicators do not reveal any heterogeneous treatment effects by gender. Control mean: mean of the outcome variable in the control group. See Table 2.1 for included covariates. MHT Correction refers to the multiple hypothesis testing procedure presented in List et al. (2019) and corrects for multiple subgroups (girls and boys) and multiple outcomes (all 4 outcomes listed). Robust standard errors in parentheses. ***/**/* indicate significance at the 1%/5%/10% level.

Table 2.6: Effects of the More Egalitarian Norm on the Expected Within-Family Gender Gap (Follow-Up Sample)

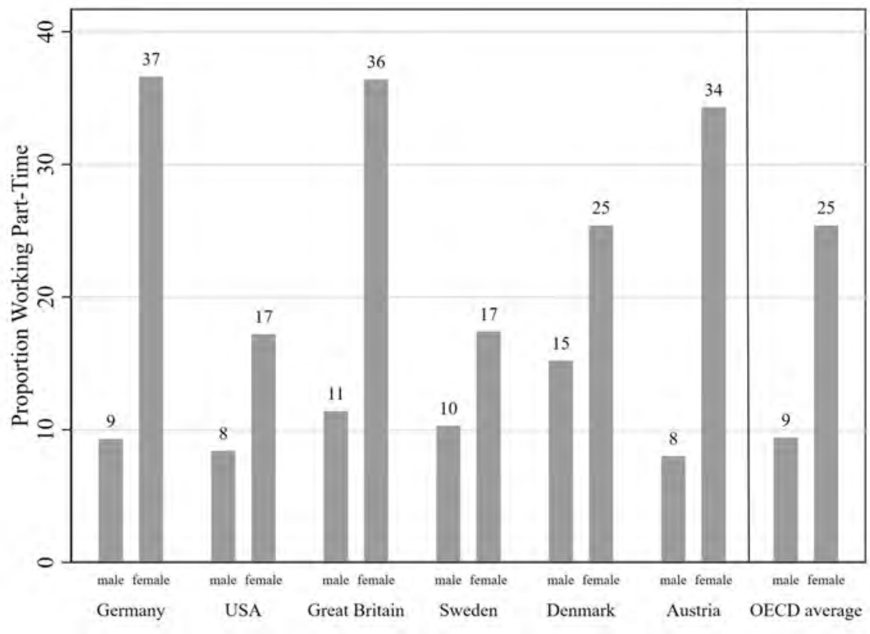
	Expected within-family gender gap		
	(1) All	(2) Girls	(3) Boys
Egalitarian treatments (pooled)	-1.321***	-1.261***	-1.398**
	(0.351)	(0.437)	(0.571)
p-values: MHT Correction			
Treatments (pooled)		0.007	1.000
Individual FEs	Yes	Yes	Yes
Observations (respondents)	1319	734	585
R-squared	0.010	0.011	0.009

Notes: OLS regressions. Egalitarian treatments (pooled): respondent in experimental groups *egalitarian information* or *salience and egalitarian information*. Dependent variables: responses to labor supply for both parents (self and partner) with higher values indicating higher labor market supply of men relative to women. Results (not shown) from interaction model between gender and treatment indicator do not reveal any heterogeneous treatment effects by gender. Sample: follow-up survey participants. Robust standard errors, adjusted for clustering at the respondent level, in parentheses. ***/**/* indicate significance at the 1%/5%/10% level.

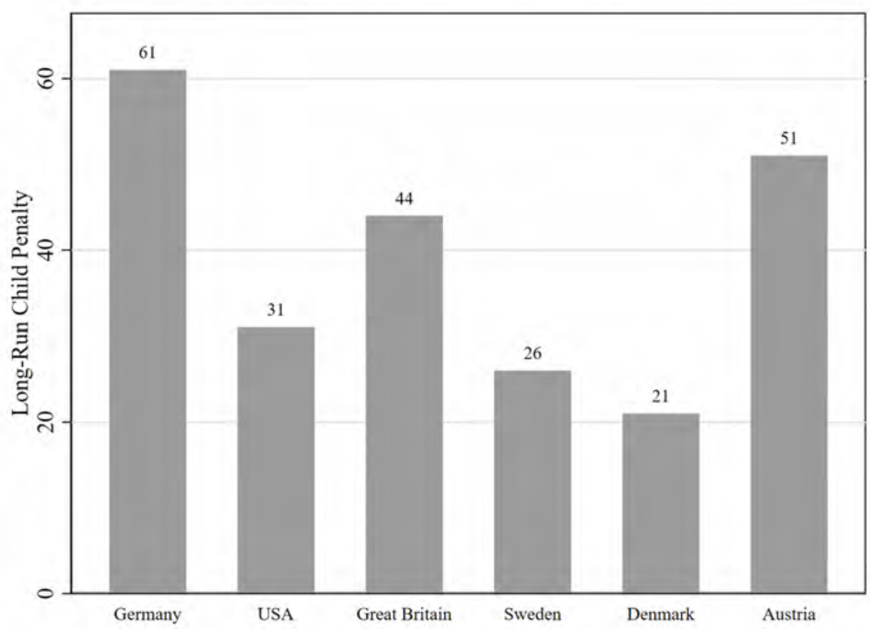
Appendix

Figure A2.1: Gender Inequalities in Labor Supply across Countries

Panel A: Gender Gaps in Part-Time Employment across Countries

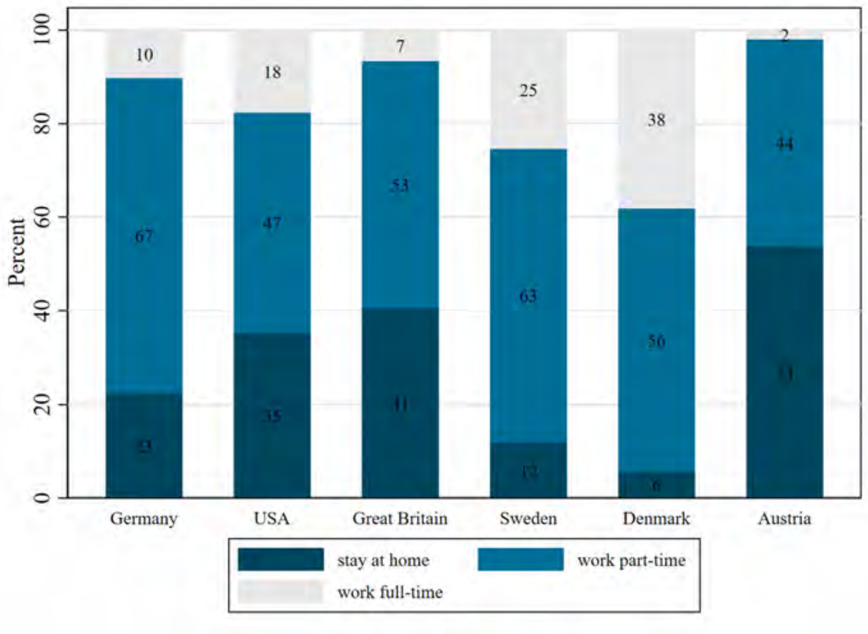


Panel B: Long-run Child Penalties across Countries



Notes: Panel A: part-time employment rate as proportion of persons employed part-time among all employed persons, by gender. Part-time employment is defined as people in employment (whether employees or self-employed) who usually work less than 30 hours per week in their main job. Data source: OECD, 2018; Panel B: long-run child penalties. Data source: Kleven et al. (2019).

Figure A2.2: Norm about Maternal Labor Supply across Countries

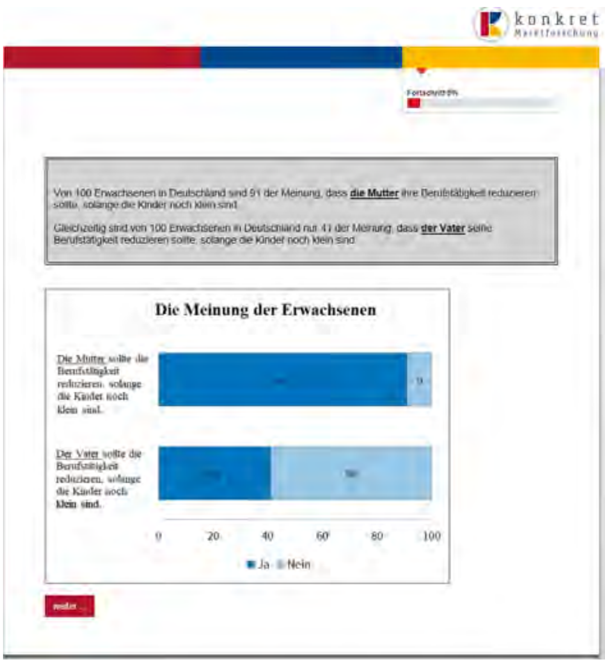


Notes: Response to the question *'Do you think women should work outside the home full-time, part-time or not at all under the following circumstances? ...When there is a child under school age.'* Weighted means. Data source: International Social Survey Program (ISSP) in 2012.

Figure A2.3: Screenshots on Norm Treatments

Panel A: Belief Elicitation about Norm Prescribing Parental Labor Supply

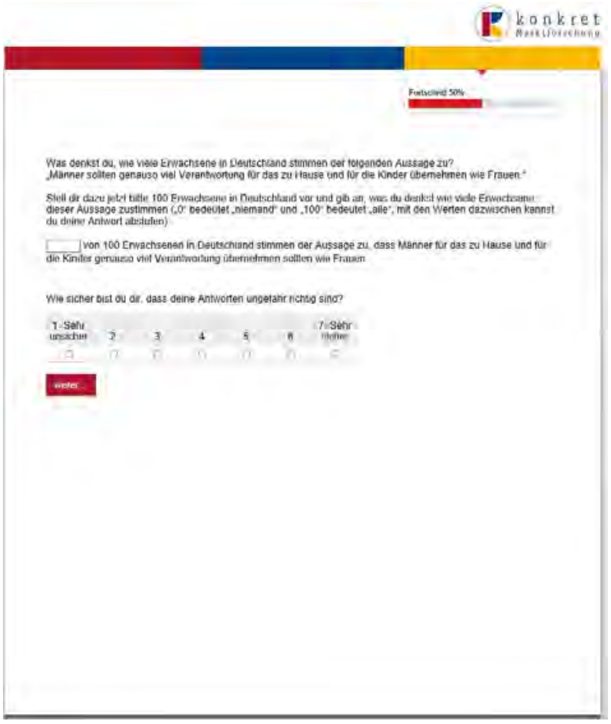
Panel B: Information about Norm Prescribing Parental Labor Supply



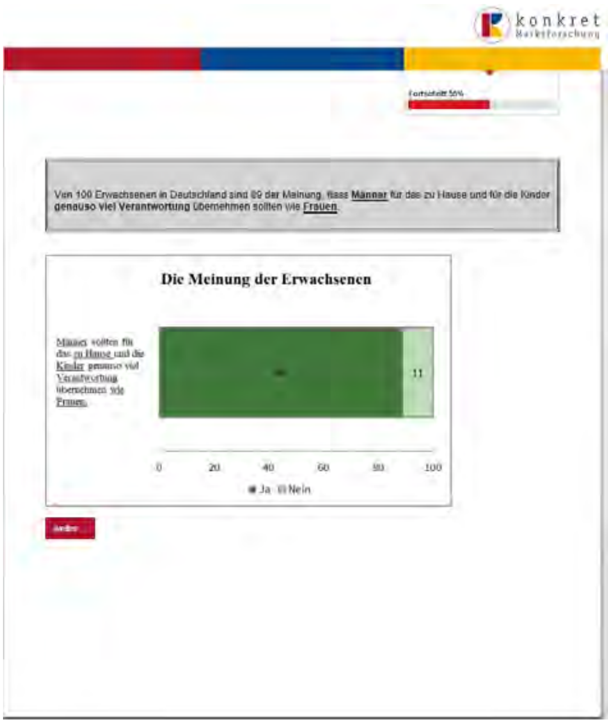
Notes: Panel A: screenshot on the belief elicitation questions in treatment *salience*. Panel B: screenshot on the information provision in treatment *information*. Respondents in treatment *salience & information* first receive the belief elicitation question (Panel A) and afterwards accurate information (Panel B).

Figure A2.4: Screenshots on More Egalitarian Norm Treatments in Follow-Up Survey

Panel A: Belief Elicitation about Norm on Shared Household Responsibility



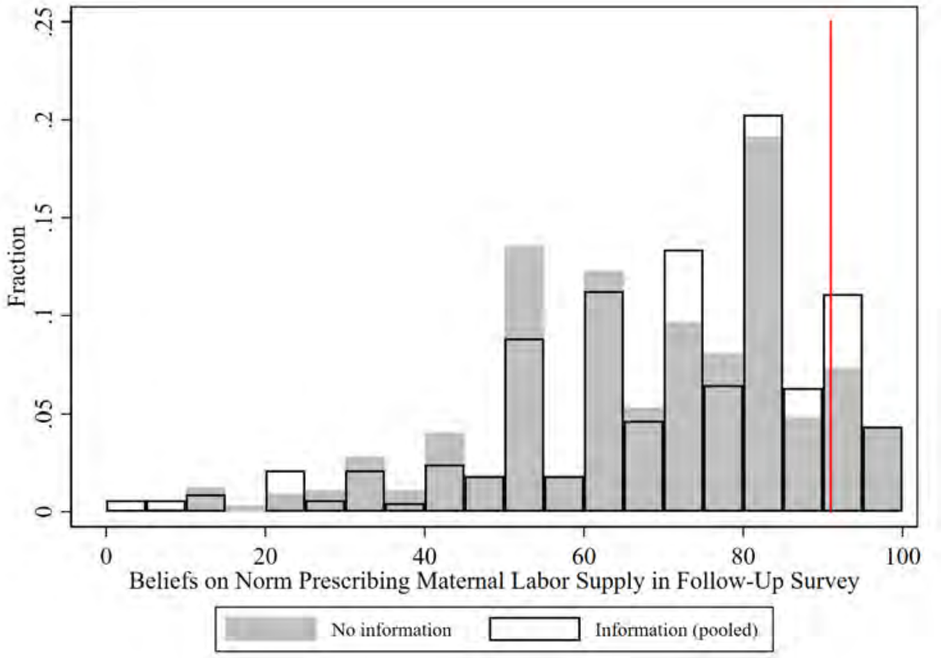
Panel B: Information about Norm on Shared Household Responsibility



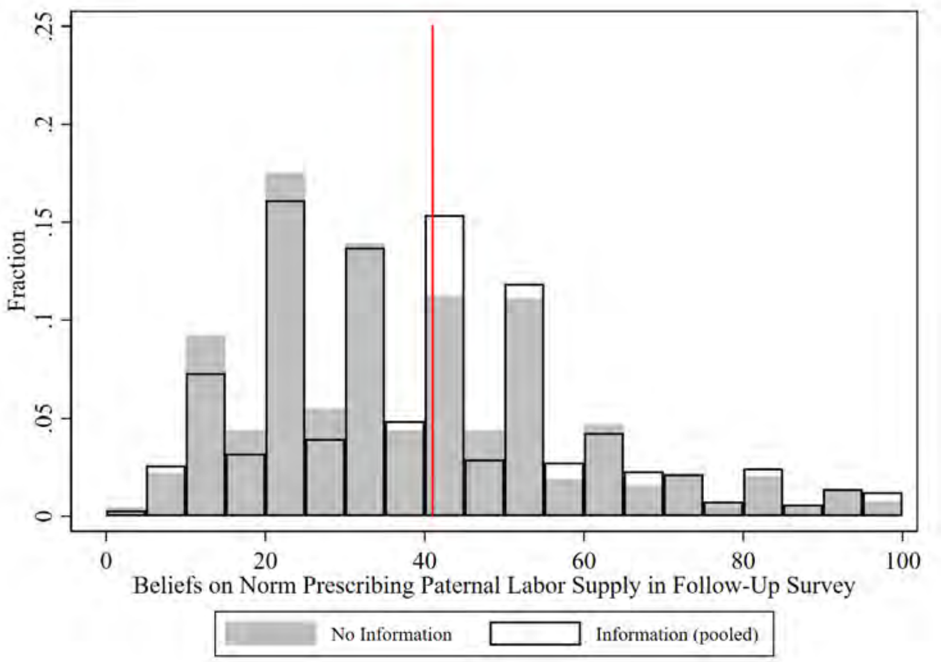
Notes: Respondents in treatment *salience and egalitarian information* first receive the belief elicitation question (Panel A) and afterwards accurate information (Panel B). Respondents in treatment *egalitarian information* are provided with accurate information (Panel B).

Figure A2.5: Distribution of Beliefs about Norm Prescribing Parental Labor Supply in Follow-Up Survey

Panel A: Norm that Mothers Should Reduce their Labor Supply when Children Are Young



Panel B: Norm that Fathers Should Reduce their Labor Supply when Children Are Young



Notes: Beliefs about the extent of the norm that parents should reduce their labor market supply when children are young elicited in the follow-up survey. Correct values indicated by vertical lines. Information (pooled): respondents in experimental groups *information* or *salience and information*). Sample: follow-up survey participants.

Table A2.1: Comparison of Survey Sample Characteristics to Microcensus Data

Characteristic	Microcensus 2015 (1)	Sample mean (2)
Female	0.488	0.582
		(0.011)
Age	15.508	15.748
		(0.026)
Living in West Germany (excl. Berlin)	0.847	0.794
		(0.009)
Attending Hauptschule/Realschule (low/middle track)	0.288	0.172
		(0.008)
Attending school with several tracks	0.156	0.138
		(0.008)
Attending Gymnasium (high track)	0.393	0.509
		(0.011)
Living with both parents	0.761	0.723
		(0.010)
At least one parent with uni degree [if living with both]	0.449	0.420
		(0.013)
Mother does not work [if living with both]	0.233	0.173
		(0.010)
Mother works full-time [if living with both]	0.207	0.405
		(0.013)
Father works full-time [if living with both]	0.875	0.914
		(0.008)
Observations	18501	2000

Notes: Column 1: means based on Microcensus data from 2015. Column 2: sample means and standard errors (in parentheses) of our survey data. Data source: German population Microcensus 2015 and own survey data.

Table A2.2: Sociodemographic Characteristics across Treatments

	Control mean (1)	Salience (2)	Diff. (3)	Information (4)	Diff. (5)	Salience & information (6)	Diff. (7)
Female	0.559	0.603	0.043	0.607	0.048	0.563	0.004
Age	15.752	15.758	0.006	15.800	0.048	15.685	-0.067
Living in west Germany	0.796	0.778	-0.019	0.802	0.005	0.799	0.003
City size ≥ 100,000	0.681	0.698	0.016	0.697	0.016	0.687	0.006
Born in Germany	0.963	0.957	-0.006	0.973	0.010	0.976	0.013
Currently at school	0.893	0.922	0.030	0.922	0.030*	0.911	0.019
Obtained degree/Current track leads to							
No degree	-0.000	-0.000	0.000	0.002	0.002	0.000	0.000
Basic degree (Hauptschulabschluss)	0.049	0.049	-0.001	0.045	-0.005	0.040	-0.010
Middle school degree (Realschulabschluss)	0.284	0.239	-0.045	0.247	-0.037	0.305	0.021
University entrance degree (Fachabitur)	0.061	0.058	-0.003	0.096	0.035**	0.067	0.006
University entrance degree (Abitur)	0.606	0.654	0.049	0.611	0.005	0.589	-0.017
Living status							
Living with both parents	0.715	0.732	0.017	0.671	-0.044	0.705	-0.010
Living with one parent	0.267	0.251	-0.016	0.292	0.026	0.272	0.005
Living without parents	0.019	0.017	-0.001	0.037	0.018*	0.024	0.005
Having siblings	0.841	0.827	-0.014	0.849	0.008	0.797	-0.043*
At least one parent with university degree	0.369	0.413	0.044	0.389	0.020	0.386	0.017
Maternal employment status							
Mother works full-time	0.419	0.426	0.008	0.423	0.005	0.437	0.018
Mother works part-time	0.367	0.357	-0.010	0.366	-0.001	0.362	-0.004
Mother housewife	0.031	0.026	-0.006	0.033	0.001	0.031	0.000
Paternal employment status							
Father works full-time	0.809	0.814	0.005	0.838	0.029	0.852	0.043*
Father works part-time	0.041	0.043	0.002	0.033	-0.008	0.041	0.001
Risk tolerance (11-point scale)	5.674	5.616	-0.059	5.538	-0.136	5.500	-0.174
Patience (11-point scale)	6.461	6.341	-0.120	6.430	-0.031	6.421	-0.040
Observations	540	463		463		508	

Notes: Group means. 'Diff.' displays the difference in means between the control group and respective treatment groups. Significance levels of 'Diff.' from linear regressions of the background variables on the respective treatment indicators. Robust standard errors in parentheses. ***/**/* indicate significance at the 1%/5%/10% level.

Table A2.3: Participation in the Follow-Up Survey

	Respondent participated in follow-up survey	
	(1)	
Treatments		
Salience	0.022	(0.031)
Information	0.017	(0.031)
Salience & information	0.056*	(0.030)
Covariates		
Female	-0.058***	(0.022)
Age	-0.041***	(0.010)
Living in west Germany	-0.022	(0.027)
City size ≥ 100,000	0.060**	(0.024)
Born in Germany	-0.068	(0.059)
Currently at school	0.086*	(0.047)
No degree	0.435***	(0.096)
Middle school degree (Realschulabschluss)	0.029	(0.055)
University entrance degree (Fachabitur)	0.073	(0.067)
University entrance degree (Abitur)	0.026	(0.056)
Living with one parent	-0.034	(0.026)
Living without parents	-0.090	(0.075)
Having siblings	-0.044	(0.028)
At least one parent with uni degree	-0.005	(0.023)
Mother works full-time	0.017	(0.023)
Father works full-time	0.058*	(0.031)
Risk tolerance (11-point scale)	-0.006	(0.005)
Patience (11-point scale)	-0.004	(0.005)
Observations	1901	
R-squared	0.037	

Notes: Dependent variable: dummy variable coded one if respondent participated in the follow-up survey. Salience/Information/Salience & information indicate membership of respective treatment group. Robust standard errors in parentheses. ***/**/* indicate significance at the 1%/5%/10% level.

Table A2.4: Sociodemographic Characteristics across the Treatments (Follow-up Sample)

	Control mean (1)	Salience (2)	Diff. (3)	Information (4)	Diff. (5)	Salience & information (6)	Diff. (7)
Female	0.535	0.570	0.035	0.581	0.046	0.544	0.009
Age	15.623	15.694	0.071	15.714	0.091	15.558	-0.065
Living in west Germany	0.789	0.769	-0.021	0.797	0.007	0.803	0.013
City size ≥ 100,000	0.690	0.717	0.027	0.733	0.043	0.704	0.014
Born in Germany	0.956	0.958	0.002	0.968	0.012	0.972	0.016
Currently at school	0.918	0.922	0.004	0.949	0.031	0.935	0.017
Obtained degree/Current track leads to							
No degree	0.000	-0.000	-0.000	0.003	0.003	0.000	0.000
Basic degree (Hauptschulabschluss)	0.047	0.054	0.007	0.039	-0.007	0.030	-0.017
Middle school degree (Realschulabschluss)	0.298	0.246	-0.052	0.241	-0.057	0.290	-0.008
University entrance degree (Fachabitur)	0.068	0.074	0.006	0.085	0.016	0.069	0.001
University entrance degree (Abitur)	0.587	0.626	0.039	0.632	0.045	0.611	0.024
Living status							
Living with both parents	0.713	0.746	0.032	0.698	-0.015	0.713	-0.001
Living with one parent	0.263	0.238	-0.025	0.279	0.016	0.268	0.004
Living without parents	0.023	0.016	-0.007	0.022	-0.001	0.020	-0.004
Having siblings	0.822	0.810	-0.011	0.857	0.036	0.772	-0.050
At least one parent with university degree	0.383	0.404	0.021	0.410	0.026	0.386	0.003
Maternal employment status							
Mother works full-time	0.421	0.453	0.032	0.410	-0.012	0.451	0.030
Mother works part-time	0.380	0.332	-0.048	0.397	0.017	0.366	-0.014
Mother housewife	0.035	0.026	-0.009	0.025	-0.010	0.039	0.004
Paternal employment status							
Father works full-time	0.815	0.840	0.025	0.834	0.019	0.868	0.052^{*}
Father works part-time	0.038	0.036	-0.002	0.038	0.000	0.037	-0.002
Risk tolerance (11-point scale)	5.611	5.518	-0.093	5.479	-0.132	5.456	-0.155
Patience (11-point scale)	6.322	6.267	-0.055	6.404	0.083	6.439	0.118
Observations	342	307		315		355	

Notes: Group means. 'Diff.' displays the difference in means between the control group and respective treatment groups. Significance levels of 'Diff.' from linear regressions of the background variables on the respective treatment indicators. Robust standard errors in parentheses. Sample: follow-up survey participants. ***/**/* indicate significance at the 1%/5%/10% level.

Table A2.5: Sociodemographic Characteristics across the More Egalitarian Treatments (Follow-up Sample)

	Control mean (1)	Egalitarian information (2)	Diff. (3)	Salience & egalitarian information (4)	Diff. (5)
Female	0.552	0.557	0.005	0.567	0.015
Age	15.624	15.675	0.051	15.652	0.028
Living in west Germany	0.794	0.787	-0.007	0.785	-0.008
City size ≥ 100,000	0.716	0.689	-0.027	0.726	0.010
Born in Germany	0.963	0.970	0.007	0.956	-0.008
Currently at school	0.922	0.932	0.009	0.952	0.029*
Obtained degree/Current track leads to					
No degree	-0.000	-0.000	0.000	0.004	0.004
Basic degree (Hauptschulabschluss)	0.041	0.037	-0.004	0.050	0.009
Middle school degree (Realschulabschluss)	0.278	0.269	-0.009	0.251	-0.027
University entrance degree (Fachabitur)	0.071	0.077	0.006	0.077	0.007
University entrance degree (Abitur)	0.610	0.617	0.007	0.618	0.008
Living status					
Living with both parents	0.719	0.727	0.008	0.700	-0.019
Living with one parent	0.261	0.249	-0.012	0.285	0.025
Living without parents	0.020	0.025	0.004	0.015	-0.006
Having siblings	0.806	0.814	0.008	0.833	0.027
At least one parent with university degree	0.394	0.393	-0.000	0.400	0.006
Maternal employment status					
Mother works full-time	0.452	0.418	-0.034	0.407	-0.045
Mother works part-time	0.357	0.410	0.053*	0.344	-0.013
Mother housewife	0.026	0.041	0.015	0.033	0.007
Paternal employment status					
Father works full-time	0.840	0.839	-0.001	0.840	-0.000
Father works part-time	0.040	0.038	-0.001	0.030	-0.010
Risk tolerance (11-point scale)	5.577	5.311	-0.265*	5.641	0.064
Patience (11-point scale)	6.374	6.470	0.096	6.178	-0.196
Observations	683	366		270	

Notes: Group means. 'Diff.' displays the difference in means between the control group and respective treatment groups. Significance levels of 'Diff.' from linear regressions of the background variables on the respective treatment indicators. Robust standard errors in parentheses. Sample: follow-up survey participants. Robust standard errors in parentheses. ***/**/* indicate significance at the 1%/5%/10% level.

Table A2.6: Labor-Supply Expectations across Sociodemographic Characteristics (Respondents in the Control Group)

	Self-expected labor supply (1)		Expected within-family gender gap (2)	
Female	-7.293***	(0.748)	-3.233***	(1.053)
Living in east Germany	3.698***	(0.910)	-4.089***	(1.009)
City size ≥ 100,000	0.372	(0.894)	-0.420	(1.175)
Aged 16 or 17	0.111	(0.830)	-0.831	(1.072)
University entrance degree (Abitur)	-0.030	(0.929)	-4.615***	(1.189)
Mother works full-time	0.346	(0.807)	-2.669***	(1.023)
Father works full-time	-0.069	(1.033)	0.225	(1.250)
Having siblings	-0.736	(1.139)	1.112	(1.235)
At least one parent with university degree	-0.510	(0.821)	-2.910***	(1.046)
Patience median or above	0.112	(0.819)	1.158	(1.051)
Risk median or above	0.557	(0.820)	0.187	(1.044)
Mean of the outcome	27.056		9.074	
Observations	540		540	

Notes: Bivariate OLS regressions, each column shows results from a different regression of the respective outcome on the respective sociodemographic characteristic. Dependent variables: (1) hours expected to work per week with child 1–6 (0=0 hours, i.e. not at all; 10= 10 hours; 20= 20 hours, i.e. part-time; 30=30 hours, 40=40 hours, i.e. full-time); (2) responses to labor supply for both parents (self and partner) with higher values indicating higher labor market supply of men relative to women. Sample: respondents in the control group. Robust standard errors in parentheses. ***/**/* indicate significance at the 1%/5%/10% level.

Table A2.7: Heterogeneous Treatment Effects on Labor-Supply Expectations by Gender

	Self-expected labor supply	Expected within-family gender gap
	(1)	(2)
Panel A: Combined treatment effects		
Treatments (pooled)	-1.805***	-1.453
	(0.658)	(0.975)
Treatments (pooled) x female	-0.675	3.811***
	(0.889)	(1.243)
Female	-7.333***	-3.358***
	(0.749)	(1.045)
Panel B: Seperate treatment effects		
Salience	-1.286	-1.191
	(0.845)	(1.278)
Information	-1.648*	-1.995
	(0.871)	(1.293)
Salience & information	-2.366***	-1.210
	(0.798)	(1.201)
Salience x female	-0.449	3.476**
	(1.120)	(1.577)
Information x female	-0.836	4.113**
	(1.165)	(1.620)
Salience & information x female	-0.824	3.883**
	(1.073)	(1.533)
Female	-7.332***	-3.363***
	(0.750)	(1.046)
Control mean	31.13	10.88
Covariates	Yes	Yes
Observations	2000	2000
R-squared	0.193	0.049

Notes: OLS regressions. Panel A: treatments (pooled): respondent in experimental groups *salience*, *information* or *salience and information*. Panel B: *salience*, *information* and *salience & information* indicate membership of respective treatment groups. Dependent variables: (1) hours expected to work per week when having a child 1–6 (0=0 hours, i.e. not at all; 10= 10 hours; 20= 20 hours, i.e. part-time; 30=30 hours, 40=40 hours, i.e. full-time); (2) responses to labor supply for both parents (self and partner) with higher values indicating higher labor market supply of men relative to women. Control mean: mean of the outcome variable in the control group. See Table 2.1 for included covariates. Robust standard errors in parentheses. ***/**/* indicate significance at the 1%/5%/10% level.

Table A2.8: Persistence of Separate Treatment Effects on Self-Expected Labor Supply (Follow-up Sample)

	Self-expected labor supply		
	(1) All	(2) Girls	(3) Boys
Salience	-2.034***	-2.613***	-1.845*
	(0.698)	(0.954)	(0.971)
Information	-2.042***	-2.548***	-1.630*
	(0.690)	(0.964)	(0.963)
Salience & information	-3.157***	-3.323***	-2.981***
	(0.648)	(0.896)	(0.920)
Salience x follow-up	2.490***	2.258**	2.774**
	(0.714)	(0.922)	(1.137)
Information x follow-up	1.461**	1.694*	1.108
	(0.694)	(0.958)	(1.008)
Salience & information x follow-up	1.970***	1.793*	2.177**
	(0.731)	(1.007)	(1.072)
Follow-up	-0.731	-0.601	-0.881
	(0.482)	(0.667)	(0.704)
Salience in follow-up	0.456	-0.355	0.929
Information in follow-up	-0.581	-0.854	-0.522
Salience & information in follow-up	-1.187*	-1.531	-0.804
Control mean	27.19	23.66	31.26
Covariates	Yes	Yes	Yes
Observations (respondents)	1319	734	585
R-squared	0.202	0.114	0.033

Notes: OLS regressions. *Salience*, *information* and *salience & information* indicate membership of respective treatment groups. Dependent variable: hours expected to work per week when having a child 1–6 (0=0 hours, i.e. not at all; 10= 10 hours; 20= 20 hours, i.e. part-time; 30=30 hours, 40=40 hours, i.e. full-time). *Salience in follow-up*/*information in follow-up*/*salience & information in follow-up* are the linear combinations of the coefficients on the *respective treatment indicators* plus *respective treatment indicator x follow-up*. Control mean: mean of the outcome variable in the omitted group (i.e. control group and treatment *salience*) reported in the main survey. See Table 2.1 for included covariates. Sample: Follow-up survey respondents. Robust standard errors, adjusted for clustering at the respondent level, in parentheses. ***/**/* indicate significance at the 1%/5%/10% level.

Table A2.9: Persistence of Information Treatment Effects on Beliefs about Norms Prescribing Parental Labor Supply (Follow-up Sample)

	Relative belief			
	(1) All	(2) Girls	(3) Boys	(4) All
Panel A: Social norm towards mothers elicited in follow-up				
Information provision (pooled)	0.025**	0.034**	0.013	0.014
	(0.012)	(0.016)	(0.018)	(0.018)
Information provision (pooled) x female				0.019
				(0.024)
Female				0.001
				(0.017)
Control mean	0.73	0.73	0.73	0.73
Covariates	Yes	Yes	Yes	Yes
Observations	1308	731	577	1308
R-squared	0.017	0.035	0.029	0.017
Panel B: Social norm towards fathers elicited in follow-up				
Information provision (pooled)	0.058**	0.038	0.086**	0.085**
	(0.027)	(0.037)	(0.041)	(0.040)
Information provision (pooled) x female				-0.049
				(0.055)
Female				0.100**
				(0.039)
Control mean	0.87	0.91	0.82	0.87
Covariates	Yes	Yes	Yes	Yes
Observations	1296	728	568	1296
R-squared	0.040	0.045	0.048	0.040

Notes: OLS regressions. Info provision: respondent in the experimental groups *information* or *salience and information*. Dependent variables: Panel A: belief about social norm towards mothers relative to correct value (=91). Panel B: belief about social norm towards fathers relative to correct value (=41). Control mean: mean of the outcome variable in the omitted group (i.e. control group *and* belief elicitation only). See Table 2.1 for included covariates. Sample: follow-up survey respondents. Robust standard errors in parentheses. ***/**/* indicate significance at the 1%/5%/10% level.

Table A2.10: Seperate Treatment Effects on Incentivized Outcomes

	University education more important for boy.	Children suffer if mother works for pay.	Being a housewife as fulfilling as working for pay.	Causes problems if a woman earns more than her husband.
	(1)	(2)	(3)	(4)
Panel A: Girls				
Salience	0.016	-0.014	-0.011	0.013
	(0.117)	(0.058)	(0.034)	(0.107)
Information	0.437***	0.104*	0.086**	0.176*
	(0.117)	(0.058)	(0.034)	(0.105)
Salience & information	0.494***	0.231***	0.095***	0.288***
	(0.118)	(0.058)	(0.034)	(0.107)
p-values: MHT Correction				
Salience	0.989	0.999	1.000	0.900
Information	0.000	0.634	0.134	0.664
Salience & information	0.000	0.000	0.072	0.126
Control mean	2.266	1.629	0.679	2.154
Covariates	Yes	Yes	Yes	Yes
Observations	1130	1156	1143	1137
R-squared	0.041	0.057	0.046	0.025
Panel B: Boys				
Salience	-0.097	-0.107	-0.007	-0.140
	(0.138)	(0.068)	(0.042)	(0.124)
Information	0.412***	0.091	0.043	0.161
	(0.146)	(0.070)	(0.042)	(0.125)
Salience & information	0.691***	0.233***	0.086**	0.369***
	(0.138)	(0.066)	(0.039)	(0.118)
p-values: MHT Correction				
Salience	0.973	0.709	0.999	0.886
Information	0.069	0.843	0.911	0.856
Salience & information	0.000	0.013	0.299	0.014
Control mean	2.329	1.634	0.683	2.040
Covariates	Yes	Yes	Yes	Yes
Observations	811	829	820	807
R-squared	0.073	0.066	0.059	0.045

Notes: OLS regressions. *Salience*, *information* and *salience & information* indicate membership of respective treatment groups. Dependent variables: beliefs about share of Germans agreeing with the statements that (1) a university education is more important for a boy than for a girl relative to correct value (=16); (2) the children suffer if the mothers works for pay relative to correct value (=33); (3) being a housewife is just as fulfilling as working for pay relative to correct value (=55); (4) it is almost certain to cause problems if a woman earns more than her husband relative to correct value (=20). Results (not shown) from full interaction model between gender and treatment indicator reveals significant heterogeneous treatment effects by gender. Control mean: mean of the outcome variable in the control group. See Table 2.1 for included covariates. MHT Correction refers to the multiple hypothesis testing procedure presented in List et al. (2019) and corrects for multiple subgroups (girls and boys), multiple treatments as well as multiple outcomes (all 4 outcomes listed). Robust standard errors in parentheses. ***/**/* indicate significance at the 1%/5%/10% level.

Table A2.11: Treatment Effects on Labor-Supply Expectations Without Child

	Self-expected labor supply	Expected within-family gender gap
	(1)	(2)
Panel A: Girls		
Treatments (pooled)	-0.468	0.267
	(0.531)	(0.360)
p-values: MHT Correction		
Treatments (pooled)	0.375	0.714
Control mean	34.37	1.09
Observations	1164	1164
R-squared	0.023	0.014
Panel B: Boys		
Treatments (pooled)	0.538	-0.391
	(0.581)	(0.631)
p-values: MHT Correction		
Treatments (pooled)	0.590	0.539
Control mean	34.45	3.49
Observations	836	836
R-squared	0.026	0.023
Panel C: All		
Treatments (pooled)	0.002	0.050
	(0.391)	(0.340)
Control mean	34.41	2.15
Observations	2000	2000
R-squared	0.021	0.031
Covariates	Yes	Yes

Notes: OLS regressions. Treatments (pooled): respondent in experimental groups *salience*, *information* or *salience and information*. Dependent variables: (1) hours expected to work per week without child (0=0 hours, i.e. not at all; 10= 10 hours; 20= 20 hours, i.e. part-time; 30=30 hours, 40=40 hours, i.e. full-time); (2) responses to labor supply without child for both spouses (self and partner) with higher values indicating higher labor market supply of men relative to women. Control mean: mean of the outcome variable in the control group. See Table 2.1 for included covariates. MHT Correction refers to the multiple hypothesis testing procedure presented in List et al. (2019) and corrects for two subgroups (girls and boys). Robust standard errors in parentheses. ***/**/* indicate significance at the 1%/5%/10% level.

Table A2.12: Separate Treatment Effects of the More Egalitarian Norm on the Expected Within-Family Gender Gap (Follow-Up Sample)

	Expected within-family gender gap		
	(1) All	(2) Girls	(3) Boys
Egalitarian information	-1.066**	-1.078*	-1.049*
	(0.426)	(0.584)	(0.622)
Salience & egalitarian information	-1.667***	-1.503**	-1.880*
	(0.590)	(0.660)	(1.055)
p-values: MHT Correction			
Egalitarian information	0.013	0.066	1.000
Salience & egalitarian information	0.009	0.063	1.000
Individual FEs	Yes	Yes	Yes
Observations (respondents)	1319	734	585
R-squared	0.011	0.011	0.010

Notes: OLS regressions. *Egalitarian information* and *salience & egalitarian information* indicate membership of respective treatment groups. Dependent variable: responses to labor supply for both parents (self and partner) with higher values indicating higher labor market supply of men relative to women. Results (not shown) from interaction model between gender and treatment indicator does not reveals significant heterogeneous treatment effects by gender. Robust standard errors, adjusted for clustering at the respondent level, in parentheses. ***/**/* indicate significance at the 1%/5%/10% level.

Online Appendix

Results from the ifo Education Survey 2018

The experiment presented in the main part of the chapter is an extended version of a previous experiment that was conducted with 1,085 adolescents within the scope of the ifo Education Survey 2018.[1] The main purpose of the ifo Education Survey was to inform the German policy debate about the opinion of adolescents on education-policy topics that are unrelated to social norms or labor-supply expectations. At the end of the questionnaire (which encompassed a total of 30 questions on education policy), we ran a similar but substantially shortened version of the experiment on the social norm prescribing parental labor supply. Particularly, the experiment randomized respondents into one of two experimental groups (the control group and treatment *salience and information*), and then elicited labor-supply expectations. Online Appendix Table O2.4 presents the results. Treatment effects in this alternative sample are remarkably similar to—and statistically indistinguishable from—those in our main sample (see Table 2.2 for comparison). Given the importance of replication for the credibility of scientific findings (e.g., Maniadis et al., 2014), it is also reassuring that our treatment effects replicate in two independent samples.

Perceived Peer Pressure as Potential Mechanism

A potential complementary mechanism through which the gender norm alters labor-market expectations is perceived peer pressure: Individuals may adhere to the social norm to meet their peer groups expectations, and thereby avoid peer punishment of non-conform behavior (e.g., Benabou and Tirole, 2011).[2]

To study the empirical relevance of this channel, we elicited respondents' beliefs about whether the following groups think that mothers and fathers, respectively, should reduce their labor supply when the children are young: (i) family, (ii) friends, and (iii) the (future) partner. For the analysis we combine these items into a z-standardized index measuring peer-group expectations (Kling et al., 2007). To gauge the relevance of these different peer groups for

[1] The ifo Education Survey is an annual opinion survey on education policy among representative samples of adults in Germany (see `https://www.ifo.de/en/survey/ifo-education-survey`). In the 2018 wave, the general-population sample was complemented by a sample of adolescents that was surveyed about education topics (see Woessmann et al., 2018). The sampling and polling was done by the polling firm *Kantar Public*, which drew respondents from a different subject pool than the one used for our main study.

[2] There are at least two further potential reason for why individuals follow norms, namely (i) because they derive direct utility from actions that maintain their identity-based self-concepts (e.g., Akerlof and Kranton, 2000) and (ii) because of 'best-practice considerations' in the sense that respondents may think that norms entail meaningful information on how to achieve certain outcomes most effectively (e.g., Cialdini and Trost, 1998; Cialdini and Goldstein, 2004). Focusing on child development and family income as two such outcomes, we find no evidence that the latter channel is relevant in our setting (results available upon request).

individual respondents, we additionally asked them how important it is for them to meet the expectations of each of these four groups.[3]

Treated girls, but not boys are more likely to believe that their peers expect mothers to reduce their labor supply. In columns 1 and 2 of Online Appendix Table O2.5, we regress the indices of respondents' beliefs about their peer groups' norms towards mothers and fathers, respectively, on the pooled treatment indicator. The treatments increase girls' beliefs that their peer groups demand that mothers reduce their labor supply by 15 percent of a standard deviation, but do not affect their corresponding beliefs regarding fathers. For boys, we find no significant pooled treatment effects.

Next, we study the connection between the importance that adolescents assign to their peer groups' opinions and labor-supply expectations. Online Appendix Table O2.6 regresses girls' and boys' self-expected labor supply on the pooled treatment indicator and its interaction with the importance respondents assign to their peer groups' expectations.[4] The coefficient on the interaction term reveals that treatment effects for girls (but not for boys) are stronger the more important they consider peer expectations.

Put together, these two sets of results suggest that treatment effects for girls may in fact be driven by their desire to adhere to their peer groups' norms. The peer-pressure channel seems to be less relevant for boys.

Importance of Different Job Attributes

To investigate whether the traditional norm affects further labor-market-related preferences that are only indirectly linked to labor supply as young parents, we also study treatment effects on adolescents' preferences for future job attributes. For that purpose, respondents rated the following eight different job attributes on a five-point scale from 'very important' to 'very unimportant': 'The job …' (i) '… can be reconciled with children.', (ii) '… enables part-time employment.', (iii) '… offers a high salary.', (iv) '… offers good career opportunities.', (v) '… offers job security.', (vi) '… is challenging.', (vii) '… gives me enough leisure time', (viii) '… is enjoyable.' In the control group, female respondents consider reconciliation with children and the possibility to work part-time more important than males (see Online Appendix Table O2.7), which is in line with Wiswall and Zafar (2020)'s finding on higher willingness to pay for work flexibility among females. Regressing the importance of job attributes on the pooled treatments reveals that the social norm decreases the importance that females assign to the reconciliation of job and children.[5] While we would expect the treatments to increase and not

[3] Beliefs about the peer groups' views as well as the groups' importance were elicited after treatment administration. We do not find any treatment effects on stated importance of the different peer groups (results available upon request).

[4] In the regressions, we again computed an index of the importance assigned to the different peer groups using the procedure by Kling et al. (2007).

[5] For the regressions, we z-standardized the five-point scale outcomes.

decrease the importance of this factor if it makes girls more likely to expect working part-time, it can be rationalized by the fact that females downward-adjust their fertility expectations in response to the treatment, which in turn makes reconciliation of work and children less important (results available upon request). The regression results in Online Appendix Table O2.7 suggest that social-norm considerations have limited overall effects on labor-market preferences.

Table O2.1: Treatment Effects on Labor-Supply Expectations among Girls: All Answer Categories

	(1) 0 hours	(2) 10 hours	(3) 20 hours	(4) 30 hours	(5) 40 hours
Panel A: Self-expected labor supply					
Treatments (pooled)	0.034***	0.042***	0.035***	-0.068***	-0.042***
	(0.008)	(0.010)	(0.009)	(0.015)	(0.010)
Salience	0.026***	0.032***	0.027***	-0.052***	-0.032***
	(0.010)	(0.012)	(0.010)	(0.019)	(0.012)
Information	0.032***	0.040***	0.033***	-0.065***	-0.041***
	(0.010)	(0.012)	(0.010)	(0.019)	(0.012)
Salience & information	0.042***	0.052***	0.043***	-0.085***	-0.053***
	(0.010)	(0.012)	(0.011)	(0.019)	(0.013)
Control mean	0.04	0.07	0.49	0.29	0.12
Covariates	Yes	Yes	Yes	Yes	Yes
Observations	1164	1164	1164	1164	1164
Panel B: Expected labor supply for partner					
Treatments (pooled)	0.001	0.003	0.011	0.005	-0.019
	(0.002)	(0.004)	(0.015)	(0.006)	(0.028)
Salience	-0.001	-0.002	-0.005	-0.002	0.010
	(0.002)	(0.005)	(0.019)	(0.008)	(0.034)
Information	0.002	0.004	0.014	0.006	-0.026
	(0.002)	(0.005)	(0.018)	(0.008)	(0.034)
Salience & information	0.003	0.006	0.022	0.009	-0.040
	(0.002)	(0.005)	(0.019)	(0.008)	(0.034)
Control mean	0.00	0.03	0.17	0.40	0.39
Covariates	Yes	Yes	Yes	Yes	Yes
Observations	1164	1164	1164	1164	1164

Notes: Results from an ordered probit model. The table reports the average marginal treatment effects. Treatment (pooled): respondents in experimental groups *salience*, *information* and *salience and information*. Dependent variable is the answer to the question: *'Imagine you are 30 years old and you have a child aged between 1 and 6 years with your partner. What do you think, how many hours per week on average would you like to work in order to earn money?'* (Panel A) or *'And how many hours per week on average would you like your partner to work in order to earn money?'* (Panel B). Control mean: mean of the outcome variable in the control group. See Table 2.1 for included covariates. Sample: girls. Robust standard errors in parentheses. ***/**/* indicate significance at the 1%/5%/10% level.

Table O2.2: Treatment Effects on Labor-Supply Expectations among Boys: All Answer Categories

	(1) 0 hours	(2) 10 hours	(3) 20 hours	(4) 30 hours	(5) 40 hours
Panel A: Self-expected labor supply					
Treatments (pooled)	0.008**	0.013**	0.053***	0.005	-0.081***
	(0.004)	(0.005)	(0.019)	(0.004)	(0.028)
Salience	0.006	0.010	0.040*	0.004	-0.060*
	(0.004)	(0.006)	(0.024)	(0.004)	(0.036)
Information	0.007*	0.011*	0.043*	0.004	-0.066*
	(0.004)	(0.006)	(0.024)	(0.004)	(0.036)
Salience & information	0.012**	0.018***	0.072***	0.007	-0.109***
	(0.005)	(0.007)	(0.023)	(0.005)	(0.034)
Control mean	0.01	0.03	0.16	0.45	0.35
Covariates	Yes	Yes	Yes	Yes	Yes
Observations	836	836	836	836	836
Panel B: Expected labor supply for partner					
Treatments (pooled)	0.007	0.006	0.000	-0.008	-0.006
	(0.014)	(0.013)	(0.001)	(0.016)	(0.012)
Salience	0.002	0.001	0.000	-0.002	-0.001
	(0.018)	(0.016)	(0.001)	(0.020)	(0.015)
Information	-0.010	-0.009	-0.001	0.011	0.008
	(0.018)	(0.016)	(0.001)	(0.020)	(0.015)
Salience & information	0.027	0.023	0.002	-0.030	-0.022
	(0.017)	(0.015)	(0.002)	(0.019)	(0.014)
Control mean	0.11	0.17	0.39	0.23	0.09
Covariates	Yes	Yes	Yes	Yes	Yes
Observations	836	836	836	836	836

Notes: Results from an ordered probit model. The table reports the average marginal treatment effects. Treatment (pooled): respondents in experimental groups *salience*, *information* and *salience and information*. Dependent variable is the answer to the question: *'Imagine you are 30 years old and you have a child aged between 1 and 6 years with your partner. What do you think, how many hours per week on average would you like to work in order to earn money?'* (Panel A) or *'And how many hours per week on average would you like your partner to work in order to earn money?'* (Panel B). Control mean: mean of the outcome variable in the control group. See Table 2.1 for included covariates. Sample: boys. Robust standard errors in parentheses. ***/**/* indicate significance at the 1%/5%/10% level.

Table O2.3: Heterogeneous Treatment Effects on Self-Expected Labor Supply by Sociodemographic Characteristics

	Self-expected labor supply		
	(1) All	(2) Girls	(3) Boys
Region			
Treatments (pooled)	-2.108***	-2.425***	-1.769**
	(0.504)	(0.683)	(0.735)
Treatments (pooled) x living in east Germany	-0.392	-0.952	-0.214
	(1.057)	(1.384)	(1.552)
City size			
Treatments (pooled)	-1.301	-2.167**	0.080
	(0.805)	(1.067)	(1.121)
Treatments (pooled) x city size ≥ 100,000	-1.295	-0.644	-2.803**
	(0.964)	(1.292)	(1.376)
Age			
Treatments (pooled)	-2.421***	-3.078***	-1.660
	(0.708)	(0.944)	(1.019)
Treatments (pooled) x aged 16 or 17	0.402	0.801	-0.269
	(0.909)	(1.228)	(1.319)
Educational attainment			
Treatments (pooled)	-2.771***	-2.830**	-2.538**
	(0.855)	(1.232)	(1.144)
Treatments (pooled) x school to uni degree	0.684	0.068	1.115
	(1.008)	(1.419)	(1.416)
Mothers' employment			
Treatments (pooled)	-2.510***	-2.775***	-2.202**
	(0.595)	(0.820)	(0.858)
Treatments (pooled) x mother w. full-time	0.765	0.381	0.946
	(0.889)	(1.199)	(1.303)
Parental education			
Treatments (pooled)	-2.774***	-2.924***	-2.445***
	(0.561)	(0.745)	(0.829)
Treatments (pooled) x parents w. uni degree	1.562*	0.899	1.546
	(0.915)	(1.267)	(1.320)
Recruitment			
Treatments (pooled)	-1.518***	-1.583**	-1.777**
	(0.540)	(0.680)	(0.883)
Treatments (pooled) x recruited via parents	-1.654*	-2.796**	-0.194
	(0.939)	(1.357)	(1.308)

Notes: OLS regressions. Treatments (pooled): respondent in experimental groups *salience*, *information* or *salience and information*. Dependent variable: hours expected to work per week when having a child 1–6 (0=0 hours, i.e. not at all; 10= 10 hours; 20= 20 hours, i.e. part-time; 30=30 hours, 40=40 hours, i.e. full-time). Living in east Germany/ city size / aged 16 or 17 / school to uni degree / mother w. full-time / parents w. uni degree / recruited via parents: Respondent belongs to respective subgroup. See Table 2.1 for included covariates. Robust standard errors in parentheses. ***/**/* indicate significance at the 1%/5%/10% level.

Table O2.4: Treatment Effects on Self-Expected Labor Supply in the ifo Education Survey

	Self-expected labor supply
	(1)
Panel A: Girls	
Salience & information	-3.114***
	(0.814)
p-values: MHT Correction	
Treatments (pooled)	0.000
Control mean	23.75
Observations	553
R-squared	0.074
Panel B: Boys	
Salience & information	-2.179***
	(0.764)
p-values: MHT Correction	
Treatments (pooled)	0.003
Control mean	31.52
Observations	532
R-squared	0.062
Panel C: All	
Salience & information	-2.566***
	(0.558)
Control mean	27.39
Observations	1085
R-squared	0.200
Covariates	Yes

Notes: OLS regressions. *Salience & information*: respondent in respective treatment group. Dependent variable: hours expected to work per week when having a child 1–6 (0=0 hours, i.e. not at all; 10= 10 hours; 20= 20 hours, i.e. part-time; 30=30 hours, 40=40 hours, i.e. full-time). Control mean: mean of the outcome variable in the control group. See Table 2.1 for included covariates. MHT Correction refers to the multiple hypothesis testing procedure presented in List et al. (2019) and corrects for multiple subgroups (girls and boys) in Panel A and B. Sample: 2018 survey participants. Robust standard errors in parentheses. ***/**/* indicate significance at the 1%/5%/10% level.

Table O2.5: Treatment Effects on Perceived Peer Pressure

	Index: Peers' opinion about labor supply of mothers	Index: Peers' opinion about labor supply of fathers
	(1)	(2)
Panel A: Girls		
Treatments (pooled)	0.133**	0.042
	(0.065)	(0.063)
p-values: MHT Correction		
Treatments (pooled)	0.163	0.873
Covariates	Yes	Yes
Observations	1163	1163
R-squared	0.052	0.033
Panel B: Boys		
Treatments (pooled)	0.032	-0.022
	(0.073)	(0.077)
p-values: MHT Correction		
Treatments (pooled)	0.874	0.772
Covariates	Yes	Yes
Observations	836	835
R-squared	0.051	0.031

Notes: OLS regressions. Treatments (pooled): respondent in experimental groups *salience*, *information* or *salience and information*. Dependent variables: (1) index summarizing respondents' belief about opinion of parents, friends and (future) partner on appropriate labor market supply for mothers following the methodology in Kling et al. (2007); (2) index summarizing respondents' belief about opinion of parents, friends and (future) partner on appropriate labor market supply for fathers following the methodology in Kling et al. (2007). Results (not shown) from full interaction model between gender and treatment indicators do not reveal any heterogeneous treatment effects by gender. See Table 2.1 for included covariates. MHT Correction refers to the multiple hypothesis testing procedure presented in List et al. (2019) and corrects for multiple subgroups (girls and boys) and multiple outcomes (all 2 outcomes listed). Robust standard errors in parentheses. ***/**/* indicate significance at the 1%/5%/10% level.

Table O2.6: Heterogeneous Treatment Effects on Self-Expected Labor Supply by Respondents' Importance to Conform to Peers' Expectations

	Self-expected labor supply	
	(1)	(2)
	Girls	Boys
Treatments (pooled)	-2.702***	-1.839***
	(0.599)	(0.653)
Treatments (pooled) x conformity index	-1.170*	0.555
	(0.608)	(0.687)
Conformity index	0.967*	-0.378
	(0.527)	(0.562)
Covariates	Yes	Yes
Observations	1164	836
R-squared	0.102	0.039

Notes: OLS regressions. Treatments (pooled): respondent in experimental groups *salience*, *information* or *salience and information*. Dependent Variable: Hours expected to work per week when having a child 1–6 (0=0 hours, i.e. not at all; 10= 10 hours; 20= 20 hours, i.e. part-time; 30=30 hours, 40=40 hours, i.e. full-time). Conformity index: index summarizing respondents' stated importance to conform to expectations of parents, friends and (future) partner following the methodology in Kling et al. (2007). See Table 2.1 for included covariates. Robust standard errors in parentheses. ***/**/* indicate significance at the 1%/5%/10% level.

Table O2.7: Treatment Effects on Preferences for Job Attributes

	Reconcil. with children	Enables part-time work	High salary	Good career opport.	Job security	Leisure time	Enjoyable	Challenge
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A: Girls								
Treatments (pooled)	-0.152**	-0.027	0.022	-0.011	0.003	-0.040	0.066	0.026
	(0.066)	(0.062)	(0.063)	(0.063)	(0.058)	(0.064)	(0.059)	(0.066)
p-values: MHT Correction								
Treatments (pooled)	0.287	1.000	1.000	1.000	0.999	1.000	0.983	1.000
Control importance	0.83	0.72	0.91	0.83	0.96	0.91	0.96	0.74
Covariates	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	1164	1164	1163	1164	1164	1164	1164	1164
R-squared	0.040	0.033	0.022	0.068	0.030	0.023	0.021	0.038
Panel B: Boys								
Treatments (pooled)	-0.037	-0.031	0.037	0.055	-0.008	-0.012	0.013	-0.001
	(0.078)	(0.079)	(0.074)	(0.073)	(0.074)	(0.079)	(0.073)	(0.075)
p-values: MHT Correction								
Treatments (pooled)	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.993
Control importance	0.76	0.53	0.91	0.83	0.95	0.88	0.95	0.78
Covariates	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	836	835	836	834	835	836	836	835
R-squared	0.031	0.031	0.020	0.043	0.031	0.029	0.024	0.042

Notes: OLS regressions. Treatments (pooled): respondent in experimental groups *salience*, *information* or *salience and information*. Dependent variables: (1)-(8) respondents' stated importance of respective job attribute on a 5 point-scale, standardized (the higher the value, the more important the respective job preference). Results (not shown) from full interaction model between gender and treatment indicators do not reveal any heterogeneous treatment effects by gender. Control importance: share of respondents in the control group reporting respective job preference to be (very) important. See Table 2.1 for included covariates. MHT Correction refers to the multiple hypothesis testing procedure presented in List et al. (2019) and corrects for multiple subgroups (girls and boys) and multiple outcomes (all 8 outcomes listed). Robust standard errors in parentheses. ***/**/* indicate significance at the 1%/5%/10% level.

3 COVID-19 and Educational Inequality: How School Closures Affect Low- and High-Achieving Students *

3.1 Introduction

To inhibit the spread of the COVID-19 pandemic, many countries closed their schools for several months during the first half of 2020. These closures affected over 90 percent of school children (1.5 billion) worldwide (UNESCO, 2020b). A defining feature of school closures is that students do not have the same support of teachers as in traditional in-person classroom teaching. Many have argued that the school closures may increase inequality between children from different family backgrounds (e.g., European Commission, 2020; UNESCO, 2020a). But another dimension of inequality that may be particularly relevant for school closures is the one between low- and high-achieving students. Out-of-school learning implies a large amount of self-regulated learning where students must independently acquire and understand the academic content without the support of trained educators. While self-regulated learning may be feasible for high-achieving students during school closures, it may be especially challenging for low-achieving students. In this chapter, we provide evidence on how the COVID-19 school closures affected the learning time and other activities of low- and high-achieving students and how parents and schools differentially compensated for the closures.

The COVID-19-related school closures, and the associated temporary discontinuation of traditional in-person teaching, represent an unprecedented disruption of students' educational careers. From an educational production perspective, the school closures induced a sharp decline in what is probably the most important school input factor to produce educational achievement: the support of trained educators. Teachers provide the traditional teaching activities such as explaining new material or providing learning-stimulating feedback. Ample evidence shows that teachers are a key ingredient for students' educational success (e.g., Rivkin et al., 2005). Our data show that direct contact with teachers evaporated during the school closures in Germany, as in many other countries (e.g., Andrew et al., 2020, for England). Instead, students mostly had to embark on self-regulated learning. Since skill formation is a process of dynamic complementarities in the sense that basic skills are necessary to acquire additional skills (e.g., Cunha et al., 2006; Cunha and Heckman, 2007), students with lower initial achievement may lack the knowledge and skill base necessary to generate additional learning gains through self-regulated learning. Consequently, if returns to time invested in independent learning activities are sufficiently low, low-achieving students will spend less time

* This chapter is joint work with Philipp Lergetporer, Katharina Werner, Ludger Woessmann and Larissa Zierow. It is based on the paper 'COVID-19 and Educational Inequality: How School Closures Affect Low- and High-Achieving Students', *CESifo Working Paper*, 2020.

on school-related activities, substituting other activities that are relatively more rewarding to them.

To test this hypothesis, we designed and ran an online survey of 1,099 parents of school-aged children in Germany in June 2020. In our detailed time-use data, we carefully elicit how many hours students spent with a range of activities per day both before and during the school closures. We distinguish between (i) *school-related* activities such as going to school or learning at home; (ii) activities generally deemed *conducive* to child development such as reading, arts, playing music, or doing sports; and (iii) activities generally deemed *detrimental* to child development such as watching TV, playing computer games, or consuming social media.[1] The retrospective panel structure of our data allows us to investigate how the closures affected the gap in learning time between low- and high-achieving students, categorized by their prior school grades. To further investigate the extent to which parents and schools compensated for changes in learning time, we additionally elicited parental involvement in home-schooling activities as well as detailed information on schools' distance-teaching activities. Complementing our analysis of inequality along the achievement dimension, we also analyze the learning-gap change between children from different family backgrounds and by gender.

We find that the school closures had a large negative impact on learning time, particularly for low-achieving students. Overall, students' learning time more than halved from 7.4 hours per day before the closures to 3.6 hours during the closures. While learning time did not differ between low- and high-achieving students before the closures, high-achievers spent a significant 0.5 hours per day more on school-related activities during the school closures than low-achievers. Most of the gap cannot be accounted for by observables such as socioeconomic background or family situation, suggesting that it is genuinely linked to the achievement dimension. Time spent on conducive activities increased only mildly from 2.9 hours before to 3.2 hours during the school closures. Instead, detrimental activities increased from 4.0 to 5.2 hours. This increase is more pronounced among low-achievers (+1.7 hours) than high-achievers (+1.0 hour). Taken together, our results imply that the COVID-19 pandemic fostered educational inequality along the achievement dimension.

The COVID-19-induced learning gap between low- and high-achieving students was not compensated by parents' activities. Already before the school closures, parents of low-achievers spent less learning time together with their children than parents of high-achievers (0.4 versus 0.6 hours per day). The school closures only exacerbated this inequality in parental involvement, as parents of low-achievers increased their time investment in joint learning by less than parents of high-achievers (+0.5 versus +0.6 hours).

[1] Time spent on educational activities has been shown to be the most productive input for cognitive skill development among different activities of children (Fiorini and Keane, 2014). Our further categorization is in line with parents' beliefs about how beneficial the different activities are for their children's development (section 3.3.2).

The activities of schools did not compensate for the learning gap between low- and high-achieving students either. During the school closures, schools and teachers only carried out a fraction of their usual teaching activities via distance teaching. For instance, only 29 percent of students had shared lessons for the whole class (e.g., by video call) more than once a week, and only 17 percent had individual contact with their teacher more than once a week. This reduction in school activities hit low-achieving students particularly hard: Compared to high-achievers, low-achievers were 13 percentage points less likely to have online lessons and 10 percentage points less likely to have individual teacher contacts more than once a week.

Looking at other dimensions of educational inequality, the COVID-19 school closures did not increase learning-time gaps by parental education, but they affected boys more than girls. While children with a university-educated parent spent significantly more time learning for school than those without a university-educated parent before the school closures, we do not find a significant difference in the reduction in learning time between both groups in response to the closures. However, school support was significantly lower for children without a university-educated parent, which suggests that the school closures may also have amplified socioeconomic inequality in educational achievement. Compared to girls (-3.5 hours), the COVID-19-induced learning disruption was more pronounced for boys (-4.0 hours), who particularly spent more time playing computer games.

By documenting how the discontinuation of in-person teaching differentially affects low- and high-achieving students, we contribute to the broad literatures on educational production (e.g., Hanushek, 2020), skill formation (e.g., Cunha and Heckman, 2007), and educational inequality (e.g., Björklund and Salvanes, 2011). Our results complement the English time-use study during COVID-19 by Andrew et al. (2020) by investigating inequality along the achievement dimension as well as compensating activities of parents and schools. Our study of a range of substituted conducive and detrimental activities also complements several other contemporaneous studies on how COVID-19-induced school closures affected learning inputs and outcomes such as online learning (e.g., Chetty et al., 2020 for online lesson completion and Bacher-Hicks et al., 2021 or household search for online learning resources in the United States) and standardized tests (e.g., Maldonado and De Witte, 2020, for Flemish Belgium and Engzell et al., 2020 for the Netherlands), neither of which has a focus on differential effects by the achievement dimension.[2] Our findings contribute to the rapidly emerging literature on effects of the COVID-19 pandemic on other economic and social outcomes such as labor markets, families, and well-being (e.g., Alon et al., 2020; Chetty et al., 2020; Fetzer et al., 2020).

The remainder of this chapter is structured as follows. Section 3.2 provides a brief conceptual framework and institutional background on schooling during the COVID-19 pandemic in Germany. Section 3.3 introduces our data and research design. Section 3.4 presents results on how the COVID-19 school closures affected learning and other activities of low- and high-

[2] For additional descriptive evidence on overall learning engagement of students during the school closures in Germany in specific samples, see Anger et al. (2020) and Huber and Helm (2020).

achieving students. Section 3.5 presents results on support structures by parents and schools. Section 3.6 reports results on differences by parental education background, child gender, and school type as additional dimensions of inequality. Section 3.7 discusses the findings, and section 3.8 concludes.

3.2 Conceptual Framework and Institutional Background

This section provides a conceptual framework (section 3.2.1) and institutional background (section 3.2.2).

3.2.1 School Closures in the Framework of an Education Production Function

To frame ideas, we conceptualize the potential effects of school closures on educational inequality in the framework of a standard education production function (e.g., Hanushek, 1986, 2020). The production of educational output is expressed as a function f of student ability A, family inputs F, and school inputs S:

$$\Delta Y_i = f(A_i, F_i, S_i) \tag{3.1}$$

where ΔY_i is the change in educational output, or learning, of student i. While educational output can be conceived generally as the acquisition of skills, ΔY_i will be approximated by student i's daily learning time in our empirical application. We will discuss the implications of this approximation for the interpretation of changes in educational inequality below.

In this framework, school closures can be thought of as a reduction in school inputs S_i. Specifically, a defining feature of school closures is that there is no teacher in the room to help students with their learning. As teachers are probably the most important school input factor for student learning (e.g., Hanushek, 1986; Rivkin et al., 2005; Chetty et al., 2014), students are missing out on key support, and their learning is left more to the discretion of themselves and their families. In standard applications, the education production function is often simplified to be additive in the different inputs. In this case, the effect of a uniform change in school inputs would have the same effect on children from different family backgrounds and different ability levels, thereby leaving educational inequality unaffected.

For school closures to affect educational inequality, either the amount or the production elasticities of the other inputs must depend on the extent of school inputs.[3] One often hy-

[3] The exposition here assumes that school closures entail the same reduction in school inputs for all students. Another way in which school closures could affect educational inequality is that the decline in effective school inputs may differ for different students, e.g., when high-SES parents are more likely to lobby for or support the implementation of better distance-teaching measures or when schools implement specific measures to reach out to low-SES or low-achieving students. Such mechanisms would give rise to differences in the extent to which schools compensate the lack of in-person teaching by other school inputs in one way or the other.

pothesized aspect is that the extent to which families compensate for reduced school inputs may depend on their socioeconomic background (SES). Their child's education may enter the utility function of high-SES parents more strongly, higher education may make them better substitute teachers, and they may have weaker budget constraints. As a consequence, high-SES parents may make sure that their child spends more time learning, may increase their family inputs more strongly, and may be in a better position (either financially or in terms of managing the curricular content) to support their child's learning activities. Formally, provided family inputs may depend on provided school inputs, and high-SES families (h) may react more strongly (in absolute terms) to a decline in school inputs than low-SES families (l):

$$\left|\frac{\partial F_i}{\partial S_i}\right|^h > \left|\frac{\partial F_i}{\partial S_i}\right|^l \tag{3.2}$$

As high-SES parents compensate more of the lost school inputs than low-SES parents, inequality in educational output will increase in the SES dimension.

Here, we emphasize another dimension of inequality, the one between students of different initial achievement. The sharp decline in teacher inputs that defines school closures implies the necessity of self-regulated learning. Outside the school context, students must acquire and understand the academic content more independently without the support of trained educators. Given dynamic complementarities in the skill formation process (e.g., Cunha et al., 2006; Cunha and Heckman, 2007; Cunha et al., 2010), the effectiveness of self-regulated learning will depend on individual students' ability and prior achievement. As a consequence, the presence or absence of school inputs, in particular teachers, will affect the production elasticities of students' own prior achievement. The easiest way to conceptualize this aspect is to depict the extent to which students with different levels of initial achievement A can add to their learning as a negative function of the extent of school inputs:

$$\frac{\partial Y_i}{\partial A_i} = g(S_i) \tag{3.3}$$

That is, the extent to which high-achieving students acquire larger learning gains compared to low-achieving students will be larger in home schooling than in classroom teaching because high-achieving students have a better skill base for self-regulated learning. As a consequence, school closures are expected to widen educational inequality along the achievement dimension.

To the extent that family SES and students' initial achievement are correlated, the two described mechanisms will exacerbate each other: Socioeconomic differences in family inputs may be one driver for the learning differences between low- and high-achieving students, and differences in initial achievement may be one driver for learning differences between children from low- and high-SES backgrounds.

In our empirical application, we proxy for students' educational outcomes by the amount of learning time as captured in a time-use survey. For the very reasons discussed, one may expect children from higher-SES families and higher-achieving students to acquire more skills per hour of learning at home than their counterparts. In this case, the true effects of school closures on the inequality in students' skill acquisition along these two dimensions are likely underestimated by any estimated effects on learning time. The same is true when disadvantaged children are more likely to substitute the reduced learning time by other activities that are otherwise detrimental rather than conducive to child development.

3.2.2 Institutional Background

Germany reported its first official COVID-19 case in late January 2020. As infection numbers continued to grow over the following weeks, federal and local governments adopted a broad range of measures to slow down the spread of the virus, such as social-distancing requirements, contact limitations, quarantine after traveling, and closures of shops and restaurants. A first district with a local spike in infections closed its schools on February 28.[4]

On March 13, 2020, the 16 federal states closed all educational institutions throughout Germany (Anger et al., 2020). Only young children (up to age 12) of parents who both work in so-called system-relevant occupations (e.g., health, public safety, public transportation, and groceries) were exempt and could attend emergency services in schools (*Notbetreuung*). The implementation of emergency services varied across the federal states. In April, the first states began relaxing the requirements for emergency-service attendance, e.g., by expanding the list of system-relevant occupations, including families in which only one parent worked in such an occupation, as well as children of single parents. Children admitted to emergency services were usually not taught regularly, but only supervised.

There was no standardized concept to implement distance teaching during the closures. The state ministers of education also did not formulate specific rules on which subjects should be prioritized during school closures. Instead, decisions regarding the organization of distance-teaching activities were left to the discretion of schools and teachers. Regardless of their specific subjects, all teachers were generally expected to engage in distance teaching. While many schools formally implemented certain distance-teaching activities, in practice teachers' activities were limited and left many students uninstructed (Anger et al., 2020).[5] Distance-teaching activities were further undermined by the lack of technical equipment in the schools and at students' homes.[6]

[4] This section provides an overview of German school policies during the COVID-19 pandemic between March and June 2020. See Appendix for some general facts about the German school system.

[5] A survey of teachers found that instruction was mostly limited to sending out assignments sheets: Less than half of teachers surveyed provided students with explainer videos, and online instruction via video was provided by fewer than one in five teachers (Robert Bosch Stiftung, 2020).

[6] Technical problems in distance teaching are not surprising in the German context: According to the European Commission (2019), the share of highly digitally equipped schools in Germany is substantially lower than the EU

With regard to student assessments, the states jointly decided that school exit exams should take place despite the pandemic. Most states postponed examinations for high-school diplomas (*Abitur*) from March to April or May. Unlike final exams, standardized student assessments scheduled for 2020 have been canceled because of the pandemic. Thus, no data are available so far to assess the impact of school closures on students' standardized test scores in Germany.[7]

In late April 2020, education ministers decided to gradually re-open schools, with starting dates and procedures differing across states. Accompanied by political controversies given the continued risk of COVID-19 outbreaks, schools initially re-opened only for graduation classes, and with strict hygiene rules such as compulsory mouth-nose masks and social distancing.[8]

Partial school operations—usually with alternating halves of students per classroom in daily or weekly shifts—were successively expanded to other grade levels during May and June (see Appendix Table A3.1 for the timing of school re-openings by state and class type). Ultimately, most students had at least a few weeks of in-person teaching before the summer break. Many students lost up to twelve weeks of in-person classroom teaching as a result of the school closures, equivalent to one third of a school year (Woessmann, 2020). Unfortunately, the education ministries do not provide more specific information about the exact number of weeks during which in-person classes were canceled during the school closures in spring 2020.

After the summer break in August/September 2020, schools opened for all students. However, there were no universal guidelines yet on how to continue school operations through distance teaching in the event of future infection hikes. To the best of our knowledge, we provide the first encompassing quantitative assessment of distance-teaching activities during the school closures in Germany.

---

average (e.g., 9 percent versus 35 percent at ISCED-level 1 institutions; 48 percent versus 72 percent at ISCED-level 3 institutions). In addition, the teacher survey by Huber and Helm (2020) shows that 56 percent disagree with the statement that the technical capacity at their school is sufficient for web-based formats.

[7] For details, see `https://www.kmk.org/presse/pressearchiv/mitteilung/detail/News/kmk-pruefungen-finden-wie-geplant-statt.html` [accessed June 2, 2021]. Student achievement tests that were scheduled for 2020 but had to be canceled include the IQB Bildungstrend, VERA 3, and VERA 8 for grades three, four, and eight.

[8] Teachers in particular were skeptical about the re-opening of schools. For example, when the federal state of Hesse announced it would return to normal school operations in all primary schools starting June 22, the teachers' union Gewerkschaft Erziehung und Wissenschaft (GEW) called this decision 'unreasonable' (see `https://www.gew-hessen.de/bildung/schule-fachgrupen/grundschulen/details?tx_news_pi1%5Baction%5D=detail&tx_news_pi1%5Bcontroler%5D=News&tx_news_pi1%5Bnews%5D=1884&cHash=74a0cf8544c8e797dd5604f315787907` [accessed June 16, 2021]). Similarly, the German Teachers' Association repeatedly warned against opening schools too quickly (see `https://www.lehrerverband.de/warnung-schuloeffnungen` [accessed June 16, 2021]).

3.3 Research Design and Data Collection

Using a survey of parents (section 3.3.1), we elicit time-use data on a broad range of students' activities for the periods both before and during the COVID-19-related school closures (section 3.3.2), complemented by information on parents' and schools' support activities.

3.3.1 The Survey

Our survey of parents of school children was fielded as part of the ifo Education Survey 2020, which provides a representative sample of the German population aged 18 to 69 years. Carried out between June 3 and July 1, 2020, by the survey company *respondi* via online access panels, the total sample consisted of 10,338 respondents. From the total sample, we asked all parents of school-aged children (N=1,099) to answer a series of questions on their youngest school-aged child before and during the COVID-19-related school closures.[9] As such, the subsample is a convenience sample of parents with students in all types of primary and secondary schools. However, due to the representativeness of the overall sample, it should provide a very good fit for students in Germany. In fact, comparing parental and child characteristics of our analysis sample to all school children in the representative German Microcensus[10] shows that the two samples are very similar in terms of observables (Appendix Table A3.2), raising confidence in the generalizability of results.[11]

The sociodemographic characteristics of the students and their surveyed parent (Appendix Table A3.3) indicate an average student age in the sample of 12.5 years and a rather even gender split. The sample is roughly evenly distributed between students in primary (grades 1-4), upper-track secondary (*Gymnasium*), and other types of secondary school. Responding parents are also roughly evenly split by gender, and 27 percent hold a university degree.

To categorize students as low- or high-achievers, we asked parents about their child's school grades in mathematics and German.[12] According to their parents, 15.7 percent and 12.1 percent of students in our sample have grade 1 (best grade) in mathematics and German, respectively, 34.6 and 41.3 percent grade 2, 26.4 and 28.9 percent grade 3, 10.4 and 6.2 percent

[9] The parent questions were quite detailed and therefore mentally taxing and time consuming. To minimize the risk that survey fatigue undermines data quality, parents with more than one child were only asked about their youngest school-aged child. Studying the youngest child helps to focus on the challenges of self-regulated learning (which are arguably greater for younger children) and on those whose returns to educational investments tend to be highest (e.g., Cunha et al., 2006).

[10] Research Data Centres of the Federal Statistical Office and the statistical offices of the Laender, Microcensus, census year 2015.

[11] Cases where parents reported that the child had zero hours of schooling on a typical weekday before Corona were excluded from the analysis sample as they cannot be identified as students.

[12] The question was worded as follows: 'What grades does your youngest child receive in the main subjects (mathematics and German) most frequently?' Respondents reported a separate grade for mathematics and German on the German grade scale (from 1='very good' to 6='failed').

grade 4, and 2.3 and 0.6 percent grade 5.[13] Computing the median of the average grade in the two subjects separately for the three school types, we classify students at or above this median as high-achievers (55.5 percent) and those below the median as low-achievers (44.5 percent).[14] Thus, our achievement measure captures children's previous educational performance relative to other children in the same school type.

A regression of a high-achiever indicator on sociodemographic characteristics (column 2 of Appendix Table A3.3) indicates few significant observable differences between low- and high-achieving students, with the exceptions that high-achievers are more likely to come from high-income households, have the parent working in home office during Corona, and be younger. Child gender, family status, and parent's work hours do not significantly predict better student grades. We control for these background variables in our regression analysis.[15]

3.3.2 Elicitation of Time-Use Information Before and During COVID-19

The core of our analysis is detailed time-use data on students' activities for the period of the COVID-19-related school closures. To be able to investigate whether any differences between low- and high-achieving students already existed before the closures or whether they emerged with the closures, we also elicited the same time-use battery retrospectively for the time before the school closures.

Inspired by the time-use module in the mother-child questionnaire of the German Socio-Economic Panel Study (Schröder et al., 2013), we carefully designed the time-use battery to capture relevant activities that students engaged in before and during the school closures. Parents had to specify how many hours (rounded to the nearest half hour) their child spent during a typical workday on each of the following activities:[16] (i) School attendance; (ii)

[13] Reassuringly, the grade distribution in our sample is similar to the distribution in the youth questionnaire of the 2018 wave of the German Socio-Economic Panel Study (GSOEP). Detailed results are available upon request.

[14] Because of the rather coarse grading in primary school (33 percent of students have the median average grade of 2.0), a relatively large fraction of primary-school students (64 percent) falls into the category of at-or-above median grades, compared to 51 and 53 percent of upper-track and other secondary-school students, respectively. 116 students (10.6 percent) had to be excluded from this sub-group analysis because they do not receive numerical grades. Most of them (106) are in primary school, where children usually do not receive numerical grades in the early grade levels. In bounding analyses, we assigned children with missing grade information hypothetical achievement levels â€“ either low or high achieving. Reassuringly, our main finding that the school closures increased the learning-time gap by student achievement turns out robust in this attrition analysis (detailed results available upon request).

[15] The small number of observable differences likely reflects that the analysis neglects any variation between school types and that it is based on a multivariate model that holds the other variables constant. In fact, regressing the high-achievement dummy on each characteristic separately (accounting only for school-type dummies) yields the following significant coefficients ($p<0.05$) in addition to the ones in column 2 of Appendix Table A3.3: parental university degree (positive), child not in household (negative), parental work hours (positive), and household income (positive). Detailed results are available upon request.

[16] Question wording: *'The following questions are about your youngest child attending school. What activities did your child do on a typical workday (Monday to Friday) before [during] the several weeks of Corona-related school*

Learning for school; (iii) Reading or being read to; (iv) Playing music and creative work; (v) Physical exercise; (vi) Watching TV; (vii) Gaming on computer or smartphone; (viii) Social media; (iv) Online media; and (x) Time-out (e.g., relaxing). We also provided an open field to specify 'Another activity.'[17] To be able to study whether and how parents adapted their home-schooling activities vis-à-vis the school closures, we also elicited how much time parents spent together with their child on the respective activities.

For our analysis, we group the activities into three categories: *school*-related activities (activities (i) and (ii)), other activities generally deemed *conducive* to child development (activities (iii)-(v)), and activities generally deemed *detrimental* (activities (vi)-(ix)). Our categorization is reflected in parents' beliefs about how beneficial each activity is for their child's development, which we elicited after the time-use batteries. Almost all parents consider the two *school*-related activities (97 and 93 percent) and the *conducive* activities (82-95 percent) beneficial (Appendix Table A3.4). In contrast, only 22-34 percent think that the different *detrimental* activities are beneficial. Importantly, these assessments do not differ substantially between parents of low- and high-achieving students, implying that any difference in time use cannot be assigned to different beliefs about the activities' developmental effects.

Complementing our time-use data, we also elicited parents' assessment of how the school closures affected their family and learning environment at home, as well as information on the distance-teaching activities undertaken by schools. The five questionnaire items on the home environment capture topics such as how the family coped with the situation, whether it was a psychological burden for the child and the parents, and an overall assessment of the child's home learning environment (see notes to Appendix Table A3.7 for question wordings). Schools' distance-teaching activities during school closures were elicited by seven questionnaire items on activities such as shared remote lessons, individual teacher contacts, use of educational videos or software, and providing work sheets (see notes to Table 3.4 for question wordings).

The survey-based, partially retrospective elicitation of information about children from their parents raises issues of validity and interpretation that we will discuss in section 3.7 below. There, we also discuss evidence that several patterns in our data are consistent with alternative data sources, which raises confidence in the validity of our main findings.

closures?' The sum of reported hours spent per day was prevented from exceeding 24 hours. In our analysis, outliers in any answer category are top-coded at 12 hours.

[17] In cases where the activity specified in the open field corresponded to existing categories, we re-coded the respective category accordingly.

3.4 Time Use of Low- and High-Achieving Students Before and During the School Closures

This section reports results on how the COVID-19 school closures differentially affected low- and high-achieving students' learning time (section 3.4.1), as well as their time investment in other conducive and detrimental activities (section 3.4.2).

3.4.1 Learning Time

To be able to investigate how the gap in learning time between low- and high-achieving students changed over time, we elicited information on time use for school-related activities on a typical workday both before and during the school closures. The school-related activities include the two sub-categories of attending school and learning for school at home.

In the full sample, the school closures more than halved students' learning time. Before the school closures, students spent on average 7.4 hours per day on school-related activities (Appendix Table A3.5). This number dropped to 3.6 hours during the closures. This reduction is due to a large decline in school attendance—from an average of 5.9 to 0.9 hours (emergency services) per day—that is hardly compensated by a much smaller increase in time spent on learning for school (from 1.5 to 2.7 hours).

Differentiating between low- and high-achieving students reveals that the school closures strongly increased educational inequality. Columns 5–8 of Table 3.1 indicate that learning time before the school closures did not differ economically or statistically significantly between students initially achieving below versus at-or-above the median (7.4 versus 7.5 hours per day).[18] By contrast, columns 1-4 show that high-achieving students spent 0.5 hours more on school-related activities during the closures (3.4 versus 3.9 hours, $p<0.01$).[19] Consequently, the increase in the learning-time gap between low- and high-achieving students relative to pre-closure times (columns 9–12) is a significant 0.4 hours per day (-4.1 versus -3.7 hours for low- and high-achievers, respectively; see also Appendix Figure A3.1). Beyond the binary achievement indicator of our baseline analysis, Appendix Figure A3.2 shows that the relationship between the reduction in learning time and student achievement is visible across the entire grade spectrum. For instance, learning time decreases by 3.6 hours in the top and 4.2 hours in the bottom of the five grade categories. Distinguishing between the two sub-categories of school-related activities, the decrease in school attendance was similar for

[18] Throughout, average results for the full sample are not a simple weighted average of high- and low achieving students because they include students who do not yet receive grades.

[19] The difference in learning time between low- and high-achieving students during the school closures is visible throughout the entire distribution (Appendix Table A3.6). For example, 43 percent of low-achievers spent at most two hours per day on school-related activities, compared to 33 percent of high-achievers. Only 22 versus 30 percent, respectively, spent more than four hours per day on learning. For comparison, before the school closures 89 percent of students spent at least five hours per day on learning.

low- and high-achievers (-5.1 versus -5.0 hours), but low-achievers increased home learning less than high-achievers (+1.0 versus +1.4 hours).

Going beyond mean differences between low- and high-achieving students, Figure 3.1 depicts the respective distributions of learning-time losses for the two groups. The distribution of low-achievers is consistently shifted to the left (towards greater learning-time losses) compared to high-achievers. A two-sample Kolmogorov-Smirnov test rejects the null hypothesis that learning-time losses do not differ by student achievement ($p=0.014$). Thus, average differences in learning-time losses as reported in Table 3.1 are not driven by extreme outliers but are rather observable throughout the distribution

The learning-time gap between low- and high-achieving students can hardly be accounted for by other observed student and parent characteristics. Table 3.2 shows results of regressions of the learning time during the school closures on a high-achiever dummy, learning time before the school closures, and a series of student and parent characteristics: the student's school type, age, gender, a single-child dummy, the responding parent's gender, education, single-parent status, home-office status and work hours during the school closures, partner at home during the school closures, household income, and a West-Germany dummy. In all cases, including the additional variables leaves the difference between high- and low-achieving students highly significant and of similar magnitude as the unconditional gap.[20] Including all controls simultaneously (column 14) reduces the difference in learning time between high- and low-achieving students by less than one fifth. Thus, most of the large gap does not reflect differences in the observed characteristics, but rather seems to capture the genuine achievement dimension.

3.4.2 Other Conducive and Detrimental Activities

Substituting the reduced learning time, both low- and high-achieving students only mildly increased the time spent on other activities that are generally viewed as conducive for child development. During the school closures, high-achievers (3.4 hours) spent significantly more time on reading, playing music, creative work, or physical exercise than low-achievers (2.8 hours; see middle panel of Table 3.1). However, most of this gap existed already before the closures, so that the difference in the increase in these conducive activities is only marginally significant (+0.2 versus +0.4 hours for low- and high-achievers, respectively, $p<0.1$).

By contrast, low-achieving students particularly used the released time to expand activities such as gaming on the computer or consuming social media. During the school closures, low-achieving students spent 6.3 hours on activities such as watching TV, playing computer games, and consuming social and online media that are generally deemed detrimental to child

[20] In fact, the only noteworthy reduction does not come from any of the measures of socioeconomic background or family situation, but rather from student age (column 3), reflecting that younger students tend to get better grades and had a smaller reduction in learning time (due to lower before-Corona levels).

development (bottom panel of Table 3.1)—nearly three hours more each day than on school-related activities. In comparison, high-achievers spent 1.5 hours less on the detrimental activities. Roughly half of this gap already existed before the school closures, so that the increase in time spent on detrimental activities was 0.7 hours larger for low- compared to high-achieving students (+1.7 versus +1.0 hours). The increase is mostly driven by increased gaps in computer gaming and social-media use, each of which increased by 0.3 hours.

Together, the results indicate that the school closures exacerbated educational inequality along the achievement dimension. The findings suggest that COVID-19 (i) increased the gap in learning time (and, mildly, in other conducive activities) between high- and low achieving students and (ii) increased detrimental activities especially among low-achieving students. Since low-achieving students are, basically by definition, less effective in turning learning-time inputs into knowledge and skills, we interpret the pronounced effect of the school closures on students' learning-time gaps as lower bound for the impact on gaps in actual learning.[21]

3.5 Compensating Activities by Parents and Schools

This section investigates to what extent parents (section 3.5.1) and schools (section 3.5.2) acted to compensate for the increased gap in learning time between low- and high-achieving students.

3.5.1 Parental Support

While parents of both low- and high-achieving students increased the time they spent together with their child on learning during the school closures, both level and increase were smaller for low-achievers.[22] During the school closures, low-achievers spent 0.3 hours per day less learning together with their parents than high-achievers (0.9 versus 1.2 hours, $p<0.01$; Table 3.3). While part of this gap already existed before the closures, it further increased by 0.1 hours during the school closures ($p<0.1$). Thus, even though parents increased the learning involvement with their children by half an hour per day during the closures, this aggravated rather than compensated for the increase in educational inequality.

By contrast, the increase in time spent together with parents on other conducive and on detrimental activities did not differ statistically significantly between low- and high-achievers.

[21] Consistently, parents of low-achievers are 14 percentage points more likely than parents of high-achievers to report that their child learned 'much less' during the school closures than usual (Appendix Table A3.7).

[22] The importance of parental inputs for children's skill development is underscored by the finding that children's educational activities are particularly productive when parents are involved (Fiorini and Keane, 2014).

Still, parents of high-achieving students also spent significantly more time with their child on other conducive activities both before and during the school closures.[23]

Parents' assessment of the environment at home reinforces the finding that low-achieving students were more affected by the COVID-19 school closures. While most parents (87 percent) think that their family has coped well with the period of school closures (Appendix Table A3.7), parents of low-achieving students evaluate the situation slightly worse than parents of high-achieving students (85 versus 90 percent, $p<0.05$). There is no significant difference between low- and high-achieving students in whether parents report that the phase of the school closures was a psychological burden for the child or for themselves (38 percent each on average). By contrast, parents of low-achievers are slightly more likely than parents of high-achievers to report that during the school closures, they argued more than usual with their child (30 versus 24 percent, $p<0.1$). They also assess the overall learning environment at home (e.g., in terms of available computers or working space) worse. These gaps hardly change when conditioning on observable child and parent characteristics (column 6).

3.5.2 School Support

During the closures, schools and teachers carried out only a fraction of their usual teaching operations via distance teaching, which led to a drastic reduction in direct communication between teachers and students. Table 3.4 indicates that only 29 percent of students on average had online lessons for the whole class (e.g., by video call) more than once a week. Only 17 percent of students had individual contact with their teacher more than once a week.[24] The main teaching mode during the school closures was to provide students with exercise sheets for independent processing (87 percent),[25] although only 37 percent received feedback on the completed exercises more than once a week. School activities strongly correlate with children's learning time during the school closures: Children in schools with above-median intensity of distance teaching (with respect to online lessons, individual teacher-student contacts, and feedback on exercises) spent a significant 0.4 hours more time on learning for school a day (2.92 hours versus 2.55 hours).

The distance-teaching measures over-proportionally reached high-achieving students. Low-achievers were 13 percentage points less likely than high-achievers to be taught in online lessons and 10 percentage points less likely to have individual contact with their teachers (column 4). Low-achievers were also less likely to be provided with educational videos or software and to receive feedback on their completed tasks. These gaps do not change noticeably

[23] In additional analyses, we find that parent involvement in learning and other conducive activities before and during the school closures decreases with child age, as does the increase in parental involvement in these activities induced by the school closures (detailed results available upon request).

[24] Across the five answer categories, 6 (4) percent had joint online lessons (individual teacher contact) on a daily basis, 23 (14) percent several times a week, 14 (16) percent once a week, 11 (22) percent less than once a week, and 45 (45) percent never.

[25] 96 percent of students received exercises at least once a week.

when conditioning on child and parental characteristics (column 6). Thus, schools were not able to compensate for the adverse effects of the closures on educational inequality. To the contrary, those students more in need of additional support to keep up learning during the school closures were less likely to benefit from distance-teaching activities.[26]

3.6 Other Dimensions of Inequality

This section investigates whether the school closures also amplified educational inequality along other dimensions than students' prior achievement, namely parents' educational background (section 3.6.1) and students' gender and school type (section 3.6.2).

3.6.1 Differences by Parents' Educational Background

In the public debate, there is concern that the COVID-19-induced school closures could aggravate educational inequality between children from different socioeconomic backgrounds (e.g., European Commission, 2020; UNESCO, 2020a). Family background has been shown to strongly impact students' educational success (e.g., Björklund and Salvanes, 2011).

While children of university-educated parents invested more time in out-of-school learning activities before COVID-19 than children of parents without a university degree, the reduction in learning time during the school closures did not differ significantly between children of parents with (-3.7 hours per day) or without (-3.8 hours) a university degree (upper panel of Table 3.5).[27] While children of university-educated parents spent marginally significantly more time on school-related activities during the closures (3.8 versus 3.6 hours), most of this gap already existed before COVID-19.[28] Children of university-educated parents did increase their time on other conducive activities more. They also spent less time on detrimental activities both before and during the closures, but the change over time was not significantly different from children of parents without a university degree.

At the same time, there are strong differences in school support during the closures by family background. For instance, children without university-educated parents were 12 percentage points less likely than children with university-educated parents to be taught in online lessons more than once a week, and 15 percentage points less likely to have individual contact with their teachers more than once a week (not shown). This pattern raises concerns that the

[26] Consistently, the share of parents reporting to be satisfied with their school's activities during the school closures was 13 percentage points lower for low- than for high-achieving students (Appendix Table A3.7).

[27] Consistently, learning time during the school closures also did not differ between students with above and below median household income. Due to longer school attendance before the closures, the decline was actually larger for students from high-income households (results available upon request).

[28] We find the same qualitative pattern of results when using a more fine-grained categorization of parental education (no degree, vocational degree, advanced vocational degree (e.g., *Meister*), and university degree). Detailed results are available upon request.

school closures might have exacerbated inequality in student achievement by children's socioeconomic background, even though the learning-time gap did not widen.

3.6.2 Differences by Students' Gender and School Type

Analysis by student gender indicates that the school closures reduced boys' learning time more than girls'. Before the closures, there was no significant gender difference in learning time (lower panel of Table 3.5). By contrast, boys spent half an hour less than girls learning at home during the school closures (3.4 versus 3.9 hours, $p<0.01$). Boys substituted learning time mostly for playing computer games, whereas girls mostly increased their time on social media, reinforcing gender differences in both dimensions. The overall gender effect of the closures may exacerbate the 'boy crisis' in education (e.g., Cappelen et al., 2019).

There are also noteworthy differences between students in primary, upper-track secondary (*Gymnasium*), and other secondary school. During Corona, primary-school students were more likely to attend emergency services in schools, which were open only to younger children (Appendix Table A3.8). Upper-track secondary-school students spent more time learning at home (3.2 hours) than their lower-track and primary-school counterparts (2.5 hours each). Still, in absolute terms, both types of secondary-school students lost learning time to a similar extent. Primary-school students expanded other conducive activities—in particular, physical exercise—more than secondary-school students, who mostly expanded gaming and social media.

3.7 Discussion

The detailed time-use survey data provide novel and otherwise unavailable information on students' learning during the COVID-19-induced school closures. Still, several points should be kept in mind in interpreting the findings. First, students' time spent on learning and other activities are imperfect proxies for how much they actually learn (e.g., Hanushek and Woessmann, 2008). Arguably, high-achieving students are more effective in turning learning time into knowledge and skills. In this case, our results likely constitute a lower bound for the impact of school closures on skill inequality by student's prior achievement.[29]

Second, survey responses could be subject to social-desirability bias. For instance, parents may inflate reported learning time because they think it is considered socially appropriate. However, research shows that social desirability does not yield major bias in anonymous

[29] In addition, an interesting interpretative question that remains unanswered from our analysis is what exact subjects were taught and at what intensity during the school closures. While some evidence speaks against a strong shift in teaching emphasis to core subjects such as mathematics or German (e.g., because teachers of all subjects were expected to engage in distance-teaching activities and because the majority of parents thinks their child learned 'much less' than usual during the school closures), an in-depth analysis of distance-teaching curricula would be interesting for future research.

online surveys as ours (e.g., Das and Laumann, 2010). In fact, parents reported that during the closures, their child spent much more time on detrimental activities such as watching TV or computer gaming than on learning. This pattern is inconsistent with a major influence of social-desirability bias on answering behavior. Furthermore, any remaining bias would imply that the large discrepancy between school-related and detrimental activities found in our data even underestimates the true difference.

Third, our analyses are partly based on retrospective reports on how much time children spent on different activities before the school closures. While we cannot rule out that selective memory leads to measurement error in the data (e.g., Zimmermann, 2020), it is reassuring that the retrospective answers are plausible in the sense that reported hours spent in school before the closures correspond closely to the hours prescribed in the school curricula. Furthermore, our retrospective data closely resemble students' self-reported learning time elicited in the 2018 wave of the German Socio-Economic Panel Study (GSOEP), which further raises confidence in the validity of our retrospective time-use data.[30]

Fourth, the survey data could suffer from measurement error because parents do not know exactly how much time their child spends on different activities. However, only 21 percent of respondents state that both they and their partner worked at least half a day outside the home during the school closures. The relatively intense parent-child contact in most households increases parents' ability to monitor their child's activities, so that most parents should be able to assess these activities reasonably well. Reassuringly, a survey of students in the final two grades of upper-track secondary school in eight German states by Anger et al. (2020) also finds that learning time during the school closures differs markedly by students' previous school grades, but not by parental educational background. This indicates that our results are unlikely driven by measurement error from lacking knowledge of parents in our data.

Fifth, survey fatigue can lead to respondents not answering some questions conscientiously. However, 500 of the 1,099 parents in our sample used the provided open answer field to type in 'another activity' in the time-use battery, which indicates that they were very conscientious in filling out the survey.

Finally, the extent to which our results for Germany are informative for other contexts is ultimately an empirical question that we cannot answer with our data. On the one hand, most countries were at least as affected by the COVID-19 pandemic as Germany, had broadly similar school-closure policies, had no previous experience with nation-wide school closures, and had no concepts in place for online school operations. Reports from many countries

[30] The GSOEP asks 12- to 15-year-olds: *'How much time do you usually spend on homework and studying for school?'* Answer categories are less than half an hour a day, half an hour to less than 1 hour a day, 1 to less than 2 hours a day, 2 to less than 3 hours a day, 3 to less than 4 hours a day, and 4 hours and more a day. The average answer is 1.1 hours of daily learning for school, compared to 1.5 hours that parents of children in the same age range report in our sample. Importantly, the GSOEP data reveals no difference in learning time between low- and high-achieving students (using our grade-based classification), which is also in line with our results.

indicate that the organization of distance-teaching activities was challenging and caused major problems not only in Germany (e.g., Andrew et al., 2020; Chetty et al., 2020; Engzell et al., 2020; Maldonado and De Witte, 2020). On the other hand, there is some indication that Germany lagged other countries in the classroom usage of digital technologies before the pandemic (e.g., Fraillon et al., 2018; Beblavý et al., 2019), raising the possibility that some other countries may have fared better in providing online teaching for their students and particularly support the low-achievers.

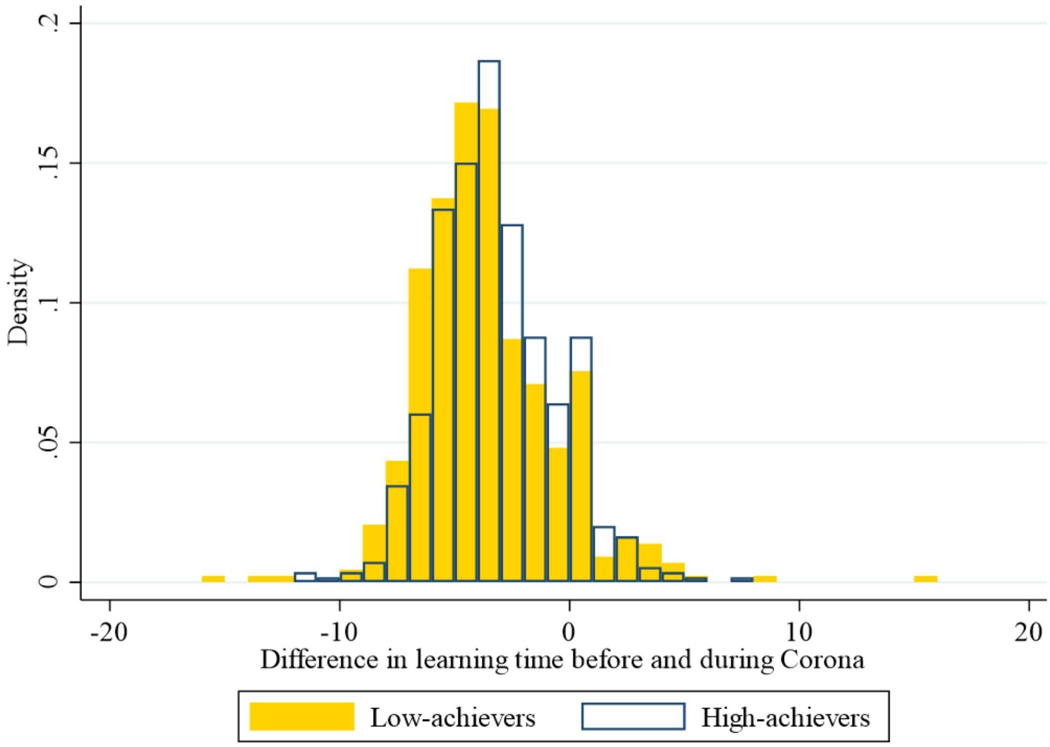
3.8 Conclusion

We present novel time-use data on the activities of more than 1,000 school children before and during the COVID-19 school closures in Germany. On average, the school closures reduced students' learning time by about half. This reduction was significantly larger for low-achieving than for high-achieving students. Especially low-achieving students substituted the learning time for detrimental activities such as watching TV and playing computer games, rather than for conducive activities. Neither parents nor schools compensated for the increased learning gap by students' prior achievement and actually provided less support for low- than for high-achieving students. The reduction in students' learning time did not vary by parents' educational background (though children without university-educated parents received less school support during the closures), but it was larger for boys than for girls.

From a policy perspective, our results call for universal and binding distance-teaching concepts for school closures that are particularly geared towards low-achieving students. Leaving the decision over whether and how to maintain teaching operations during school closures at schools' or teachers' discretion has proven largely unsuccessful in our setting. In fact, proposals to instruct teachers to maintain daily contact with their students, require all schools to switch to online teaching if in-person classes are not possible, and enable online teaching by compulsory teacher training and providing digital equipment to students who cannot afford them have overwhelming majority appeal in the German electorate (Woessmann et al., 2020). Our results suggest that it is particularly the low-achieving students who suffer when support of teachers is lacking, so that any attempt to support their learning when schools have to close is likely to reduce future educational inequality.

Figures and Tables

Figure 3.1: Distribution of Reduction in Learning Time by Student Achievement



Notes: Difference in average hours spent on school activities on a typical workday between the period before the school closures and the period of school closures due to COVID-19. Low- versus high-achievers: students with an average grade in mathematics and German below versus at-or-above the median for their respective school type. A two-sample Kolmogorov-Smirnov test rejects equality of the two depicted distributions with a p-value of 0.014. Data source: ifo Education Survey 2020.

Table 3.1: Activities of Low- and High-Achieving Students Before and During the School Closures

	During Corona				Before Corona				Difference During-Before			
	Low-Achiev. (1)	High-Achiev. (2)	Gap (3)	Std. Err. (4)	Low-Achiev. (5)	High-Achiev. (6)	Gap (7)	Std. Err. (8)	Low-Achiev. (9)	High-Achiev. (10)	Gap (11)	Std. Err. (12)
School Activities												
Aggregate	3.36	3.85	0.496	(0.151)***	7.42	7.50	0.079	(0.130)	-4.07	-3.65	0.416	(0.180)**
Attending School	0.82	0.92	0.103	(0.133)	5.93	5.93	-0.003	(0.116)	-5.11	-5.01	0.105	(0.177)
Learning for School	2.54	2.93	0.393	(0.102)***	1.49	1.58	0.082	(0.067)	1.04	1.35	0.311	(0.108)***
Conducive Activities												
Aggregate	2.79	3.37	0.58	(0.128)***	2.61	3.01	0.403	(0.107)***	0.19	0.36	0.177	(0.107)*
Reading	0.63	0.86	0.237	(0.046)***	0.54	0.74	0.201	(0.039)***	0.09	0.12	0.036	(0.041)
Music and Creative Work	0.66	0.82	0.164	(0.061)***	0.53	0.65	0.117	(0.046)**	0.13	0.17	0.047	(0.047)
Physical Exercise	1.51	1.69	0.179	(0.080)**	1.53	1.62	0.085	(0.067)	-0.03	0.07	0.094	(0.077)
Detrimental Activities												
Aggregate	6.29	4.84	-1.452	(0.210)***	4.58	3.82	-0.762	(0.156)***	1.71	1.02	-0.691	(0.146)***
Watching TV	1.50	1.37	-0.126	(0.070)*	1.24	1.18	-0.059	(0.058)	0.26	0.20	-0.067	(0.051)
Gaming	1.87	1.32	-0.55	(0.101)***	1.23	0.99	-0.244	(0.068)***	0.64	0.34	-0.306	(0.068)***
Social Media	1.77	1.18	-0.593	(0.097)***	1.22	0.90	-0.321	(0.067)***	0.55	0.28	-0.272	(0.067)***
Online Media	1.15	0.97	-0.184	(0.067)***	0.89	0.76	-0.137	(0.047)***	0.26	0.21	-0.046	(0.056)

Notes: Average hours spent on different activities on a typical workday. During Corona: period of school closures due to COVID-19. Before Corona: period before the school closures. Low-versus high-achievers: students with an average grade in mathematics and German below versus at-or-above the median for their respective school type. Std. err.: standard errors stemming from regressions of hours spent on each activity on a high-achiever indicator. Significance levels: *** p<0.01, ** p<0.5, * p<0.1. Data source: ifo Education Survey 2020.

Table 3.2: Gap in Learning Time between Low- and High-Achieving Students Conditional on Student and Parent Characteristics

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
High-Achiever	0.478***	0.474***	0.417***	0.455***	0.478***	0.460***	0.463***	0.483***	0.460***	0.478***	0.475***	0.492***	0.461***	0.240**
	(0.149)	(0.149)	(0.151)	(0.148)	(0.149)	(0.149)	(0.149)	(0.149)	(0.150)	(0.149)	(0.149)	(0.149)	(0.149)	(0.038)
School Activities Before Corona	0.224***	0.218***	0.244***	0.219***	0.224***	0.229***	0.222***	0.225***	0.227***	0.224***	0.224***	0.232***	0.219***	0.368**
	(0.036)	(0.037)	(0.037)	(0.036)	(0.038)	(0.036)	(0.036)	(0.038)	(0.036)	(0.037)	(0.036)	(0.037)	(0.036)	(0.151)
Upper Sec. School (*Gymnasium*)		0.12												0.146
		(0.191)												(0.232)
Other Sec. School		-0.286												0.687***
		(0.183)												(0.254)
Age			-0.053**											-0.095***
			(0.023)											(0.032)
Girl				0.477***										0.486***
				(0.147)										(0.147)
Single Child					-0.062									-0.142
					(0.152)									(0.153)
Parent Female						-0.286*								-0.299*
						(0.148)								(0.161)
Parent has Univ. Degree							0.185							0.113
							(0.167)							(0.188)
Single Parent								-0.079						-0.006
								(0.205)						(0.222)
Child not in Household								0.146						0.043
								(0.271)						(0.281)
Parent in Home Office									0.183					0.170
									(0.157)					(0.174)
Parent Work Hours										0.000				-0.001
										(0.004)				(0.005)
Partner at Home											0.172			0.172
											(0.189)			(0.196)
Household Income												-0.001*		-0.012**
												(0.000)		(0.005)
West Germany													-0.399**	-0.390**
													(0.184)	(0.185)
Constant	1.692	1.805	2.260	1.504	1.719	1.809	1.665	1.683	1.616	1.692	1.662	1.881	2.058	3.132
Observations	983	983	983	982	983	983	983	983	982	983	983	980	983	978
R-squared	0.048	0.053	0.053	0.057	0.048	0.051	0.049	0.048	0.050	0.048	0.048	0.050	0.052	0.090

Notes: Dependent variable: average hours spent on 'attending school' and 'learning for school' on a typical workday during the period of school closures due to COVID-19. Before Corona: period before the school closures. Low- versus high-achievers: students with an average grade in mathematics and German below versus at-or-above the median for their respective school type. Significance levels: *** p<0.01, ** p<0.5, * p<0.1. Data source: ifo Education Survey 2020.

Table 3.3: Parental Involvement in Activities of Low- and High-Achieving Students

	During Corona				Before Corona				Difference During-Before			
	Low-Achiev. (1)	High-Achiev. (2)	Gap (3)	Std. Err. (4)	Low-Achiev. (5)	High-Achiev. (6)	Gap (7)	Std. Err. (8)	Low-Achiev. (9)	High-Achiev. (10)	Gap (11)	Std. Err. (12)
School Activities												
Aggregate	0.89	1.20	0.311	(0.079)***	0.42	0.62	0.193	(0.044)***	0.47	0.59	0.118	(0.069)*
Conducive Activities												
Aggregate	1.07	1.47	0.398	(0.099)***	0.78	1.11	0.325	(0.087)***	0.29	0.36	0.073	(0.077)
Reading	0.22	0.34	0.121	(0.033)***	0.18	0.30	0.124	(0.030)***	0.04	0.04	-0.002	(0.026)
Music and Creative Work	0.20	0.28	0.086	(0.033)***	0.17	0.23	0.060	(0.028)**	0.03	0.06	0.026	(0.030)
Physical Exercise	0.66	0.85	0.191	(0.063)***	0.44	0.58	0.142	(0.050)***	0.22	0.27	0.050	(0.057)
Detrimental Activities												
Aggregate	1.36	1.45	0.094	(0.132)	1.03	1.23	0.200	(0.109)*	0.32	0.22	-0.106	(0.090)
Watching TV	0.68	0.73	0.047	(0.058)	0.52	0.62	0.101	(0.049)**	0.16	0.11	-0.053	(0.047)
Gaming	0.23	0.24	0.003	(0.044)	0.18	0.22	0.037	(0.035)	0.05	0.02	-0.035	(0.033)
Social Media	0.24	0.24	-0.005	(0.053)	0.18	0.20	0.016	(0.040)	0.06	0.04	-0.021	(0.039)
Online Media	0.19	0.24	0.049	(0.034)	0.15	0.19	0.046	(0.028)	0.05	0.05	0.003	(0.030)

Notes: Average hours parents spent with their child on different activities on a typical workday. During Corona: period of school closures due to COVID-19. Before Corona: period before the school closures. Low- versus high-achievers: students with an average grade in mathematics and German below versus at-or-above the median for their respective school type. Std. err.: standard errors stemming from regressions of hours spent on each activity on a high-achiever indicator. Significance levels: *** p<0.01, ** p<0.5, * p<0.1. Data source: ifo Education Survey 2020.

Table 3.4: Schools' Distance-Teaching Activities During the School Closures for Low- and High-Achieving Students

				Unconditional Gap		Conditional Gap	
	Average	Low-Achiev.	High-Achiev.	Gap	Std. Err.	Gap	Std. Err.
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Shared Lessons (e.g., by Video Call)	0.29	0.24	0.37	0.131	(0.029)***	0.131	(0.031)***
Individual Contact with Teacher	0.17	0.13	0.23	0.102	(0.025)***	0.081	(0.026)***
Educational Videos or Texts	0.53	0.47	0.59	0.118	(0.032)***	0.115	(0.034)***
Educational Software	0.43	0.40	0.47	0.078	(0.032)**	0.068	(0.034)**
Child Received Exercises	0.87	0.84	0.89	0.049	(0.022)**	0.042	(0.023)*
Child Had to Submit Exercises	0.51	0.51	0.55	0.033	(0.032)	0.054	(0.033)
Child Received Feedback on Exercises	0.37	0.34	0.42	0.078	(0.031)**	0.096	(0.033)***

Notes: Probability that the respective activity was conducted 'daily' or 'several times a week' (residual category includes 'once a week', 'less than once a week', and 'never'). Question wording: 'Which activities did the teachers/school of your child carry out during the several weeks of Corona-related school closures? Shared lessons for the whole class (e.g., by video call or telephone); Individual contact with my child (e.g., by video call or telephone); My child should watch provided educational videos or read texts; My child should use educational software or programs; My child should work on provided exercises; My child had to submit completed exercises; Teachers gave feedback on the completed exercises.' Low- versus high-achievers: students with an average grade in mathematics and German below versus at-or-above the median for their respective school type. Std. err.: standard errors stemming from regressions of an indicator that the respective activity was conducted at least several times a week on a high-achiever indicator. Conditional gap: see Table 3.2 for controls. Significance levels: *** p<0.01, ** p<0.5, * p<0.1. Data source: ifo Education Survey 2020.

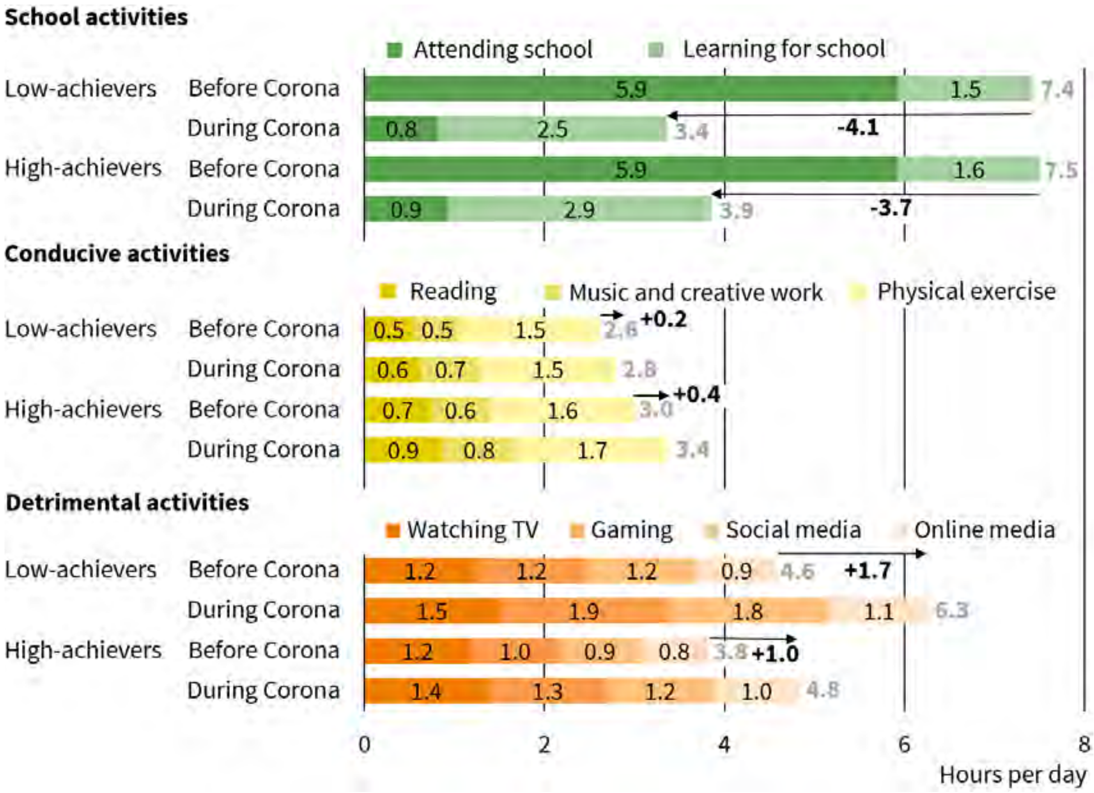
Table 3.5: Student Activities Before and During the School Closures by Parental Education and by Students' Gender

	During Corona				Before Corona				Difference During-Before			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
	Low-Ed.	High-Ed.	Gap	Std. Err.	Low-Ed.	High-Ed.	Gap	Std. Err.	Low-Ed.	High-Ed.	Gap	Std. Err.
School Activities												
Aggregate	3.55	3.82	0.275	(0.162)*	7.37	7.55	0.178	(0.136)	-3.83	-3.73	0.097	(0.189)
Attending School	0.85	1.04	0.190	(0.143)	5.91	5.92	0.013	(0.122)	-5.06	-4.88	0.177	(0.185)
Learning for School	2.70	2.78	0.085	(0.107)	1.46	1.63	0.165	(0.070)**	1.23	1.15	-0.080	(0.114)
Conducive Activities												
Aggregate	3.10	3.48	0.380	(0.138)***	2.86	2.98	0.122	(0.114)	0.24	0.50	0.258	(0.113)**
Reading	0.73	0.87	0.141	(0.048)***	0.63	0.76	0.128	(0.042)***	0.10	0.11	0.013	(0.043)
Music and Creative Work	0.70	0.95	0.249	(0.064)***	0.57	0.73	0.161	(0.048)***	0.13	0.22	0.088	(0.051)*
Physical Exercise	1.67	1.66	-0.010	(0.088)	1.66	1.50	-0.166	(0.072)**	0.01	0.16	0.156	(0.083)*
Detrimental Activities												
Aggregate	5.48	4.54	-0.934	(0.223)***	4.17	3.41	-0.759	(0.164)***	1.31	1.13	-0.175	(0.150)
Watching TV	1.48	1.25	-0.237	(0.072)***	1.26	1.04	-0.221	(0.060)***	0.23	0.21	-0.016	(0.053)
Gaming	1.55	1.33	-0.225	(0.106)**	1.10	0.91	-0.189	(0.070)***	0.46	0.42	-0.036	(0.070)
Social Media	1.42	1.01	-0.409	(0.102)***	1.04	0.72	-0.318	(0.070)***	0.38	0.29	-0.092	(0.068)
Online Media	1.02	0.96	-0.062	(0.070)	0.78	0.75	-0.031	(0.050)	0.24	0.21	-0.031	(0.057)
	Boy	Girl	Gap	Std. Err.	Boy	Girl	Gap	Std. Err.	Boy	Girl	Gap	Std. Err.
School Activities												
Aggregate	3.36	3.89	0.525	(0.143)***	7.40	7.44	0.039	(0.121)	-4.04	-3.55	0.486	(0.168)***
Attending School	0.88	0.91	0.026	(0.127)	5.91	5.92	0.016	(0.109)	-5.02	-5.01	0.010	(0.164)
Learning for School	2.48	2.98	0.499	(0.094)***	1.50	1.52	0.022	(0.063)	0.98	1.46	0.476	(0.100)***
Conducive Activities												
Aggregate	3.08	3.34	0.260	(0.123)**	2.85	2.94	0.087	(0.102)	0.23	0.40	0.173	(0.101)*
Reading	0.72	0.82	0.102	(0.043)**	0.65	0.68	0.032	(0.038)	0.07	0.14	0.071	(0.038)*
Music and Creative Work	0.65	0.90	0.253	(0.057)***	0.55	0.68	0.128	(0.043)***	0.10	0.22	0.125	(0.046)***
Physical Exercise	1.71	1.62	-0.096	(0.079)	1.65	1.58	-0.073	(0.064)	0.06	0.04	-0.023	(0.075)
Detrimental Activities												
Aggregate	5.57	4.85	-0.716	(0.199)***	4.19	3.72	-0.477	(0.147)***	1.38	1.14	-0.239	(0.134)*
Watching TV	1.41	1.43	0.013	(0.065)	1.20	1.19	-0.008	(0.054)	0.21	0.23	0.021	(0.048)
Gaming	1.97	0.98	-0.987	(0.090)***	1.34	0.73	-0.611	(0.060)***	0.63	0.25	-0.376	(0.062)***
Social Media	1.19	1.44	0.254	(0.091)***	0.87	1.03	0.162	(0.062)***	0.32	0.41	0.092	(0.061)
Online Media	1.00	1.00	0.004	(0.063)	0.78	0.76	-0.020	(0.044)	0.22	0.24	0.024	(0.051)

Notes: Average hours spent on different activities on a typical workday. During Corona: period of school closures due to COVID-19. Before Corona: period before the school closures. Low-ed: parents without a university degree. High-ed: parents with a university degree. Std. err.: standard errors stemming from regressions of hours spent on each activity on a high-ed and female indicator, respectively. Significance levels: *** p<0.01, ** p<0.5, * p<0.1. Data source: ifo Education Survey 2020.

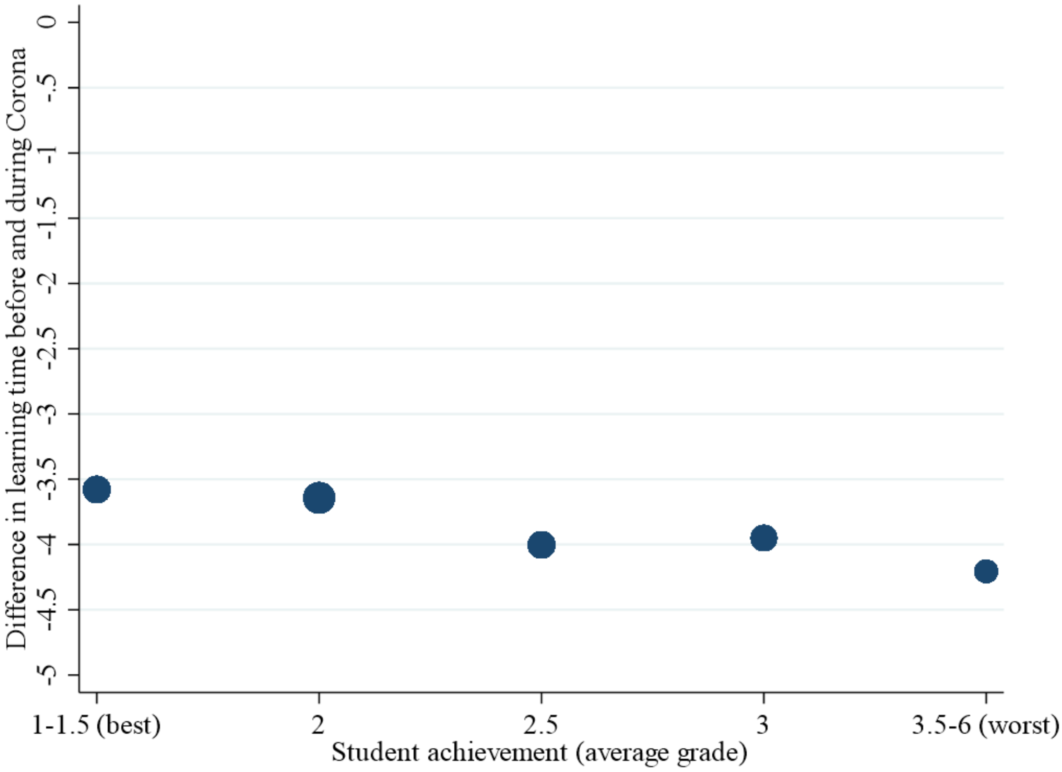
Appendix

Figure A3.1: Activities of Low- and High-Achieving Students Before and During the School Closures



Notes: Average hours spent on different activities on a typical workday. During Corona: period of school closures due to COVID-19. Before Corona: period before the school closures. Low- versus high-achievers: students with an average grade in mathematics and German below versus at-or-above the median for their respective school type. See Table 3.1 for details. Data source: ifo Education Survey 2020.

Figure A3.2: Reduction in Learning Time by Student Achievement



Notes: Difference in average hours spent on school activities on a typical workday between the period before the school closures and the period of school closures due to COVID-19. Student achievement (average grade): average of school grade in mathematics and German. Size of markers indicates number of observations. Average grades range from 1 (best grade) to 6 (worst grade). To ensure sufficient size of each category, observations are grouped as follows: grade 1.5 or better (20 percent of the sample), grade 2 (28 percent), grade 2.5 (20 percent), grade 3 (18 percent), and grade 3.5 or worse (14 percent). Data source: ifo Education Survey 2020.

Table A3.1: Timing of School Re-openings by State and Class Type

State	Transfer Classes		Graduation Classes		All other Classes	
	Re-opening Date	School Operations	Re-opening Date	School Operations	Re-opening Date	School Operations
Baden-Wuerttemberg	18 May 2020	partial	4 May 2020	partial	15 June 2020	partial
Bavaria	11 May 2020	partial	27 April 2020	partial	15 June 2020	partial
Berlin	4 May 2020	partial	20 April 2020	partial	1 June 2020	partial
Brandenburg	4 May 2020	partial	27 April 2020	partial	25 May 2020	partial
Bremen	4 May 2020	partial	27 April 2020	partial	15 June 2020	partial
Hamburg	4 May 2020	partial	27 April 2020	partial	1 June 2020	partial
Hesse	18 May 2020	partial	27 April 2020	partial	2 June 2020	partial
Lower Saxony	4 May 2020	partial	27 April 2020	partial	15 June 2020	partial
Mecklenburg-West Pomerania	4 May 2020	partial	27 April 2020	partial	1 June 2020	partial
North Rhine-Westphalia	7 May 2020	partial	27 April 2020	partial	2 June 2020	partial
Rhineland-Palatine	25 May 2020	partial	25 May 2020	partial	8 June 2020	partial
Saarland	4 May 2020	partial	4 May 2020	partial	8 June 2020	partial
Saxony	6 May 2020	full	6 May 2020	partial	18 June 2020	partial
Saxony-Anhalt	4 May 2020	partial	4 May 2020	partial	15 June 2020	partial
Schleswig-Holstein	6 May 2020	partial	27 April 2020	partial	1 June 2020	partial
Thuringia	11 May 2020	partial	4 May 2020	partial	2 June 2020	partial

*Notes:*Transfer classes (*Uebertrittsklassen*) are in the last year of primary school, which in most states corresponds to grade 4. Graduation classes end secondary school in that year (which can be grade 9, 10, 12, or 13, depending on the type of school). The re-opening dates for all other classes refer to the date when all classes had the opportunity to return to school. 'Partial' school operations mean that not all students in the respective classes were in school at the same time, but â€“ in accordance with school-specific rules â€“ were in school part of the time and otherwise at home. Source: `https://deutsches-schulportal.de/bildungswesen/schuloeffnung-das-haben-die-laenderchefs-entschieden/` [access June 7, 2021].

Table A3.2: Comparison of Analysis Sample to Microcensus Data

	Microcensus (1)		Analysis Sample (2)	
Child Characteristics				
School Type				
Primary Sschool	0.335	(0.002)	0.361	(0.014)
Upper-Track Secondary School (*Gymnasium*)	0.301	(0.002)	0.301	(0.014)
Other Secondary School	0.364	(0.002)	0.338	(0.014)
Age	12.070	(0.016)	12.480	(0.106)
Girl	0.491	(0.002)	0.483	(0.015)
Living with Both Parents	0.783	(0.002)	0.800	(0.012)
Parent Characteristics				
Educational Attainment				
Mother with *(Fach-)Abitur*	0.362	(0.002)	0.437	(0.021)
Father with *(Fach-)Abitur*	0.410	(0.003)	0.474	(0.021)
Working status				
Mother Works Full-time	0.211	(0.002)	0.233	(0.013)
Father Works Full-time	0.876	(0.002)	0.671	(0.015)
West Germany	0.832	(0.002)	0.795	(0.012)
Observations	49,621		1,099	

*Notes:*Means; standard errors in parentheses. Column (1): all children aged below 20 years in general schools in the Microcensus 2015 (representative of the German population). Column (2): our analysis sample, referring to youngest school-aged child of parents in our survey data. Data sources: Microcensus 2015 and ifo Education Survey 2020.

Table A3.3: Sample Characteristics

	Sample Mean	Regr. of High-Achiever on Sample Charact.	
		Coef.	Std. err.
	(1)	(2)	(3)
Child Characteristics			
School type			
Elementary School	0.361		
Upper Second. School (*Gymnasium*)	0.301	-0.014	-0.001
Other Secondary School	0.338	0.033	(0.049)
Age	12.480	-0.024	(0.007)***
Girl	0.484	0.038	(0.031)
Single Child	0.383	-0.010	(0.033)
Parent Characteristics			
Female	0.490	-0.047	(0.034)
University Degree	0.273	0.015	(0.040)
Single Parent	0.166	0.026	(0.047)
Child not in Household	0.080	-0.109	(0.060)*
Parent in Home Office+	0.342	0.105	(0.037)***
Work Hours	29.110	0.000	(0.001)
Partner at Home++	0.185	0.020	(0.042)
Household Income	3370.4	0.002	(0.011)
West Germany	0.795	-0.063	(0.039)
Observations	1,099	978	
R-squared		0.059	

Notes: Column (1): sample means. Columns (2)–(3): dependent variable: dummy for high-achieving student (average grade in mathematics and German at or above the median for respective school type). In the regression, work hours and household income are divided by 100. Significance levels: *** p<0.01, ** p<0.5, * p<0.1. Data source: ifo Education Survey 2020. + Parent in home office: responding parent reports a positive number of hours working from home during the period of school closures. ++ Partner at home: dummy=1 if additional adult in household who works less than 20 hours per week during period of school closures, 0 otherwise.

Table A3.4: Parental Assessment of Whether Activities are Beneficial for Child Development

				Unconditional Gap		Conditional Gap	
	Average	Low-Achiev.	High-Achiev.	Gap	Std. Err.	Gap	Std. Err.
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Attending School	0.97	0.96	0.97	0.006	(0.012)	-0.000	(0.012)
Learning for School	0.93	0.91	0.95	0.034	(0.016)**	0.032	(0.017)*
Reading	0.89	0.83	0.91	0.077	(0.021)***	0.074	(0.022)***
Music and Creative Work	0.82	0.78	0.82	0.039	(0.025)	0.029	(0.027)
Physical Exercise	0.95	0.94	0.95	0.009	(0.015)	0.015	(0.015)
Watching TV	0.30	0.27	0.35	0.083	(0.030)***	0.066	(0.031)**
Gaming	0.22	0.19	0.25	0.060	(0.026)**	0.055	(0.028)**
Social Media	0.24	0.23	0.27	0.036	(0.028)	0.029	(0.030)
Online Media	0.34	0.33	0.39	0.063	(0.031)**	0.048	(0.033)

Notes: Dummy=1 for respondents who say activity is 'very beneficial' or 'rather beneficial' for the further development of their child (on a five-point scale from 'not beneficial at all' to 'very beneficial'). Low- versus high-achievers: students with an average grade in mathematics and German below versus at-or-above the median for their respective school type. Std. err.: reports standard errors of regression from dummy=1 for high-achievers on hours in each category. Conditional gap: see Table 3.2 for controls. Significance levels: *** p<0.01, ** p<0.5, * p<0.1. Data source: ifo Education Survey 2020.

Table A3.5: Average Student Activities Before and During the School Closures

	During Corona (1)	Before Corona (2)	Difference (3)
School activities			
Aggregate	3.62	7.42	-3.80
Attending School	0.90	5.92	-5.01
Learning for School	2.72	1.51	1.21
Conducive Activities			
Aggregate	3.20	2.89	0.31
Reading	0.77	0.67	0.10
Music and Creative Work	0.77	0.61	0.16
Physical Exercise	1.67	1.62	0.05
Detrimental Activities			
Aggregate	5.22	3.96	1.26
Watching TV	1.42	1.20	0.22
Gaming	1.49	1.04	0.45
Social Media	1.31	0.95	0.36
Online Media	1.00	0.77	0.23

Notes: Average hours spent on different activities on a typical workday. During Corona: period of school closures due to COVID-19. Before Corona: period before the school closures. Data source: ifo Education Survey 2020.

Table A3.6: Distribution of School-Related Activities During the School Closures

	At Most...								
	0 Hours (1)	1 Hour (2)	2 Hours (3)	3 Hours (4)	4 Hours (5)	5 Hours (6)	6 Hours (7)	7 Hours (8)	8 Hours (9)
All	0.023	0.144	0.378	0.568	0.742	0.818	0.881	0.918	0.954
Low-Achievers	0.03	0.188	0.435	0.613	0.783	0.849	0.902	0.936	0.961
High-Achievers	0.015	0.104	0.326	0.516	0.701	0.791	0.872	0.91	0.954

Notes: Hours spent on ‘attending school’ or ‘learning for school’ on a typical workday during the period of school closures due to COVID-19. Low- versus high-achievers: students with an average grade in mathematics and German below versus at-or-above the median for their respective school type. Data source: ifo Education Survey 2020.

Table A3.7: Parental Assessment of Home Environment and Child's Learning

				Unconditional Gap		Conditional Gap	
	Average	Low-Achiev.	High-Achiev.	Gap	Std. Err.	Gap	Std. Err.
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Family Coped Well	0.87	0.85	0.90	0.049	(0.021)**	0.060	(0.022)***
Psychological Burden for Child	0.38	0.39	0.36	-0.030	(0.031)	-0.060	(0.033)*
Psychological Burden for Parent	0.38	0.37	0.34	-0.028	(0.031)	-0.046	(0.032)
Argued More with Child	0.28	0.30	0.24	-0.055	(0.028)*	-0.080	(0.030)***
Assessment of Home Learning Environment	3.86	3.70	4.01	0.312	(0.063)***	0.289	(0.067)***
Satisfied with School Activities	0.57	0.49	0.62	0.131	(0.032)***	0.113	(0.034)***
Child Learned Much Less	0.64	0.72	0.58	-0.142	(0.031)***	-0.135	(0.032)***

Notes: Rows 1–4 and 7: probability that statement 'fully applies' or 'rather applies' (on a five-point scale from 'does not apply at all' to 'fully applies'); question wording: 'Our family coped well with the situation during the school closures.'; 'The phase of school closures was a great psychological burden for my child/for me.'; 'I argued with my child during the school closures more than usual.'; 'My child has learned much less during the school closures than usual in school.' Row 5: average grade provided on 5-point scale (1='insufficient', 5='very good'); question wording: 'How would you evaluate your child's learning environment at home during the period of several weeks of Corona-related school closure, e.g., in terms of available computers or space to work?' Row 6: probability that respondents are 'very satisfied' or 'satisfied' (on a five-point scale from 'very unsatisfied' to 'very satisfied'); question wording: 'Overall, how satisfied are you with the activities your child's school carried out during the several weeks of Corona-related school closure?' Low- versus high-achievers: students with an average grade in mathematics and German below versus at-or-above the median for their respective school type. Std. err.: standard errors stemming from regressions of the respective outcome variable on a high-achiever indicator. Conditional gap: see Table 3.2 for controls. Significance levels: *** p<0.01, ** p<0.5, * p<0.1. Data source: ifo Education Survey 2020.

Table A3.8: Student Activities Before and During the School Closures by School Type

	During Corona			Before Corona			Difference During-Before		
	Primary	Secondary		Primary	Secondary		Primary	Secondary	
		Upper-Track	Other		Upper-Track	Other		Upper-Track	Other
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
School Activities									
Aggregate	3.62	3.91	3.37	6.98	7.97	7.40	-3.36	-4.06	-4.03
Attending School	1.08	0.75	0.85	5.63	6.17	5.99	-4.55	-5.42	-5.14
Learning for School	2.54	3.16	2.52	1.35	1.80	1.41	1.19	1.36	1.11
Conducive Activities									
Aggregate	3.89	2.84	2.79	3.41	2.50	2.69	0.48	0.34	0.09
Reading	0.93	0.74	0.62	0.82	0.56	0.59	0.10	0.18	0.03
Music and Creative Work	0.93	0.75	0.61	0.76	0.53	0.53	0.17	0.23	0.08
Physical Exercise	2.03	1.35	1.56	1.83	1.41	1.57	0.21	-0.06	-0.02
Detrimental Activities									
Aggregate	3.71	5.85	6.29	2.90	4.17	4.92	0.81	1.68	1.37
Watching TV	1.37	1.45	1.45	1.16	1.14	1.28	0.21	0.31	0.17
Gaming	1.11	1.48	1.91	0.83	0.94	1.37	0.28	0.54	0.55
Social Media	0.54	1.73	1.76	0.39	1.21	1.32	0.15	0.52	0.44
Online Media	0.69	1.19	1.16	0.52	0.87	0.95	0.17	0.32	0.21

Notes: Average hours spent on different activities on a typical workday. During Corona: period of school closures due to COVID-19. Before Corona: period before the school closures. Primary: students in primary school. Upper-track: students in upper-track secondary school (Gymnasium). Other: students in other secondary school. Data source: ifo Education Survey 2020.

General Overview of the German School System

To provide context for the presented results, this appendix briefly presents some stylized facts about the German school system.

Germany's education system is decentralized, with each of the 16 states holding legislative and executive power over their respective school system. Although there are some differences between states, the general structure of the school system is similar across states. In general, enrollment in primary school is based on the catchment area in which a child lives. Generally based on their achievement in the fourth and final grade of primary school, children are usually sorted into one of two or three secondary-school tracks at age ten. The exact designations vary from state to state, but the possible tracks typically include a basic track (five or six years), a middle track (six years), and a high track (eight or nine years). The high track leads to the university entrance qualification (*Abitur*). Only a small share of 11 percent of schools in Germany are private schools, and many of these schools have ecclesiastic operators.

Educational inequality in Germany is quite high. For example, comparing PISA test scores of 15-year-olds in mathematics, students from families with low socioeconomic status (defined as being in the lowest decile of the PISA Index of Economic, Social and Cultural Status) lag behind their high-SES peers (top decile) by a test-score difference equivalent to four years of schooling (Lergetporer et al., 2020). Differences in PISA test scores by students' socioeconomic background in Germany are the third largest among all OECD countries (OECD, 2020).

4 School Track Decisions and Teacher Recommendations: Evidence from German State Reforms*

4.1 Introduction

Many educational systems around the globe employ some sort of school tracking, i.e. the streaming of students into different school tracks according to their career goals or educational needs (e.g., Betts, 2011; OECD, 2013). While some countries, like Germany or Austria, separate students into different schools as early as age 10, other countries, e.g. Ireland or the Netherlands, do so at later ages. Yet others, like the U.S. or the U.K., have a more comprehensive schooling system but still stream students into different tracks within schools.

Common to these tracking systems is that they need to rely on admission or placement policies that govern the allocation of students into the different tracks. Some placement polices are non-selective, ensure free choice of school tracks and students sort themselves into the different streams according to their preferences. Other placement policies are highly selective, base track admission on prior performance and students may only attend higher tracks if they show proof of academic accomplishments.[1] Surprisingly, we know very little about the impact of selective placement policies for school track assignment despite their common and diverse usage.[2] In this chapter, I study how changes in admission polices influence students' educational outcomes in the short and medium run.

---

* Among others, this paper uses data from the IGLU, PISA and IQB National Assessment Studies. The data were made available by the Research Data Center at the Institute for Educational Quality Improvement (FDZ at IQB) with project numbers 1803-05a/b, March 2018, 1910-08a/b, October 2019 and 2004-09a, April 2020. PISA-2000 was conceived in Germany as a national research program by the German PISA Consortium (Jürgen Baumert, Eckhard Klieme, Michael Neubrand, Manfred Prenzel, Ulrich Schiefele, Wolfgang Schneider, Klaus-Jürgen Tillmann, Manfred Weiß). The lead was Professor Dr. Jürgen Baumert, Max Planck Institute for Educational Research, Berlin. Results of the primary research are published in Baumert et al. (2001b, 2002, 2003). The survey instruments are documented in Kunter et al. (2002). I thank the German PISA Consortium and the Research Data Center (FDZ) in Berlin for permission and support of the secondary analysis.

[1] 43 percent of students in OECD countries attend selective schools whose admission depends on prior achievements, mostly in the form of good academic performances or recommendations of feeder schools. The respective share varies heavily by country ranging from more than 90 percent in the Netherlands and Japan to less than five percent in Spain and Finland. (e.g. OECD, 2013).

[2] So far, the tracking literature has mainly focused on investigating the impact of earlier vs. later tracking. The main take-away from this literature is that students' educational outcomes depend more on parental background if tracking takes place early (e.g., Meghir and Palme, 2005; Hanushek and Woessmann, 2006; Malamud and Pop-Eleches, 2011; Pekkala Kerr et al., 2013; Matthewes, 2020).

To that end, I exploit German state-level variation in the tracking procedure. In Germany, tracking takes place early, at the transition from primary to secondary schools (at age 10). While all students receive a recommendation by their primary school teacher on which school track she advises the child to pursue, the extent to which this recommendation is used as necessary criterion for being admitted to academic schools (the highest track) varies by state.[3] States with binding teacher recommendations enact quite selective placement policies since children can only attend academic schools if they have a recommendation to do so. States with non-binding recommendations enact rather generous placement policies since each student still receives a recommendation but may attend any school regardless of the recommendation outcome. Over time, several states reformed the binding nature of teacher recommendations, with some moving from non-binding to binding, some from binding to non-binding and others moving back and fourth. Thus, the reforms induce between-state variation over time in whether teacher recommendations are used as selection criterion for school track admission. As all students in all states always receive a recommendation by their primary school teacher, the reforms allow me to isolate the effect of their binding nature from any informational values that recommendations may also have.

There are several reasons why selective placement polices, such as binding teacher recommendations, affect students' outcomes in the short to medium run. First—while still attending primary school—binding recommendations can serve as incentive for children and parents to increase students' academic performance in order to be accepted to the academic schools (see e.g., Benabou and Tirole, 2000; Lindo et al., 2010, for a more general discussion on performance standards). Second, if teachers are better able to asses a child's academic potential than parents, binding recommendations may lead to a more efficient allocation of students to the different school tracks. Over time, achievements gains from early incentive provision can develop further in secondary schools as skill formation is a dynamic process (e.g., Cunha et al., 2006; Cunha and Heckman, 2007). Both would lead to improved educational outcomes for students in the longer run. On the other hand, hidden psychological costs from incentivizing students, e.g. crowding out of intrinsic motivation, may also dominate—especially in the longer run (see e.g., Benabou and Tirole, 2000). Likewise, parents may be better informed about their child's ability and use this information as basis for their decision. Then, binding recommendations would lead to worse educational outcomes.

To analyze the reform effects empirically, I combine information on state reforms which took place during the 1990s and 2000s with several data sources. First, I use individual student-level data on fourth graders, stemming from the 2001 and 2006 extensions of the Progress in International Literacy Study (IGLU-E) as well as the 2011 and 2016 National Assessment Studies (NAS) to analyze short-term effects on students' achievement in primary school, i.e. *prior to* track assignment. Second, I use administrative school data from the German Statistical

[3] The use of teacher recommendation is not Germany-specific. Other countries, such as Italy or the Netherlands, also employ teacher recommendations to facilitate school track assignment (e.g., Checchi and Flabbi, 2013; Timmermans et al., 2018).

Offices to analyze medium-term effects on academic school attendance in grades five to nine, i.e. *after* track assignment. I complement the analyses with individual student-level data of ninth graders, stemming from the 2000, 2003 and 2006 extensions of the Programme for International Student Assessment (PISA-E) as well as the 2009, 2012 and 2015 NAS to investigate academic school attendance and performance of students attending ninth grade.

Using a differences-in-differences approach that controls for fixed differences between states and years, I investigate the effect of selective placement polices by comparing outcomes of students attending school in states that reformed the bindingness of teacher recommendations to outcomes of students attending school in states that did not implement such reforms. I find that binding teacher recommendations have a substantial impact on students' academic achievement in primary school before track assignment. Conditional on state and school-year fixed effects as well as a rich set of sociodemographic controls, reading (math) achievement is 5.6 (12.2) percent of a standard deviation higher for students who require a teacher recommendation for academic school attendance than for students who have free choice of secondary schools.

Detailed time-use data for 9–10 year old children from the German Socio-Economic Panel (GSOEP) allow me to shed some light on the mechanisms underlying these effects. I find that children spend significantly more time with reading as well as with activities deemed as being rather *conducive* for child development, such as doing sports or making music, which suggests that the achievement effects are likely due to increased time investments. Similarly, children spend more time with their family, suggesting that parents respond accordingly by supporting their child more often after the reforms. Parents also consult their child's teacher more frequently which likely indicates increased 'lobbying' for academic track recommendations.

Next, I show that binding teacher recommendations affect educational outcomes of students in secondary schools as well. Under binding recommendations, students are slightly less likely (albeit not statistically significantly) to attend academic schools in fifth grade. Beyond that, I show that these small and negative effects—measured immediately after tracking—mask important positive effects in the medium run. In particular, reform effects gradually *increase* throughout grades and ninth grade academic track attendance is significant 1.8 percentage points *higher* under binding recommendations. This pattern suggests that the reforms reduce the rates at which students transfer to lower track schools during their secondary school career. Less 'downgrading' to lower track schools eventually increase academic school attendance rates by grade nine. Further analyses show that ninth grade students also perform slightly better in standardized reading tests (albeit not statistically significant) and have better grade point averages in the subjects German and math ($p<0.1$).

Finally, I explore effect heterogeneities by students' socioeconomic background. For the sub-sample of students with information on family background, reform effects are by and large homogeneous across different subgroups.

A series of robustness tests support the main results. Results remain robust to controlling for a rich set of contemporaneous school reforms, economic and education input factors. Moreover, there are hardly any significant differences in pre-trends between reforming and non-reforming states. I also show that results are robust to including state-level controls for government ideology. Furthermore, students who are exposed to binding teacher recommendations do not enter primary school later, nor are they more likely to strategically repeat a grade. Finally, I implement the diagnostic tools by de Chaisemartin and D'Haultfœuille (2020) to show that biases arising from negative weights are likely not an issue in my setting.

This chapter mainly contributes to two strands of the existing literature. First, it extends the literature on school tracking. There is substantial heterogeneity across countries in the extent and age at which students are tracked (Betts, 2011). In Germany, tracking takes place as early as age 10 and students are traditionally tracked in one out of three school types whereas in the US, for instance, students are assigned to different courses within schools. Most of the tracking literature investigates the effects of earlier vs. later tracking (e.g., Meghir and Palme, 2005; Hanushek and Woessmann, 2006; Malamud and Pop-Eleches, 2011; Pekkala Kerr et al., 2013; Borghans et al., 2020; Canaan, 2020) and finds that later tracking increases educational attainment and wages in adulthood, mostly for students from disadvantaged families.[4] Yet, very little is known about the impact of institutional features within the tracking procedure.

Second, the chapter relates to the literature on general admission requirements which are ubiquitous in education. Decisions to receive a scholarship, to participate in advanced courses, or to be promoted into the next grade often depend on some sort of prior student achievement. Consequently, researchers have examined the impact of admission requirements in various contexts. In the short run, empirical studies have found that admission requirements often incentivize students to meet the criteria, thereby increasing overall performance (e.g., Angrist and Lavy, 2009; Pallais, 2009; Jackson, 2010; Behrman et al., 2015; Barrow and Rouse, 2016; Lichtman-Sadot, 2016). Few studies find zero or negative effects for specific student subgroups, mostly from the lower end of the achievement distribution (e.g., Leuven et al., 2010; Lindo et al., 2010). In the longer run, positive effects of admission requirements can be dominated by hidden psychological costs (see e.g., Benabou and Tirole, 2000). Empirical findings are therefore rather inconclusive. While some studies find persistent positive effects of admission requirements (e.g., Jackson, 2010; Leuven et al., 2010) , others find zero (e.g., Abdulkadiroğlu et al., 2014) or even negative (Lindo et al., 2010) effects on students' educational or labor-market realizations.

[4] For Germany, Matthewes (2020) finds positive achievement effects from decreasing tracking intensity by combining the two lower tracks into one comprehensive track. Conversely, Piopiunik (2014) analyses a reform-induced increase in tracking intensity in the state of Bavaria and finds large achievement losses for students who are subject to more intense tracking after the reform. Finally, Dustmann et al. (2017) report zero effects of academic track attendance on educational attainment and earnings in the long run, using an instrumental variable approach which identifies a local treatment effect (LATE) of track assignments for students at the margin between two tracks.

This chapter combines the two literature strands by examining admission policies in the context of school tracking. In that sense, closely related is the study by Guyon et al. (2012) who exploit a reform in Northern Ireland that sharply increased the proportion of students admitted into so-called *grammar schools* (the highest, most selective track). Before the reform, students were selected based on their performance on a national test whereas after the reform admission was left more to the parents' choice. The authors find large and significant improvements in students' long-term educational outcomes due to the reform. However, because identification relies on comparing cohorts right before and after reform implementation, it remains unclear whether students could anticipate the sudden change in placement policies and were able to respond accordingly in earlier grades.

Several papers have investigated the German reforms on the bindingness of teacher recommendations. The sociological literature has mainly focused on evaluating single reforms that took place in one particular state, either in cross-section analyses (e.g., Neugebauer, 2010; Dollmann, 2011, 2016) or within a differences-in-differences framework (e.g., Jähnen and Helbig, 2015; Roth and Siegert, 2015, 2016). These studies generally find none to small negative effects of binding recommendations on academic school attendance, shortly after the transition to secondary schools (usually in grade five). Similarly, an economics paper by Osikominu et al. (2021) analyzes a single reform in the state of Baden-Wuerttemberg in 2011 and finds small negative effects of binding teacher recommendation on academic track attendance in fifth grade. Finally, the contemporaneous paper by Bach and Fischer (2020) investigates short-term effects on students in primary school using several identification strategies: (triple) differences-in-differences specifications that exploit the two latest reforms in Baden-Wuerttemberg and Saxony-Anhalt as well as value-added approaches that investigate effects of binding teacher recommendations more comprehensively. My results that binding teacher recommendations affect student achievement and study effort of fourth graders are broadly in line with their findings. However, I depart from previous studies in two important ways. Contentwise, I provide the first comprehensive analysis of binding teacher recommendations in the short *and* medium run. I find that binding teacher recommendations do not only improve educational outcomes among students in primary school, but effects extend to students in secondary schools. In terms of academic school enrollment, I show that small negative effects in fifth grade mask important positive attendance effects that evolve over time. Particularly, binding teacher recommendations seem to reduce the incidences of students transferring to lower track schools throughout their secondary school careers, resulting in positive attendance effects by grade nine. Methodologically, I apply a 'generalized' differences-in-differences approach that does not only identify from reforms in one or two states, but simultaneously exploits up to 10 state-level changes. Thus, I paint a detailed picture of how reforms affect students across the country. The analyses in this chapter are built on a novel data base that combines data on fourth graders from IGLU-E and NAS as well as on ninth graders from PISA-E and NAS, respectively. As these studies draw representative samples of students in all states and mandate participation, estimated effects are informative for the average student body.

Overall, my results suggest that selective placement policies for school track assignment—in form of binding teacher recommendations—can lead to persistent improvements of students' educational performance in the short and medium run. These findings have important implications for the scientific and political discourse by providing direct evidence that institutional features within the tracking procedure are important for human-capital formation.

The remainder of the paper is structured as follows: Section 4.2 provides detailed background information on the school system and the tracking procedure in Germany. Section 4.3 introduces the empirical model and the data. Section 4.4 presents main results on the short and medium-term effects of introducing binding teacher recommendations. Section 4.5 reports robustness checks, and section 4.6 concludes.

4.2 Institutional Background

This section first provides an overview of the German school system. It then describes the role of teacher recommendations in general as well as the reforms on the bindingness of teacher recommendations in particular. This section first provides an overview of the German school system. It then describes the role of teacher recommendations in general as well as the reforms on the bindingness of teacher recommendations in particular.

4.2.1 German School System

In Germany, responsibility for the school system and therefore decisions regarding educational policies are vested in the 16 federal states. The German constitution even prohibits the federal government to exert influence on the educational policies of the states. Yet, a general assembly of all state ministers of education called *Kultusministerkonferenz (KMK)* aims to harmonize education policies countrywide. Consequently, the general outlook of the school system is fairly uniform, as are degrees or teacher employment conditions. However, some other education policies may still differ across states or may be implemented or abolished at different points in time.[5]

Figure 4.1 provides an overview of the German school system. Compulsory schooling extends from the age of five or six until the age of 18. The comprehensive primary school takes four years (in Berlin and Brandenburg six years) and provides basic education in mathematics, German and several other science and social subjects. In primary school, students are usually taught all main subjects by the same teacher.

Upon completion of primary school, students move on to secondary schools. At this point, children are assigned to one of three different tracks: the basic and intermediate track last

[5] It is thus not surprising that several economic research papers have exploited the characteristics of the German federal system to evaluate educational reforms implemented over time (e.g., Pischke, 2007; Pischke and Wachter, 2008; Dustmann et al., 2017; Marcus and Zambre, 2019; Matthewes, 2020; Obergruber and Zierow, 2020)

though grades nine and ten, respectively, and prepare students for apprenticeship training or other forms of vocational education. The academic track ends with grade 13 (or 12) and leads to the university entrance qualification *Abitur*. While nowadays many different school types (including comprehensive school models) incorporate both the basic and the intermediate track, the academic track is primarily offered by the academic school *Gymnasium*.[6] In general, switching to higher track schools is possible, but rather uncommon. In 2000, only 1.5 percent of students switched to a higher track at any grade throughout grades 5–9. However, the best students from the lower tracks often transfer to higher tracks after successful degree completion.

4.2.2 The Role of Teacher Recommendations

As school track decisions are made at the end of primary school, the transition from primary to secondary school marks an important milestone in the students' further educational careers. To formally structure this transition—particularly to academic schools—entrance examinations were conducted in most of the western German states until the 1960s. As public critique grew in the 1950s (Gass-Bolm, 2005), the state ministers of education decided to facilitate the selection process for the academic schools (Herrlitz et al., 2009). Over time, more and more states abolished entrance examinations and replaced them with teacher recommendations to guide school track decisions for students and their parents (Helbig and Nikolai, 2015). Teacher recommendations were also adopted by the East German states after reunification in 1991. Today, all children in Germany obtain a recommendation at the end of their primary school.

Recommendations are issued by the students' primary school teachers and entail explicit information on which school type she thinks the child should pursue. They are issued in the students' final year of primary school, shortly after the first semester, i.e. in February or March. Recommendations are mainly based on students' grades in their mid-term report card. In some states teachers additionally base their recommendation on their assessment of the student's socio-emotional maturity (Baumert et al., 2010).[7] Appendix Figure A4.1 reports which criteria determine the recommendations. Students' math and German grades are very strong or strong determinants for nearly all teachers. Among soft skills, commitment, concentration and self-reliance are ranked highest with again almost all teachers considering them as (very) strong criteria.

While teacher recommendations are issued to all students in Germany, their bindingness differs across states. In some states, teacher recommendations are non-binding, i.e. the

[6] Roughly 40 percent of students transition to the academic schools (own calculations). Academic schools are also the most important school type on which students receive the university entrance qualification.

[7] For instance, Bavaria uses the students' grade as the only criterion. The Bavarian regulation states that *[i]n order to obtain a recommendation for the Gymnasium (academic track) a student needs to have a GPA of at most 2.33 in the subjects German, math as well as science and local history* (with grades ranging from 1 to 6 and lower values indicating better grades). In Schleswig-Holstein, for instance, the *recommendation shall be based on students' maturity considering the students' current grades* (Kultusministerkonferenz, 2015).

content of the recommendations is purely informative and students can transfer to any school type regardless of the recommendation outcome. In other states, teacher recommendations are binding, i.e., students can only attend the academic schools if they have a recommendation to do so.[8] In these states, if students wish to transfer to the academic schools without an appropriate recommendation, they must pass additional entrance examinations or trial lessons (Kultusministerkonferenz, 2015). Conversely, if students wish to attend a lower track than recommended, they can always do so.

4.2.3 Reforms on the Bindingness of Teacher Recommendations

Since the general adoption of teacher recommendations, states have frequently reformed their binding nature: Several states have abolished binding recommendations to replace them with non-binding ones, and vice versa while other states have switched back and forth. Figure 4.2 and Appendix Table A4.1 present a summary of the state regulations for the time period considered in this chapter since the 1990s.

State reforms are usually accompanied by emotional public debates. Arguments for binding recommendations come from the conservative or liberal camp and are based on the idea of the *Gymnasium* as an elite school (Fokken, 2020): Accordingly, the government should strengthen the academic schools and prevent a decline in performance due to the presence of unsuitable students (e.g., Die Welt, 2014). In addition, teachers are supposedly better able to evaluate the potential of their students than their parents (e.g., Breyton, 2018). Opposing arguments come from the left-leaning camp (Fokken, 2020). The prevailing view here is that non-binding recommendations can provide access to the academic schools for broad groups of the population. Moreover, the bindingness may put strong pressure on students in third and fourth grade and four years may be too short a time to make binding statements about students' future potential (e.g., Schenk, 2010; Otto and Schenk, 2011).

Given the lively political debate, one would expect to see reforms in the bindingness of teacher recommendations particularly after ideological changes in state governments. And indeed, seven of the ten reforms since the early 1990s were implemented after governmental changes (see Table A4.1). Among those seven, all four that involved a change from binding to non-binding recommendations were introduced after a more social government had replaced a more conservative one. Conversely, the remaining three reforms including a change from non-binding to binding recommendations were introduced after a more conservative government had replaced a more social one. To address the potential issue of non-random reform introduction, my main specification controls for a variety of different educational input factors which are observed at the state-year level (see section 4.3.1 for details). Section 4.5 also analyzes potential effects of government ideology on student outcomes. Reassuringly, the main

[8] The bindingness mainly applies to the academic school *Gymnasium*: In almost all states with binding recommendations, students with a recommendation for the basic track can nevertheless attend an intermediate track school (Kultusministerkonferenz, 2015).

effects of binding teacher recommendations are robust to controlling for state government ideology.

4.3 Empirical Strategy and Data

In this section, I first describe the empirical strategy and then present the data used for the analyses.

4.3.1 Empirical Strategy: Differences-in-Differences Approach

My identification strategy exploits variation in the implementation and abolition of binding teacher recommendations across German states and over time in a differences-in-differences framework. By controlling for fixed differences between states and years, I compare outcomes of students attending school in states that changed the bindingness of teacher recommendations to outcomes of students attending school in states that did not. The empirical model can be formalized by the following equation:

$$Y_{ist} = \alpha + \beta Bindrec_{st} + \gamma R_{st} + \delta E_{st} + \lambda X_{ist} + \eta_s + \mu_t + \epsilon_{ist} \tag{4.1}$$

where Y_{ist} is the outcome of interest (e.g., student achievement) for student *i* who attends school in state *s* and is tracked into different schools in school year t. $Bindrec_{st}$ is the treatment indicator which equals 1 if the recommendation is binding and varies at the state and school year level. To account for differences across states and over time, I include state (η_s) and school-year (μ_t) fixed effects. Since treatment varies at the state level, I use a conservative inference and cluster standard errors at the state level to account for potential correlation of error terms within states across years (Athey and Imbens, 2018).[9] Moreover, regressions are weighted by students' sampling probabilities, giving equal weight to each wave.

During the observation period, Germany has undergone several major education policy changes that potentially affect students' outcomes in primary and secondary schools. To rule out any biases arising from omitting these reforms, I include R_{st}, a vector of reform indicators that vary by school year and state. The vector entails the duration of academic schools—whether the academic schools take eight or nine years—,[10] the intensity of school tracking,[11]

[9] I additionally present wild cluster bootstrap p-values, relying on Roodman et al. (2019) in all main tables. These p-values account for a limited number of clusters when analyzing at most sixteen German states.

[10] Several states reduced the length of the academic school while simultaneously increasing the instruction hours in the remaining years. The effects of those reforms have been investigated by Andrietti and Su (2018) or Marcus and Zambre (2019).

[11] Several states reformed whether students attending the two lower tracks are taught comprehensively or further streamed into two separate tracks. The effects of such reforms have been investigated by Matthewes (2020).

the duration of primary school—whether primary schooling takes four or six years—,[12] the basis of the teacher recommendation—whether they are only based on students' grades or also on their socio-emotional maturity—,[13] as well as whether the recommendation has to be explicitly requested.[14]

To further avoid biases from the fact that the timing of recommendation reforms may not be random to the economic and educational performance of the states, I include E_{st} and X_{ist} as additional controls. E_{st} is a vector of economic and educational measures that vary by school year and state. These entail GDP as *overall economic performance measure*, but also average school spending, the number of classes in primary school, average class size in primary school, average hours of instruction in primary school as well as share of full-time employed primary school teachers as *school input factors*. The vector also entails the average share of students starting primary school late. X_{ist} is a vector of various school- and student-level characteristics, including student gender, immigration background, parental occupation, and books at home as well as community location and public school status.

The key identifying assumption is the standard differences-in-differences assumption: Conditional on the rich set of included control variables at the student, school and state level, in the absence of reforms the change in student outcomes in states that reformed the bindingness of teacher recommendations would have been similar to the change in student outcomes in states that did not reform at a given point in time.[15] I will come back to a detailed discussion of potential violations of this assumption in section 4.5.

A related important assumption is that there is no student selection into treatment status. Since recommendation policies vary at the state level, self-selection would imply that students attend a school in a different state. As between-state mobility is relatively low among school-aged children, I argue that the potential for selection bias due to sorting is very low.

4.3.2 Data

To analyze the effects of binding teacher recommendations empirically, I combine information on state reforms with the following three data sources: (i) individual-level data on students in

[12] Except for the states Berlin and Brandenburg where primary school takes six years, students are tracked after grade four. A few sates experimented with 'later tracking' by introducing (and again abolishing) so-called 'orientation grades' where students were comprehensively taught until grade six and subsequently tracked into the different school types. These reforms are described by Helbig and Nikolai (2015).

[13] While most states have a standing rule on the criteria used for the outcome of the recommendations, in a few cases states have reformed those criteria together with the bindingness of recommendations.

[14] This was only the case in Bavaria until 2008. In all other states and years, students automatically receive a teacher recommendation before transition to secondary schools.

[15] Further relaxing the identifying assumption by including state-specific linear time trends is difficult given the data structure. As I am limited to four (in the fourth grade sample) and six (in the ninth grade sample) time-series observations for each state, adding a linear time trend for each state renders coefficients too imprecise for clear inference.

fourth and ninth grade from nation-wide assessment studies, (ii) individual-level data on 9–10 year old children from the German Socio-Economic Panel and (iii) administrative state-level data on the education system from the German Statistical Office. This section presents the four components in turn.

Data on State-wide Educational Reforms

I collect data on a series of state-wide educational reforms related to the tracking procedure. First, I compile information on the bindingness of teacher recommendation using the following sources: For the reforms before 2010, I draw on Helbig and Nikolai (2015). For the subsequent reforms, I gather information from Kultusministerkonferenz (2015), newspaper articles (e.g., Otto and Schenk, 2011), plenary protocols from sessions of the state parliaments as well as individual correspondences with the 16 state ministries of education.

Similarly, I compile information on several other educational reforms which took place during the observation period. The corresponding reform indicators serve as control variables and include the following: the basis of the teacher recommendations (in particular whether they are only based on students' performance or also on their general maturity), whether the recommendation has to be explicitly requested, the duration of primary schools (four vs. six years) as well as the intensity of school tracking (i.e., whether students attending the two lower tracks are taught comprehensively or further tracked into two different school types). Finally, I use information on the duration of the academic schooling (eight vs. nine years) from Marcus and Zambre (2019).

Student Assessments

To analyze short-term reform effects on primary school students, I combine data from the German extension of the Progress in International Reading Literacy Study (IGLU-E) with data from the National Assessment Study (NAS).[16] Both studies are repeated cross-sections, testing students at the end of fourth grade (between April and July). While both studies assess students in German (reading), NAS additionally assesses math. The studies were administered in 2001 and 2006 (IGLU-E) as well as in 2011 and 2016 (NAS) and are representative for all German states.[17] Neither IGLU nor NAS follow individual students over time. However, repeated testing of fourth graders allow me to build a pseudo-panel of German states observed every five years. Reading scores are generally comparable across tests and waves as NAS was explicitly designed to emulate IGLU tests (Pietsch et al., 2009; Bos et al., 2012).[18] In each study and wave, random samples of primary schools are drawn and within each school, one class randomly

[16] For further details see Bos et al. (2007, 2010), Stanat et al. (2014), Schipolowski et al. (2019), and Stanat et al. (2019).

[17] While all 16 German states participate in all four waves, in 2001 only seven states draw larger sample sizes that were fully representative.

[18] In addition, Böhme et al. (2014) compare reading scores among students who have been tested in both studies and show a very high correlation (0.86) between the test scores produced by the IGLU and the NAS items.

participates in the tests (see Richter et al., 2014; Schipolowski et al., 2019, for more details on test administration).

To analyze medium-term effects on students in secondary school, I combine data from the German extension of the Programme for International Student Assessment (PISA-E) and the National Assessment Study (NAS).[19] Whereas the international version of the PISA test samples 15 year old students, the German extension tests ninth graders. PISA-E and NAS thus build repeated cross-sections, testing students at the end of grade nine (between May and July). The tests have been administered in 2000, 2003 and 2006 (PISA-E) as well as in 2009, 2012 and 2015 (NAS) and are representative for ninth graders in the 16 German states.[20] Again, neither PISA nor NAS follow individual students over time, but repeated testing allow me to build a pseudo-panel of German states observed every three years.

While PISA regularly tests relevant skills in math *and* reading, NAS alternates tested domains every other wave. Consequently, NAS 2009 and 2015 assess reading while NAS 2012 assesses math. Therefore, all results on reading achievement are conducted without students tested in 2012.[21] Achievement scores are generally comparable across studies and waves as the NAS was explicitly designed to emulate the PISA-E testing procedure (Hartig and Frey, 2012; Böhme et al., 2014). In each study and wave, random samples of schools were drawn to be representative at the federal state level and within each school, one class randomly participated in the tests (see Baumert et al., 2001a, 2004; Sachse et al., 2012; Lenski et al., 2016; Schipolowski et al., 2018a,b, for more details on test administration).

Tables 4.1 and 4.2 present student-level descriptive statistics of fourth and ninth graders, respectively. I consider a student as subject to binding teacher recommendations if the recommendation was binding in her current state of school attendance at the time she was tracked into the secondary schools.[22] The fourth grade sample consists of approximately 70,000 students who were streamed into secondary schools between 2001 and 2018 and the ninth grade sample consists of more than 220,000 students who were streamed into the secondary

[19] NAS replaced PISA-E after 2006. See Baumert et al. (2002), Prenzel et al. (2007), Baumert et al. (2009), Prenzel et al. (2010), Köller et al. (2011), Pant et al. (2015), and Stanat et al. (2018) for details.

[20] In 2006, the KMK decided to replace the state-level representative samples of PISA-E by the NAS (Kultusministerkonferenz, 2006). Since then, PISA is still conducted in Germany but with much smaller sample sizes to only represent the overall student body of 15 year old in Germany.

[21] While it is in principal also possible to evaluate math achievement for the subset of students tested in 2000, 2003, 2006 and 2012, I abstain from the respective analyses due to the following two reasons: (i) By excluding students assessed in 2009 and 2015, I loose a substantial amount of observations (more than 100,000). (ii) Unlike the reading test, the math test was re-scaled in 2003, which renders the comparability of PISA 2000 to the remaining waves unclear.

[22] In the fourth grade sample, a students' school-year of tracking depends on the duration of primary school. In the ninth grade sample, a students' school-year of tracking depends on the duration of primary school in her current state of school attendance as well as on self-reported grade retention. Important for identification, retention was not affected by reforms on the bindingness of teacher recommendations (results available upon request).

schools between 1994 and 2012. Student assessments were accompanied by comprehensive school, student and parent questionnaires covering a wide range of questions on sociodemographic characteristics and family background. While test participation is always compulsory for students, completing the student questionnaire is only compulsory in some states and completing the parent questionnaire is always voluntary. As a result, non-response rates for the family background questionnaires are much larger than for the test items (for example, response rates to the parent questionnaire are 72 percent in the fourth grade sample and 80 in the ninth grade sample). I select a core set of student and school level controls available in each wave and harmonize control variables across fourth and ninth grade samples.[23]

The German Socio-Economic Panel

To further investigate behavioral responses of students and their parents, I use data from the German Socio-Economic Panel (GSOEP). Since 2010, the GSOEP contains a mother-child questionnaire administered to parents of 9–10 year old children (see Schröder et al., 2013, for further information). The questionnaire collects detailed information on children's daily lives and is supplemented with questions on background characteristics of children and parents.

I focus on the following variables: First, I exploit detailed time-use information on leisure activities:[24] (i) reading; (ii) watching TV; (ii) playing on the computer; (iv) surfing on the internet; (v) listening to music; (vi) making music; (vii) dancing or theater; (viii) doing sports; (ix) doing technical work; (x) drawing, and (xi) spending time with the family. On a 5-point scale parents could indicate how much time their children spend on each activity.

Following Grewenig et al. (2020), I group activities into four categories: reading (which is directly related to students' reading test scores), activities rather *detrimental* to child development (activities (ii)–(v)), activities rather *conducive* to child development (activities (vi)–(x)) and *family* activities (activity (xi)).[25] Second, I exploit information on whether parents report frequently consulting their childrens' teachers.[26] The sample consists of approximately 4,400 students who were streamed to secondary schools between 2005 and 2017.

Administrative School Data

Finally, I use data on general schools (*allgemeinbildende Schulen*) provided by the German Statistical Office (Statistisches Bundesamt, 1991-2016). The administrative data comprises annual state-level information on the number of students in each track and grade for the years

[23] Due to lack of availability I use school-level controls only for the analyses of the fourth graders, but not of the ninth graders.
[24] I only use items which have been consistently asked throughout all GSOEP waves.
[25] Grewenig et al. (2020) show that categorization of *detrimental* and *conducive* activities reflects parental beliefs about how beneficial those activities are for child development.
[26] Teacher's consulting is elicited as follows: *'How often do you or other family member seek contact with the school?'* Respondents could tick a box if they *'frequently consult teachers outside of regular meeting hours.'*

1991 to 2016. To obtain state-wide information on the share of academic track students, I divide the number of academic school students in each grade and school year by the total number of students in the respective grade-year cell. I enrich this data with the share of issued recommendations for the academic schools collected through personal correspondence with the 16 federal ministries of education.[27] The administrative data further includes comprehensive state-wide information on school input factors which are used as additional controls (see section 4.3.1 for details).

4.4 Main Results

In this section, I discuss the main results. First, I consider short-term effects on students in primary school. Second, I estimate medium-term effects on students in secondary schools. Finally, I explore potential effect heterogeneities with respect to the students' family background.

4.4.1 Students in Primary School

This section sheds light on the academic performance of students in primary school *before* school track assignment. I first analyze how the reform affects achievement of students in fourth grade. Then, I turn to examining behavioral responses of students, parents and teachers.

Student Achievement

I start the discussion on short-term effects with student achievement in reading and math among fourth graders. If binding teacher recommendation indeed serve as incentive to improve academic performance prior to track assignment, one would expect to find positive reform effects on the outcomes discussed here.

Panel A of Table 4.3 presents the main differences-in-differences results on student achievement, using equation 4.1. The dependent variables are grade four test scores in reading, standardized to have mean zero and standard deviation one to facilitate interpretation. Column (1) shows the basic results, controlling only for state and school year of transition fixed effects. Reading achievement among students who require a respective teacher recommendation to attend academic schools is 6.4 percent of a standard deviation higher than that of students with free choice of secondary schools.

In columns (2)–(4), I gradually include controls that account for potential differences (i) in educational reforms that were contemporaneously implemented, (ii) in the overall economic

[27] Overall, I receive data on recommendations for eight states and multiple school years (108 state-year observations).

condition and schooling input factors, and (iii) in the sociodemographic composition of the student body within states and over time. The estimates in column (4)—controlling for the full set of background characteristics—are somewhat smaller in size than the estimates in the first column, suggesting that the unadjusted differences-in-differences estimates are slightly upward biased. Most importantly, all estimates remain positive and statistically significant. In the full specification, introducing binding recommendations is associated with a significant increase in reading achievement by 5.6 percent of a standard deviation.

To explore the dynamic of reforms effects over time, Panel A of Appendix Figure A4.2 additionally depicts non-parametric event-study estimates which are obtained by including an indicator for the first cohorts with binding teacher recommendations as well as lead and lag indicators besides state and school year fixed effects. The depicted pre-reform effect is in line with the common trend assumption. The coefficient on the lead dummy is economically and statistically insignificant, suggesting that students in states that switched to binding teacher recommendations were on similar pre-reform trends as students in states that remained with free choice of secondary schools. Though shy of significance, the first cohort exposed to binding teacher recommendations experiences an increase in reading test scores and the improvement seems to remain persistent over time.

Panel B of Table 4.3 presents differences-in-differences results on math achievement which is only assessed in 2011 and 2016. Consequently, identification here stems from changes in the achievement of students in states that reformed their teacher recommendations between 2011 and 2016 (Baden-Wuerttemberg and Saxony-Anhalt). I find that binding teacher recommendations substantially increase students' math scores. Reassuringly, reform-effect patterns through columns (1) to (5) are remarkably similar to those found for reading. In the full specification, math achievement among students who require a respective teacher recommendation to attend academic schools is 12.2 percent of a standard deviation higher than that of students with free choice of secondary schools.

In sum, the results in Table 4.3 suggest that selective placement polices for school track assignment—such as binding teacher recommendations—can increase students' academic achievement in primary school even before track assignment takes place. The depicted achievement effects are substantial. Considering the rule of thumb that average student learning in a year is equivalent to about one-quarter to one-third of a standard deviation, the reform effects amount to what students roughly learn during a fifth (reading) to a third (math) of a school year.

Behavioral Responses

I now turn to investigating behavioral responses as potential mechanisms for the achievement effects revealed in the previous section. Behavioral responses could stem from students, parents and teachers which are investigated in turn.

First, I examine students' responses. One obvious explanation for the positive achievement effects is that binding teacher recommendations incentivize students to put more effort into studying. To explore this channel, I draw upon detailed time-use information on various leisure activities of 9–10 year old students collected as part of the GSOEP. Panel A of Table 4.4 depicts reform effects on children's time spent with reading, standardized to have mean zero and standard deviation one. When teacher recommendations become binding, children spend significantly more time with reading (column(4)), suggesting that the reform effects on students' test scores are indeed driven by increased time investments into the development of reading skills. Similarly, Panel B depicts reform effects on an index, summarizing child's time spent with other leisure activities deemed as being rather *conducive* for skill development, using the method by Kling et al. (2007). The index includes making music, dancing, sports, technical work, and drawing. The results suggest that reforms also significantly increase the amount of time that children spend with these rather conducive activities. In contrast, Panel C depicts reform effects on time spent with activities deemed as being rather *detrimental* to child development, i.e., watching TV, playing computer, surfing on the internet or listening to music. The estimates indicate zero to slightly negative (albeit not significant) reform effects on time spent on these rather detrimental activities. Thus, the bindingness reforms do not only serve as incentives for children to spend more time with activities directly related to skills taught in school (e.g., reading), but also with other conducive leisure activities more indirectly related to skill development.

Next, I examine parental responses. Similar to students, binding recommendations may serve as incentives for parents to support the skill development of their children. To explore this channel, I investigate how much time children spend with their family. Panel D of Table 4.4 shows significant positive reform effects, suggesting that binding teacher recommendations also encourage parental time investments. Besides, the reforms may induce parents to exert influence on the content of the teacher recommendation. For instance, parents could seek out the child's teacher more frequently to 'lobby' for an academic school recommendation. Panel E of Table 4.4 therefore analyzes reform effects on whether parents consult their child's teacher on a regular basis. Column (4) reveals that parents are indeed 13 percentage points more likely to frequently consult their child's teacher when recommendations become binding which is a large and significant increase from a baseline share of 43.4 percent.

Finally turning to teachers' responses, I examine the number of recommendations for academic schools. There are several reasons why teachers would increase academic recommendations in response to the reform. First, they may become more lenient when the future career of their students depends more on their assessment. Second, they may reward the students' achievement gains discussed in section 4.4.1. Finally, they may simply yield to the parents' lobbying efforts discussed above. Appendix Table A4.2 depicts reform effects on the number of academic school recommendations at the state level. Although not statistically significant, the table provides suggestive evidence that—if anything—the share of academic school recommendations increases.

In sum, results suggest that the achievement gains in primary school are likely due to increased time investments in students' skill development by children and parents. Parents also seek out the child's teacher more frequently. These efforts are then rewarded by teachers who tend to issue more academic school recommendations in response to the bindingness reforms.

4.4.2 Students in Secondary School

This section investigates educational outcomes of students in secondary school. I first analyze reform effects on academic school attendance. Then, I examine academic performance of students through ninth grade, i.e. several years *after* track assignment.

Academic School Attendance

Reform effects on educational outcomes in secondary school are ex ante less clear. On the one hand, students can be allocated to the different tracks more efficiently if teachers are better able to assess the child's academic potential than parents. Likewise, the positive achievement effects from primary school (see section 4.4.1) can spillover to secondary schools. Both would lead to improved educational outcomes in the medium run. On the other hand, psychological costs of incentivizing students early can dominate in the longer run or parents may hold superior information on the child's academic potential. In these cases, binding recommendations would lead to worse educational outcomes. With respect to school attendance, binding recommendation may also prevent some students from transitioning to the highest track, resulting in fewer students attending academic schools.

To investigate the relevance of the opposing effects, Table 4.5 presents reform impacts on academic school attendance in grades five to nine, using state-level data from the German Statistical Office. Since I only observe the average share of students attending academic school in a given state and over time, the corresponding estimates are based on an adjusted 'state-level' version of equation 4.1 which controls for potential differences (i) in educational reforms that were contemporaneously implemented and (ii) in the overall economic condition and schooling input factors.[28]

Column (1) shows negative, statistically insignificant reform effects of one percentage point on the share of fifth grade students, attending academic schools. Finding negative attendance effects directly after track assignment is consistent with earlier studies (e.g., Jähnen and Helbig, 2015; Osikominu et al., 2021). But I also show that the effects observed in fifth grade mask important positive attendance effects in the medium run. Specifically investigating academic school attendance through grades five to nine in columns (2) to (5), I find that reform effects gradually increase across grades. In grade nine, the introduction of binding recommendations significantly ($p<0.1$) *increases* academic school attendance by 1.2 percentage points. This

[28] Basic differences-in-differences results, controlling for state and school year of transition fixed effects only, are very similar to those depicted in Table 4.5 (results available upon request).

pattern suggests that the bindingness reforms decrease the rates at which students transfer to lower track schools during their secondary school career.[29] The reduced incidences of students 'downgrading' school tracks subsequently manifest themselves in higher academic school attendance rates by grade nine.

Data from PISA-E and NAS allow me to additionally analyze academic school attendance at the individual student level. Panel A of Table 4.6 depicts the corresponding reform effects using equation 4.1. Overall, the estimates confirm significant positive effects of binding teacher recommendations on the share of ninth grade students attending academic schools. In all specifications, effects are highly significant ($p<0.05$) and amount to roughly 2 percentage points.[30] Panel B of Appendix Figure A4.2 additionally depicts the evolution of reform effects over time by applying a non-parametric event-study specification. Again, the depicted pre-reform patterns are mostly in line with the common trend assumption as coefficients on the most recent lead dummies (back to 10 years prior reform implementation) are economically and statistically insignificant. In contrast, the cohorts exposed to binding teacher recommendations experience an increase in ninth-grade academic school attendance and the corresponding increase seems to gradually phase in over time.

In sum, I find that positive effects of binding teacher recommendations indeed dominate in the medium run. This also suggests that selective placement polices for school track assignment may not only lead to short-term improvements in educational outcomes but may be also beneficial for the students' future educational path.

Academic Performance

To complement the above analyses on academic track attendance, I now turn to investigating students' academic performance in ninth grade. The results in Panel B of Table 4.6 suggest that the implementation of binding teacher recommendations increases reading achievement among ninth graders. Effects are statistically indistinguishable from zero, but estimates are still sizable (2 to 5 percent of a standard deviation) and represent about 50 to 75 percent of the achievement gains found for primary school students. Panel C of Table 4.6 reports reform effects on students' grade point average (GPA) in German and math, standardized to have mean zero and standard deviation 1. In the main specification in column (4), GPA among

[29] Later transitioning to lower track school is rather common in Germany. In 2000, 13 percent of all students have switched to lower track schools by grade nine. Interestingly, the vast majority (70 percent) of those 'downgraders' has done so during grades seven to nine (i.e., after they have spend some years on the higher track schools)—which is consistent with the depicted reform effect patterns on academic school attendance. In contrast, 'upgrading' to higher track schools is very uncommon as only 1.5 percent have ever done so by grade nine.

[30] Attentive readers may notice that I lose a few observations when including economic controls in column (3). This is because I observe average hours of instruction in primary school (one of the included control variables) starting in 1992. The few dropped observations are students who transferred earlier due to multiple grade retention. Importantly, grade retention is not affected by the reforms (results available upon request).

students who require a respective teacher recommendation for academic school attendance is 3.8 percent of a standard deviation ($p<0.1$) better than among students with free choice of school tracks.

I conclude that binding recommendations lead to persistent improvements in students' educational outcomes in the medium run.

4.4.3 Heterogeneities by Socioeconomic Background

This section explores heterogeneities in reform effects by students' socioeconomic background. While the results so far suggest that binding teacher recommendations increase overall educational outcomes, it remains an empirical question whether reforms differently affect students from different family backgrounds. On the one hand, binding teacher recommendations may particularly favor advantaged students. Because parents with a high socioeconomic status (SES) have better resources (e.g., money, knowledge or time) to support the skill development of their children, high SES students may respond more strongly to the incentives provided by binding recommendations. Similarly, these parents could push their children harder to receive a recommendation for the academic track because they aspire them to obtain a high socioeconomic status, as well. In the medium run, efficiency gains, arising from the changes in the allocation of students to school tracks, may be particularly pronounced among advantaged students since high SES parents overrule the outcome of the recommendation more frequently when there is free parental choice.

On the other hand, binding teacher recommendations may particularly favor disadvantaged students. Since they generally perform worse and invest less into skill development at baseline, any incentives arising from binding recommendations have more room to effectively improve academic achievement among low SES students. For the same reason, disadvantaged students can subsequently profit more from spillover effects throughout secondary school.

To analyze the empirical relevance of the opposing effects, I investigate whether reform effects differ along the following dimensions: whether the student has less than 100 books at home, whether the parents' occupational status according to Ganzeboom et al. (1992) is below the median, whether parents have a university (entrance) degree,[31] and whether the student is a first or second generation migrant. For each characteristic, I extend equation 4.1 to include a full interaction between the respective socioeconomic variable and the reform indicator.

The estimates for the fourth and ninth grade samples are presented in Table A4.3. As I can only estimate effects for the sub-sample of students with available information on their family

[31] Because data on parental education was not consistently collected in IGLU-E, PISA-E and NAS, I investigate heterogeneities by whether parents have a university degree in the fourth grade sample and by whether parents have a university entrance degree (*Abitur*) in the ninth grade sample.

background, findings regarding the heterogeneity should be interpreted with caution.[32] By and large, the table depicts homogeneous effects across student subgroups.[33] The coefficients on the interaction term between reform indicator and socioeconomic characteristic is statistically distinguishable from zero in three out of 20 cases. Reform effects on reading and math achievement in fourth grade are larger for students with less than 100 books at home and GPA improvements in ninth grade are smaller for migrant students.

4.5 Robustness

This section challenges the robustness of the main results. It first discusses the validity of the identifying assumption. Then, it confirms the comparability of tests items used in PISA and NAS. Finally, it evaluates the properties of the differences-in-differences estimator by performing the diagnostic test proposed by de Chaisemartin and D'Haultfœuille (2020).

4.5.1 The Validity of the Identifying Assumption

The differences-in-differences model identifies the effect of binding teacher recommendations on students' educational outcomes from policy changes within states and over time. Accordingly, a causal interpretation of the estimated effects relies on the assumption that in the absence of reforms the changes in outcomes in states that reformed the bindingness of teacher recommendations would have been similar to the changes in outcomes in states that did not reform. The event-study graphs presented in section 4.4 provide suggestive evidence in favor of the parallel trend assumption as they mostly depict small and insignificant differences in pre-trend outcomes between reforming and non-reforming states. Though reassuring, parallel trends cannot rule out the occurrence of contemporaneous policy changes. Any state-specific variation over time that (i) is correlated with the timing structure of the bindingess reforms and (ii) contemporaneously affects students' outcomes can still pose a thread to identification. Hence, the main specification also includes a variety of reform indicators that account for major educational reforms conducted during the observation period as well as a rich set of school and economic input factors. If the main results are driven by systematic differences in these inputs, the estimated coefficients on binding teacher recommendation should approach zero when further including them as controls. However, finding that reform effects remain stable throughout all specifications strongly supports the robustness of my findings (see section 4.4 for details).

[32] While test-taking is mandatory for all students, providing further information on the family background is voluntary. Appendix Table A4.4 depicts differences in educational outcomes between students with and without available information on their socioeconomic background. Students with missing values (who are by definition excluded from the heterogeneity analyses) perform significantly worse across all depicted dimensions than students with non-missing values.

[33] Classifying students with missing information as students with a low socioeconomic background and re-running the heterogeneity analyses leads to similar conclusions (results available upon request).

Given that reforms to binding recommendations are mostly implemented by conservative governments (see Appendix Table A4.1), a remaining major concern is that conservative governments undertake other unobserved policy actions which coincide with the reform under study. To get a first impression for the relevance of potential government effects, I regress government ideology at the time of transition to secondary school on student outcomes in a differences-in-differences setting. Appendix Tables A4.5 and A4.6 show that conservative prime ministers and ministers of education are associated with better educational outcomes of students' in fourth and ninth grade. The positive performance effects remain partly robust when including the full set of controls (see columns (2) and (4)). Next, I account for government ideology in the main differences-in-differences specification. Appendix Table A4.7 depicts the reform effects on students' educational outcomes in fourth and ninth grade, when controlling for the ideology of the prime minster (see column(1)) or the ideology of the education minister (see column(2)). Reassuringly, the estimated coefficients are of similar magnitude to those presented in section 4.4.

Another concern is strategic retention. To mitigate potential consequences of not meeting the requirements for the academic recommendation, students may either enroll late or strategically repeat grades. Both would lead to students being older at the time of test-taking. Panel A of Appendix Table A4.8 depicts respective reform effects. I find significant negative effects of about one month on age at test taking in fourth grade. The effects likely reflect recent attempts to lower school starting age.[34] For ninth grade students, I find small and insignificant effects on age at test taking. Thus—if anything—age effects run contrary to reported achievements gains from binding teacher recommendations, implying that strategic retention does not pose a major threat to identification.

4.5.2 Test Comparability

For the analyses on reading achievement, I combine data from various studies that assess students in Germany. The respective analyses produce meaningful results only if test scores are comparable across test and waves. By exploiting a specific feature of NAS 2012, I can shed some light on the comparability of PISA and NAS test scores. All PISA participants who attended ninth grade in 2012 were automatically sampled for NAS. For the sub-sample of students who participated in both tests, I can therefore merge additional information on student achievement as assessed by PISA (21 percent of the overall 2012 NAS sample)[35].

First, I compare the math achievement of students who participated in both studies in 2012. To the extent that the findings for math scores are informative for reading scores, I can explore whether the two tests produce similar results. I find a strong correlation (0.82) between the

[34] Between 2005 and 2010, several state governments (e.g., North Rhine-Westphalia, Baden-Wuerttemberg, Brandenburg, Berlin) changed the cutoff date for school enrollment. Students affected by the new cutoff dates are part of the fourth grade samples, but not yet of the ninth grade sample which also explains zero age effects for ninth graders.

[35] see Prenzel et al. (2015) for the PISA 2012 data

math scores from both studies. In addition, Appendix Figure A4.3 plots the distribution of within-student differences in test scores. Reassuringly, differences appear to be normally distributed around zero.

Second, the merged data set allows me to observe reading achievement for selected students in 2012, namely those who also participated in PISA. Appendix Table A4.9 reports reform effects on students' reading achievement, exploiting the merged data set. Reassuringly, the effects do not change substantially when the reading test scores of the 2012 PISA participants are included.

4.5.3 de Chaisemartin and D'Haultfoeuille Diagnostics

My empirical strategy falls into the category of two-way fixed effects differences-in-differences estimations whose estimates are a weighted sum of the average treatment effect in each state and school year. When the treatment effect is constant across states and over time, the depicted regressions estimate the effect of binding teacher recommendations under the standard *common trends* assumption. However, the weighted sum may contain negative weights which is a problem when the average treatment effects (ATEs) are heterogeneous across states or school years (see de Chaisemartin and D'Haultfœuille, 2020).

To evaluate the extent of negative weights in my setting, I perform the diagnostic test by de Chaisemartin and D'Haultfœuille (2020) on the simple model, which controls for state and school-year fixed effects. Appendix Table A4.10 shows the results on all student-level outcomes.[36] Overall, I obtain 22–84 ATTs with negative weights attached to 14–41 percent of them. Further exploring the causes of the negative weights, I conduct the same analyses without the 'always treated' students in the states Bavaria, Thuringia, and Saxony. Results show that in the fourth grade sample, 0 out of 10 ATTs receive a negative weight, in the ninth grade sample only 3 out of 35–36 ATTs (see Appendix Table A4.11). Reassuringly, after excluding the 'always treated' from the main specification, results remain robust (see Appendix Table A4.12). Although I loose some power due to decreased sample sizes, point estimates are similar in magnitude to those presented in section 4.4.

4.6 Conclusion

This chapter studies whether and how selective placement policies which are widely used to determine school track assignment, influence students' educational outcomes in the short and medium run. To that end, I exploit state-level reforms that changed the selectivity of

[36] Since the analyses on math achievement in fourth grade exploit changes in test scores measured in 2011 and 2016 (two time periods only), negative weights cannot bias the results in the math specification. Consequently, Table A4.10 is performed on the remaining outcomes.

admission policies, namely whether children require a recommendation from their primary school teacher to attend the academic schools (the highest track) or not.

In the short run, I find that binding teacher recommendations have a substantial impact on academic achievement of students in primary school *before* track assignment. Conditional on state and school year fixed effects and a rich set of sociodemographic controls, reading (math) achievement is 5.6 (12.2) percent of a standard deviation higher for students who require a respective teacher recommendation for academic school attendance than for students who have free choice of secondary schools. Further analyses show that these achievement gains likely reflect increased time investments in the students' skill development undertaken by children and their parents.

Subsequently, I show that binding teacher recommendations also affect students' educational outcomes in secondary schools. Under binding teacher recommendations, students are slightly less likely to attend academic schools in fifth grade (not statistically significant at conventional levels). However, these small and negative effects which are measured directly after tracking has taken place, veil important positive effects in the medium run. Notably, reform effects increase gradually throughout grades five to nine. In ninth grade, academic track attendance is even significant 1.8 percentage points higher under binding recommendations. Therefore, binding recommendations seem to reduce the incidence of students transferring to lower tracks during their secondary school career. I further show that ninth grade students perform slightly better in standardized reading tests (not significant) and have slightly better grade point averages ($p<0.1$).

The political debate in Germany around teacher recommendations has mostly revolved around the normative argument that broad groups in the population should be granted access to the academic schools. Consequently, the most recent reforms have abolished binding recommendations and guaranteed children and parents free choice of secondary schools. My findings, however, challenge this line of argumentation by providing evidence that free choice actually *reduces* academic school attendance in the medium run and can, therefore, harm students' academic performances.

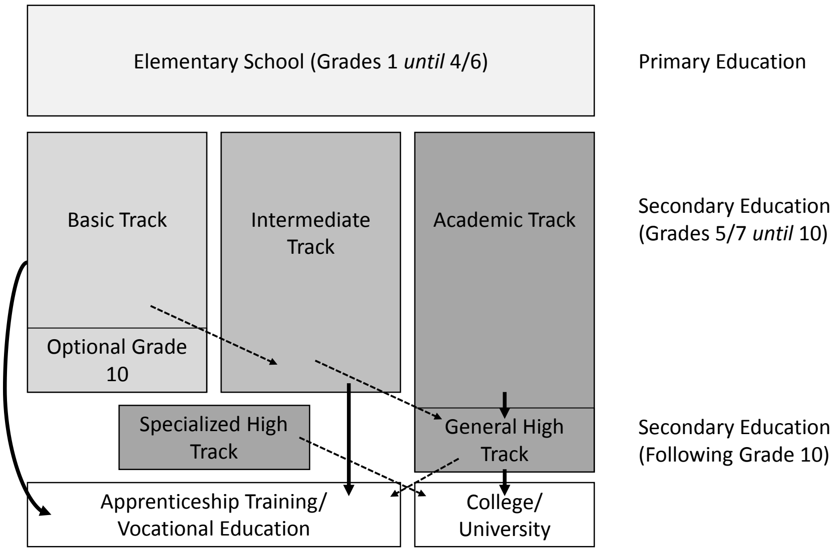
Even though the German early-tracking regime is somewhat special in its rigor, the results on selective placement polices bear broader implications. Many school systems around the world employ some sort of (selective) tracking: On average, fifteen-year-old students in the OECD have access to three education programs and 43 percent of them attend schools which consider 'students' records of academic performance' or 'recommendation of feeder schools' for admission (OECD, 2013). In this context, my chapter presents direct evidence that the design of placement policies plays an important role for the formation of human capital and proves worthwhile to be considered by policy-makers.

While the economic literature has mostly focused on the effects of earlier vs. later tracking so far (e.g., Meghir and Palme, 2005; Hanushek and Woessmann, 2006; Malamud and Pop-

Eleches, 2011; Pekkala Kerr et al., 2013; Piopiunik, 2014; Borghans et al., 2020; Canaan, 2020; Matthewes, 2020), this chapter argues that changes in the institutional feature of the tracking procedure can influence students' educational performances as well. The findings particularly highlight that changing the selectivity of placement policies do not only affect outcomes of students in the medium to long run, but they also induce behavioral changes among students, parents and teachers before track assignment takes place. Since the mentioned papers have mainly observed long-term educational outcomes and are thus limited in their ability to isolate short-term effects, investigating more systematically whether and how other aspects of the tracking procedure, e.g. the timing of tracking or the number of offered tracks, induce short-term responses is an interesting avenue for further research.

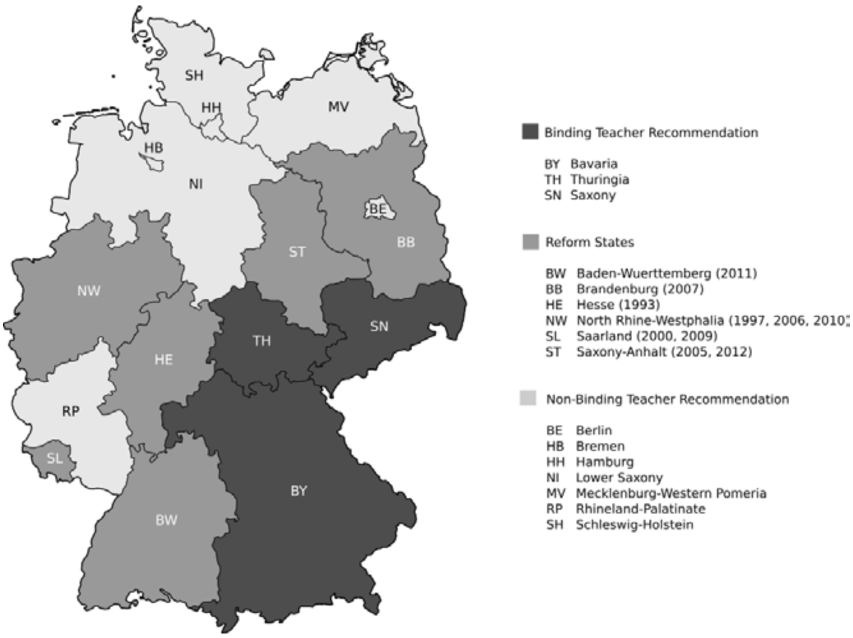
Figures and Tables

Figure 4.1: The German School System



Notes: The figure gives an overview of the school system in Germany. After elementary school which takes 4 years (only in a few states 6 years), students are tracked into three different school types: the basic and intermediate track last to grades 9 and 10, respectively, and prepare students for apprentice-ship training or other sorts of vocational education. The academic track ends with grade 13 (or 12) and leads to the university entrance qualification. Later track switching is possible, enabling graduates from the basic and intermediate track to continue on the next higher track, respectively, and/or obtaining their university entrance qualification via the specialized high track.

Figure 4.2: State Variation in the Bindingness of Teacher Recommendations



Notes: The figure gives an overview of the variation in the bindingness of teacher recommendations in Germany (for the period 1990 until 2017).

Table 4.1: Descriptive Statistics: Fourth Grade Students

Variable	Mean	Std. Dev.	Min.	Max.	Obs.
Reform Indicator					
Binding Recommendation	0.299	0.458	0	1	76,886
Outcome Variables					
Reading Score (Raw)	500.165	99.050	-36.91487	852.1359	70,615
Math Score (Raw)	481.125	108.163	-94.592	869.943	60,941
Student and School Characteristics					
Age (in Months)	124.262	8.021	73	177	73,217
Female	0.490	0.500	0	1	72,820
More than 100 Books at Home	0.598	0.490	0	1	66,175
Highest Occupational Status (ISEI) of Parents	50.389	18.054	10	90	53,335
First or Second Generation Migrant	0.255	0.436	0	1	64,675
Parents with University Degree	0.299	0.458	0	1	49,848
School located in Urban Area	0.839	0.367	0	1	71,554
Public School	0.968	0.177	0	1	72,341
Reform Controls					
Recommendation Only Based on Students' Grades	0.481	0.500	0	1	76,886
Four Years of Primary School	0.775	0.418	0	1	76,886
Comprehensive School Besides Academic School	0.587	0.492	0	1	76,886
Academic School Takes Eight Years	0.725	0.447	0	1	76,886
Economic Controls					
GDP per Capita (in 1000 Euros)	32.930	9.731	16.323	61.045	76,886
Average School Spending per Capita	1044	340	562	1934	76,886
Share of Students Enrolling late into Primary School	0.074	0.047	0.008	0.215	76,886
Number of Classes in Primary School	9042	8697	995	34,237	76,886
Average Class Size in Primary School	21.031	1.504	16.968	24.607	76,886
Average Number of Lessons in Primary School	31.017	3.469	24.241	43.148	76,886
Share of Full-Time Employed Primary School Teachers	0.491	0.138	0.060	0.726	76,886

Notes: Descriptive statistics (mean, standard deviation, minimum and maximum) for treatment, outcome and control variables. Data source: Progress in International Reading Literacy Study (IGLU-E) 2001 and 2006, National Assesment Study (NAS) 2011 and 2016.

Table 4.2: Descriptive Statistics: Ninth Grade Students

Variable	Mean	Std. Dev.	Min.	Max.	Obs.
Reform Indicator					
Binding Recommendation	0.338	0.473	0	1	212,706
Outcome Variables					
Academic School Attendance	0.328	0.469	0	1	212,706
Reading Score (Raw)	493.594	96.298	8.484	976.180	164,261
German Grade	3.039	0.886	1	6	201,905
Math Grade	3.17	1.045	1	6	201,387
Student Characteristics					
Age (in Months)	187.489	6.813	139	256	212,412
Female	0.493	0.500	0	1	212,124
More than 100 Books at Home	0.486	0.998	0	1	175,913
Highest Occupational Status (ISEI) of Parents	49.833	17.394	11.01	90	172,980
First or Second Generation Migrant	0.245	0.430	0	1	199,054
Parents with University Entrance Degree	0.414	0.492	0	1	169,687
Reform Controls					
Recommendation Only Based on Students’ Grades	0.488	0.500	0	1	212,706
Four Years of Primary School	0.782	0.413	0	1	212,706
Intensity of School Tracking	0.289	0.453	0	1	212,706
Academic School Takes Eight Years	0.429	0.495	0	1	212,706
Economic and Education Controls					
GDP per Capita (in 1000 Euros)	26.980	8.767	8.896	53.644	212,706
Share of Students Enrolling Late into Primary School	0.072	0.031	0.006	0.183	212,606
Number of Classes in Primary School	26.980	8.767	8.896	53.644	212,706
Average Class Size in Primary School	21.734	1.693	16.968	25.082	212,706
Average Number of Lessons in Primary School	8.431	2.901	22.782	41.099	212,370
Share of Full-Time Employed Primary School Teachers	0.450	0.157	0.060	0.894	212,706

Notes: Descriptive statistics (mean, standard deviation, minimum and maximum) for treatment, outcome and control variables. Data source: Program for International Student Assessment (PISA-E) 2000, 2003, and 2006 and National Assessment Study (NAS) 2009, 2012, and 2015.

Table 4.3: Reform Effects on Reading Achievement among Fourth Grade Students

	(1)	(2)	(3)	(4)
Panel A: Standardized Reading Achievement				
Binding Recommendation	0.064*	0.068*	0.100**	0.056**
	(0.031)	(0.035)	(0.041)	(0.026)
	[0.176]	[0.142]	[0.165]	[0.183]
Control Mean	-0.067	-0.067	-0.067	-0.067
Observations	70,615	70,615	70,615	70,615
R-squared	0.015	0.016	0.016	0.218
Panel B: Standardized Math Achievement				
Binding Recommendation	0.170***	0.145***	0.208***	0.122***
	(0.048)	(0.030)	(0.057)	(0.037)
	[0.201]	[0.062]	[0.186]	[0.154]
Control Mean	-0.084	-0.084	-0.084	-0.084
Observations	60,856	60,856	60,856	60,856
R-squared	0.024	0.028	0.029	0.193
State FEs	Yes	Yes	Yes	Yes
School Year FEs	Yes	Yes	Yes	Yes
Reform Controls	No	Yes	Yes	Yes
Economic Controls	No	No	Yes	Yes
Individual Controls	No	No	No	Yes

Notes: Differences-in-differences regressions weighted by students' sampling probability, including state and school year of transition fixed effects. Binding recommendation: teacher recommendation was binding in the school year of transition from primary to secondary school. Dependent variable: (Panel A) Standardized test scores in reading. (Panel B) Standardized test scores in math. Control mean: mean of the outcome variable for students not subject to binding teacher recommendations. Control variables: reform controls include basis for recommendation, whether recommendation needs to be requested explicitly, duration of primary schooling, intensity of tracking system, and duration of academic track school. Economic controls include GDP, average school spending, average share of students enrolling late into primary school, number of classes in primary school, average class size in primary school, average hours of instruction in primary school, and share of full-time employed primary school teachers. Individual controls include gender, migration background, parental occupation, number of books at home, community location, public school status, wave fixed effects and imputation dummies. Inference: standard errors clustered at the state level in parentheses. ***/**/* indicate significance at the 1% /5% /10% level. Square brackets additionally present p-values from wild cluster bootstrap by Roodman et al. (2019). Data source: Progress in International Reading Literacy Study (IGLU-E) 2001 and 2006, National Assesment Study (NAS) 2011 and 2016.

Table 4.4: Reform Effects on Time Invested into Students' Skill Development (at the Age of 10)

	(1)	(2)	(3)	(4)
Panel A: Child's Time Spent with Reading				
Binding Recommendation	0.052	0.048	0.130	0.174**
	(0.087)	(0.065)	(0.075)	(0.068)
	[0.789]	[0.817]	[0.527]	[0.371]
Observations	4,324	4,324	4,324	4,324
R-squared	0.016	0.016	0.017	0.058
Panel B: Child's Time Spent with Conducive Activities				
Binding Recommendation	0.087**	0.092*	0.090*	0.178***
	(0.032)	(0.046)	(0.047)	(0.053)
	[0.154]	[0.198]	[0.161]	[0.129]
Observations	4,215	4,215	4,215	4,215
R-squared	0.021	0.021	0.024	0.118
Panel C: Child's Time Spent with Detrimental Activities				
Binding Recommendation	0.011	0.001	-0.069	-0.085
	(0.156)	(0.150)	(0.178)	(0.165)
	[0.955]	[1.000]	[0.791]	[0.817]
Observations	4,148	4,148	4,148	4,148
R-squared	0.011	0.012	0.014	0.025
Panel D: Child's Time Spent with Family				
Binding Recommendation	0.305***	0.312***	0.370***	0.389***
	(0.054)	(0.060)	(0.059)	(0.057)
	[0.134]	[0.174]	[0.079]	[0.099]
Observations	4,318	4,318	4,318	4,318
R-squared	0.013	0.013	0.014	0.049
Panel E: Parents Frequently Consult Teacher				
Binding Recommendation	0.136**	0.134**	0.135***	0.130***
	(0.053)	(0.053)	(0.042)	(0.040)
	[0.123]	[0.149]	[0.150]	[0.141]
Control Mean	0.434	0.434	0.434	0.434
Observations	4,345	4,345	4,345	4,345
R-squared	0.027	0.027	0.029	0.041
State & School Year FEs	Yes	Yes	Yes	Yes
Reform Controls	No	Yes	Yes	Yes
Economic Controls	No	No	Yes	Yes
Individual Controls	No	No	No	Yes

Notes: Differences-in-differences regressions with state and school year of transition fixed effects. Binding recommendation: teacher recommendation was binding in the school year of transition from primary to secondary school. Dependent variables: time invested, the higher the value the more invested. (Panel A) Time spent with reading. (Panel B) Index summarizing child's time spent with making music, dancing or theatre, doing sports, doing technical work, or drawing, following Kling et al. (2007). (Panel C) Index summarizing child's time spent with watching TV, playing on the computer, surfing on the internet or listening to music, following Kling et al. (2007). (Panel D) Child's time spent together with family, standardized. (Panel E) Dummy variable (=1 if parent regularly meets teacher). Control mean: mean of the outcome variable for students not subject to binding teacher recommendations. Control variables: see Table 4.3 for included reform and economic controls. Individual controls include child's gender, child's age, respondents' gender, whether child is respondent's own child, migration background, parental educational, and imputation dummies. Inference: standard errors clustered at the state level in parentheses. ***/**/* indicate significance at the 1% /5% /10% level. Square brackets additionally present p-values from wild cluster bootstrap by Roodman et al. (2019). Data source: German Socioeconomic Panel (GSOEP), stacked mother-child questionnaires for 9–10 year old children.

Table 4.5: Reform Effects on Academic School Attendance

	Share of Students Attending Academic Schools				
	5th Grade (1)	6th Grade (2)	7th Grade (3)	8th Grade (4)	9th Grade (5)
Binding Recommendation	-0.010	-0.003	0.008	0.010*	0.012*
	(0.012)	(0.013)	(0.006)	(0.005)	(0.006)
	[0.470]	[0.859]	[0.227]	[0.152]	[0.144]
Control Mean	0.377	0.359	0.355	0.336	0.324
Observations	309	297	376	360	344
Observations (Federal States)	14	14	16	16	16
R-squared	0.774	0.739	0.855	0.867	0.889
State FEs	Yes	Yes	Yes	Yes	Yes
School Year FEs	Yes	Yes	Yes	Yes	Yes
Reform & Economic Controls	Yes	Yes	Yes	Yes	Yes

Notes: Differences-in-differences regressions with state and school year of transition fixed effects. Binding recommendation: teacher recommendation was binding in the school year of transition from primary to secondary school. Dependent variables: share of students attending academic schools. Control variables: see Table 4.3 for included reform and economic controls. Inference: standard errors clustered at the state level in parentheses. ***/**/* indicate significance at the 1% /5% /10% level. Square brackets additionally present p-values from wild cluster bootstrap by Roodman et al. (2019). Data source: German Statistical Office 1991-2016.

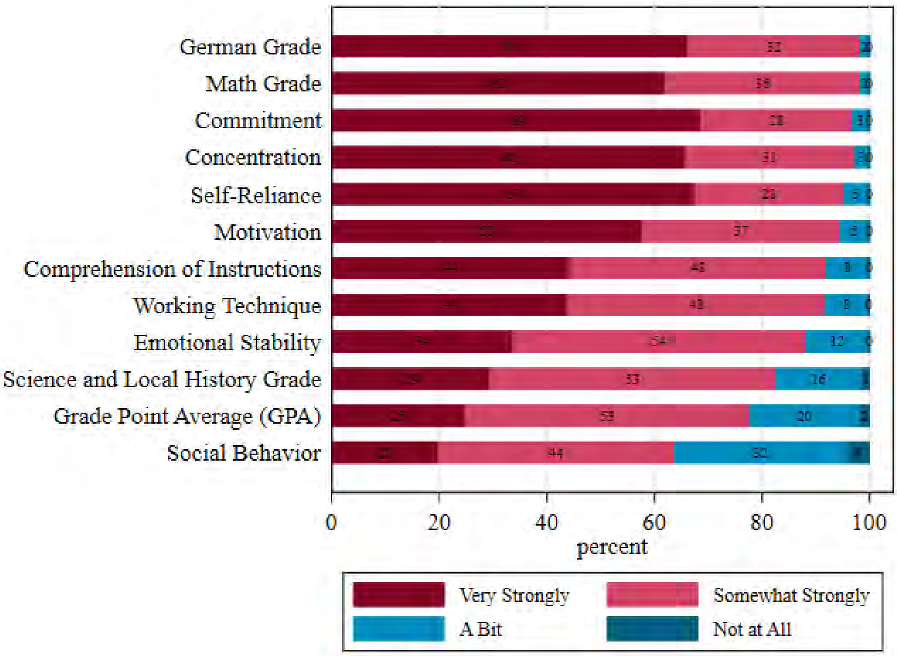
Table 4.6: Reform Effects on Academic School Attendance and Academic Performance among Ninth Grade Students

	(1)	(2)	(3)	(4)
Panel A: Academic School Attendance				
Binding Recommendation	0.017***	0.025***	0.023***	0.018**
	(0.004)	(0.005)	(0.008)	(0.006)
	[0.114]	[0.072]	[0.249]	[0.266]
Control Mean	0.335	0.335	0.335	0.335
Observations	208,405	208,405	207,969	207,969
R-squared	0.014	0.015	0.014	0.196
Panel B: Standardized Reading Achievement				
Binding Recommendation	0.041	0.055	0.046	0.019
	(0.033)	(0.043)	(0.036)	(0.025)
	[0.306]	[0.293]	[0.585]	[0.550]
Control Mean	0.009	0.009	0.009	0.009
Observations	163,346	163,346	162,940	162,940
R-squared	0.014	0.014	0.014	0.234
Panel C: Grade Point Average				
Binding Recommendation	0.041	0.049*	0.050*	0.038*
	(0.025)	(0.023)	(0.025)	(0.022)
	[0.450]	[0.239]	[0.251]	[0.210]
Control Mean	3.186	3.186	3.186	3.186
Observations	197,252	197,252	196,846	196,846
R-squared	0.046	0.047	0.047	0.091
State FEs	Yes	Yes	Yes	Yes
School Year FEs	Yes	Yes	Yes	Yes
Reform Controls	No	Yes	Yes	Yes
Economic Controls	No	No	Yes	Yes
Individual Controls	No	No	No	Yes

Notes: Differences-in-differences regressions weighted by students' sampling probability, including state and school year of transition fixed effects. Binding recommendation: teacher recommendation was binding in the school year of transition from primary to secondary school. Dependent variables. (Panel A) Academic school attendance. (Panel B) Standardized test score in reading. (Panel C) Standardized grade point average, the higher the value the better the GPA. Control mean: mean of the outcome variable for students not subject to binding teacher recommendations. Control variables: see Table 4.3 for included reform and economic controls. Individual controls include gender, migration background, parental occupation, number of books at home, wave fixed effects, and imputation dummies. Inference: standard errors clustered at the state level in parentheses. ***/**/* indicate significance at the 1% /5% /10% level. Square brackets additionally present p-values from wild cluster bootstrap by Roodman et al. (2019). Data source: Program for International Student Assessment (PISA-E) 2000, 2003, and 2006 and National Assesment Study (NAS) 2009, 2012, and 2015.

Appendix

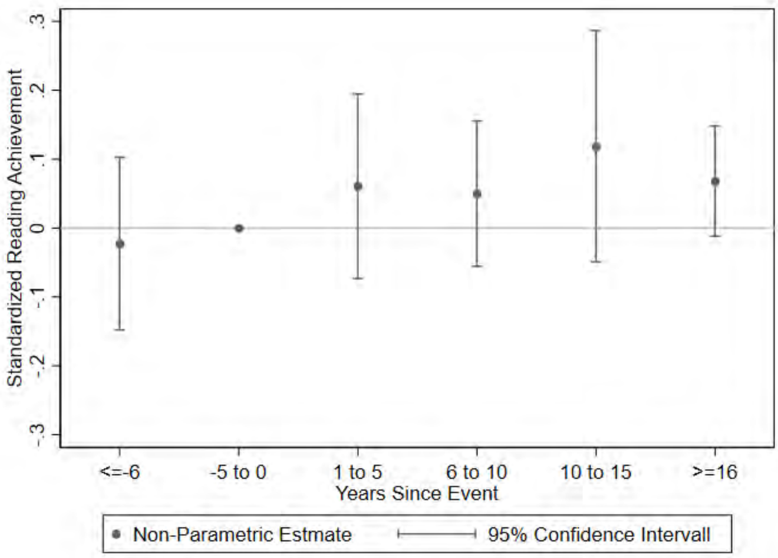
Figure A4.1: The Basis for Teacher Recommendations



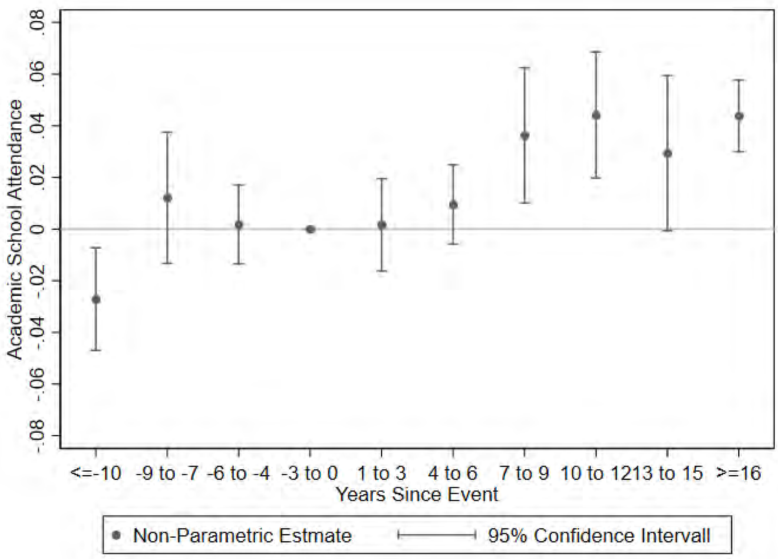
Notes: Response to the following question: *'Thinking about the different students: How strongly do the following factors determine your recommendation?'* Sample: fourth grade teachers in German. Weighted responses. Source: IGLU 2006.

Figure A4.2: Non-Parametric Event-Study Estimates on Students' Academic Performance

Panel A: Standardized Reading Achievement in Fourth Grade

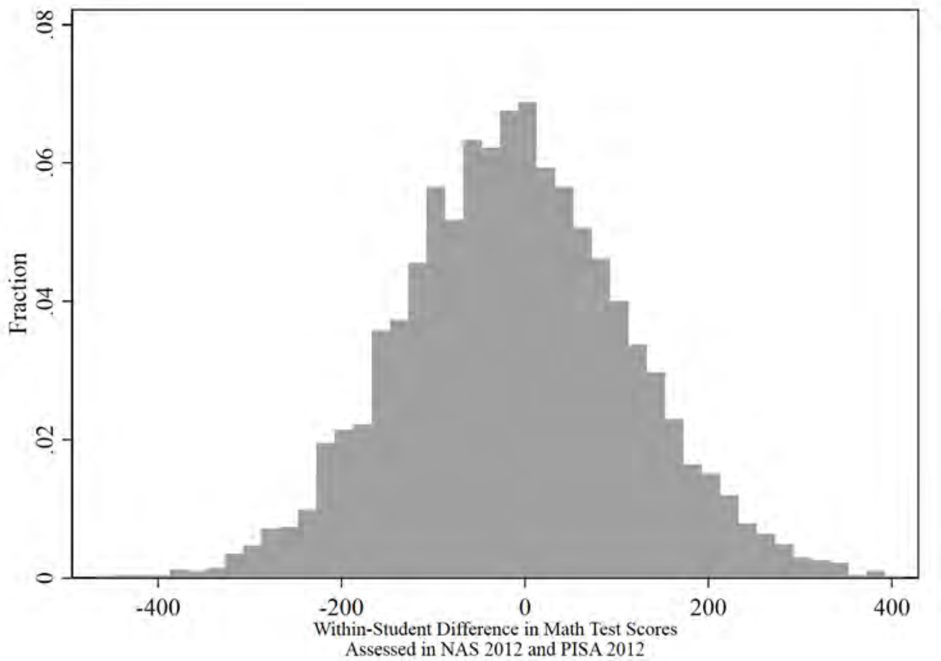


Panel B: Academic School Attendance in Ninth Grade



Notes: Coefficients from non-parametric event-study regressions and their 95 percent confidence intervals weighted by students' sampling probability, including state and school year of transition fixed effects. Dependent variables: (Panel A) Standardized reading achievement in fourth grade. (Panel B) Academic school attendance in ninth grade. Inference: standard clustering at state level. Data source: Progress in International Reading Literacy Study (IGLU-E) 2001 and 2006, National Assessment Study (NAS) 2011 and 2016. Program for International Student Assessment (PISA-E) 2000, 2003, and 2006 and National Assessment Study (NAS) 2009, 2012, and 2015.

Figure A4.3: Distribution of Math Test Scores in NAS and PISA



Notes: In each of the two tests, achievement is mapped on a scale with mean of 500 and a standard deviation of 100 test-score points. Outcome: within-student difference between math scores assessed by NAS 2012 and PISA 2012 ('0' indicates no difference in test scores). Sample: students who participated in NAS 2012 and PISA 2012.

Table A4.1: Reforms on the Bindingness of Teacher Recommendations and Ruling Parties, by State

State	Year of Reform	Reform Type: Recommendat. Change to	Ruling Parties in Legislation Period Before the Reform	Ruling Parties in Legislation Period of the Reform
Hesse	1993	non-binding	CDU/FDP (1987-1991)	SPD/Gruene (1991-1995)
North Rhine-Westphalia	1997	non-binding	SPD /Gruene (1990-1995)	SPD/Gruene (1995-2000)
Saarland	2000	binding	SPD (1994-1999)	CDU (1999-2004)
Saxony-Anhalt	2005	binding	SPD (1998-2002)	CDU/FDP (2002-2006)
North Rhine-Westphalia	2006	binding	SPD/Gruene (2000-2005)	CDU/FDP (2005-2010)
Brandenburg	2007	binding	SPD/CDU (1999-2004)	SPD/CDU (2004-2009)
Saarland	2009	non-binding	CDU (2004-2009)	CDU/FDP/SPD/Gruene (2009-2012)
North Rhine-Westphalia	2010	non-binding	CDU/FDP (2005-2010)	SPD/Gruene (2010-2012)
Baden-Wuerttemberg	2011	non-binding	CDU/FDP/DVP (2006-2011)	Gruene/SPD (2011-2016)
Saxony-Anhalt	2012	non-binding	CDU/SPD (2006-2011)	CDU/SPD (2011-2016)

Table A4.2: Reform Effects on Issued Recommendations for Academic Schools

	Share of Students with Recommendation for Academic Schools		
	(1)	(2)	(3)
Binding Recommendation	0.060	0.064	0.051
	(0.040)	(0.050)	(0.044)
	[0.287]	[0.453]	[0.348]
Control Mean	0.374	0.374	0.374
Observations	108	108	108
Observations (Federal States)	8	8	8
R-squared	0.741	0.741	0.813
State FEs	Yes	Yes	Yes
School Year FEs	Yes	Yes	Yes
Reform Controls	No	Yes	Yes
Economic Controls	No	No	Yes

Notes: Differences-in differences regressions with state and school year of transition fixed effects. Binding recommendation: teacher recommendation was binding in the school year of transition from primary to secondary school. Dependent variable: recommendations issued for the academic school as share of all recommendations issued. Control variables: see Table 4.3 for included reform and economic controls. Inference: standard errors clustered at the state level in parentheses. ***/**/* indicate significance at the 1% /5% /10% level. Square brackets additionally present p-values from wild cluster bootstrap by Roodman et al. (2019). Sample: Baden-Wuerttemberg, Bavaria, Berlin, Mecklenburg Western Pomerania, North Rhine Westphalia, Rhineland Palatinate, Saxony. Data source: various State Ministries of Education, Germany.

Table A4.3: Heterogeneous Reform Effects by Socioeconomic Background

	Fourth Grade Students		Ninth Grade Students		
	Read. Achiev. (1)	Math Achiev. (2)	Acad. School (3)	Read. Achiev. (4)	GPA (5)
Panel A: Books at Home					
Binding Recommendation	-0.013	0.081	0.011	-0.023	0.027
	(0.037)	(0.055)	(0.006)	(0.027)	(0.027)
Bind. Recomm. x Less than 100 Books	0.065**	0.103***	0.002	-0.023	0.036
	(0.023)	(0.030)	(0.007)	(0.027)	(0.039)
Observations	63,772	55,083	173,969	129,259	165,439
Panel B: Highest Occupational Status (ISEI) of Parents					
Binding Recommendation	0.026	0.083	0.032***	0.023	0.009
	(0.032)	(0.055)	(0.006)	(0.024)	(0.021)
Bind. Recomm. x ISEI below Median	0.009	0.047	-0.012	-0.006	0.057
	(0.028)	(0.032)	(0.008)	(0.024)	(0.045)
Observations	51,654	44,149	171,014	145,717	163,288
Panel C: Parental Education					
Binding Recommendation	0.018	0.146**	0.045***	-0.005	-0.004
	(0.035)	(0.063)	(0.011)	(0.033)	(0.034)
Bind. Recomm. x Parens w\o Uni Degree	0.014	0.017	-0.032*	0.029	0.056
	(0.023)	(0.032)	(0.016)	(0.019)	(0.052)
Observations	48,372	41,409	168,188	137,097	160,219
Panel D: Migration Status (First or Second Generation Migrant)					
Binding Recommendation	0.028	0.082**	0.021**	0.031	0.043*
	(0.028)	(0.031)	(0.007)	(0.025)	(0.023)
Bind. Recomm. x Migrant	-0.008	0.019	-0.004	-0.014	-0.051**
	(0.044)	(0.031)	(0.013)	(0.029)	(0.019)
Observations	62,352	53,868	197,065	152,354	187,452
State FEs	Yes	Yes	Yes	Yes	Yes
School Year FEs	Yes	Yes	Yes	Yes	Yes
Reform & Economic & Individual Controls	Yes	Yes	Yes	Yes	Yes

Notes: Differences-in-differences regressions weighted by students' sampling probability, including state and school year of transition fixed effects. Binding recommendation: teacher recommendation was binding in the school year of transition from primary to secondary school. Dependent variables. (1) Standardized test scores in reading, fourth grade. (2) Academic school attendance, ninth grade. (3) Standardized test scores in reading, ninth grade. (4) Standardized grade point average, the higher the value the better the GPA, ninth grade. Control variables: see Tables 4.3 and 4.6 for included reform, education and individual controls. Inference: standard errors clustered at the state level in parentheses. ***/**/* indicate significance at the 1% /5% /10% level. Square brackets additionally present p-values from wild cluster bootstrap by Roodman et al. (2019). Data source: Progress in International Reading Literacy Study (IGLU-E) 2001 and 2006, National Assessment Study (NAS) 2011 and 2016. Program for International Student Assessment (PISA-E) 2000, 2003, and 2006 and National Assessment Study (NAS) 2009, 2012, and 2015.

Table A4.4: Academic Performance and Missing Information on Socioeconomic Background

	Fourth Grade Students		Ninth Grade Students		
	Read. Achiev. (1)	Math Achiev. (2)	Acad. School (3)	Read. Achiev. (4)	GPA (5)
Panel A: Books at Home					
Missing: Books at home	-0.702***	-0.515***	-0.097***	-0.405***	-0.054***
	(0.028)	(0.023)	(0.006)	(0.014)	(0.013)
Obseravtions	70,615	60,202	213,613	168,554	201,689
Panel B: Highest ISEI					
Missing: HISEI	-0.476***	-0.444***	-0.187***	-0.663***	-0.264***
	(0.016)	(0.015)	(0.004)	(0.013)	(0.009)
Obseravtions	70,615	60,202	213,613	168,554	201,689
Panel C: Parental Education					
Missing: Parental Education	-0.352***	-0.330***	-0.140***	-0.456***	-0.188***
	(0.015)	(0.014)	(0.004)	(0.011)	(0.009)
Obseravtions	70,615	60,202	213,613	168,554	201,689
Panel D: Migration Status					
Missing: Migration Status	-0.634***	-0.500***	-0.175***	-0.653***	-0.246***
	(0.024)	(0.022)	(0.005)	(0.017)	(0.015)
Obseravtions	70,615	60,202	213,613	168,554	201,689
Study FEs	Yes	Yes	Yes	Yes	Yes

Notes: Ordinary least square regressions weighted by students' sampling probability, including wave fixed effects. Dependent variables. (1) Standardized test scores in reading, fourth grade. (2) Standardized test scores in math, fourth grade. (3) Academic school attendance, ninth grade. (4) Standardized test scores in reading, ninth grade. (5) Standardized grade point average, the higher the value the better the GPA, ninth grade. Inference: robust standard errors in parentheses. ***/**/* indicate significance at the 1% /5% /10% level. Data source: Progress in International Reading Literacy Study (IGLU-E) 2001 and 2006, National Assessment Study (NAS) 2011 and 2016. Program for International Student Assessment (PISA-E) 2000, 2003, and 2006 and National Assessment Study (NAS) 2009, 2012, and 2015.

Table A4.5: Effects of Government Ideology on Students' Academic Performance in Fourth Grade

	(1)	(2)	(3)	(4)
Panel A: Standardized Reading Achievement				
Conservative Prime Minister	0.058	0.021		
	(0.036)	(0.030)		
	[0.188]	[0.567]		
Conservative Education Minister			0.058*	0.014
			0.058*	0.014
			[0.060]	[0.676]
Observations	70,615	70,615	70,615	70,615
R-squared	0.015	0.218	0.015	0.218
Panel B: Standardized Math Achievement				
Conservative Prime Minister	0.133**	-0.015		
	(0.061)	(0.042)		
	[0.164]	[0.761]		
Conservative Education Minister			0.132**	0.002
			(0.055)	(0.038)
			[0.085]	[0.967]
Observations	60,856	60,856	60,856	60,856
R-squared	0.024	0.192	0.024	0.192
State FEs	Yes	Yes	Yes	Yes
School Year FEs	Yes	Yes	Yes	Yes
Reform & Economic & Individual Controls	No	Yes	No	Yes

Notes: Differences-in-differences regressions weighted by students' sampling probability, including state and school year of transition fixed effects. Dependent variables: (Panel A) Standardized test scores in reading. (Panel B) Standardized test scores in math. Control variables: see Table 4.3 for included reform, education and individual controls. Inference: Standard errors clustered at the state level in parentheses. ***/**/* indicate significance at the 1% /5% /10% level. Data source: Progress in International Reading Literacy Study (IGLU-E) 2001 and 2006, National Assessment Study (NAS) 2011 and 2016.

Table A4.6: Effects of Government Ideology on Students' Academic Performance in Ninth Grade

	(1)	(2)	(3)	(4)
Panel A: Actual Academic School Attendance				
Conservative Prime Minister	0.006	0.008		
	(0.008)	(0.005)		
	[0.479]	[0.226]		
Conservative Education Minister			0.003	0.002
			(0.006)	(0.006)
			[0.724]	[0.766]
Observations	208,405	207,969	208,405	207,969
R-squared	0.014	0.196	0.014	0.196
Panel B: Standardized Reading Achievement				
Conservative Prime Minister	0.058**	0.028		
	(0.027)	(0.039)		
	[0.108]	[0.611]		
Conservative Education Minister			0.033	-0.005
			(0.029)	(0.036)
			[0.371]	[0.922]
Observations	163,346	162,940	163,346	162,940
R-squared	0.014	0.234	0.014	0.234
Panel C: Grade Point Average				
Conservative Prime Minister	0.062***	0.043		
	(0.018)	(0.027)		
	[0.030]	[0.368]		
Conservative Education Minister			0.069***	0.045
			(0.015)	(0.029)
			[0.005]	[0.348]
Observations	197,252	196,846	197,252	196,846
R-squared	0.047	0.091	0.047	0.091
State FEs	Yes	Yes	Yes	Yes
School Year FEs	Yes	Yes	Yes	Yes
Reform & Economic & Individual Controls	No	Yes	No	Yes

Notes: Differences-in-differences regressions weighted by students' sampling probability, including state and school year of transition fixed effects. Dependent variables: (Panel A) Academic school attendance. (Panel B) Standardized test scores in reading. (Panel C) Standardized grade point average, the higher the value the better the GPA. Control variables: see Table 4.6 for included reform, education and individual controls. Inference: Standard errors clustered at the state level in parentheses. ***/**/* indicate significance at the 1% /5% /10% level. Data source: PProgram for International Student Assessment (PISA-E) 2000, 2003, and 2006 and National Assessment Study (NAS) 2009, 2012, and 2015.

Table A4.7: Reform Effects on Students' Academic Performance, Controlling for Government Ideology

	(1)	(2)
Panel A: Standardized Reading Achievement (Fourth Grade)		
Binding Recommendation	0.054*	0.047
	(0.028)	(0.029)
	[0.234]	[0.245]
Observations	70,615	70,615
R-squared	0.218	0.218
Panel B: Standardized Math Achievement (Fourth Grade)		
Binding Recommendation	0.152***	0.174**
	(0.050)	(0.066)
	[0.203]	[0.182]
Observations	60,856	60,856
R-squared	0.193	0.193
Panel C: Actual Academic School Attendance (Ninth Grade)		
Binding Recommendation	0.020**	0.015*
	(0.007)	(0.007)
	[0.252]	[0.131]
Observations	207,969	207,969
R-squared	0.196	0.197
Panel D: Reading Achievement (Ninth Grade)		
Binding Recommendation	0.016	0.029
	(0.027)	(0.028)
	[0.616]	[0.419]
Observations	162,940	162,940
R-squared	0.234	0.234
Panel E: Grade Point Average (Ninth Grade)		
Binding Recommendation	0.023	0.023
	(0.020)	(0.023)
	[0.391]	[0.484]
Observations	196,846	196,846
R-squared	0.091	0.091
State & School Year FEs	Yes	Yes
Reform & Economic & Individual Controls	Yes	Yes
Conservative Prime Minister	Yes	No
Conservative Education Minister	No	Yes

Notes: Differences-in-differences regressions weighted by students' sampling probability, including state and school year of transition fixed effects. Binding recommendation: teacher recommendation was binding in the school year of transition from primary to secondary school. Dependent variables. (Panel A) Standardized test scores in reading, fourth grade. Panel B) Standardized test scores in math, fourth grade. (Panel C) Academic school attendance, ninth grade. (Panel D) Standardized test scores in reading, ninth grade. (Panel E) Standardized grade point average. Control variables: see Tables 4.3 and 4.6 for included reform, education and individual controls. Inference: Standard errors clustered at the state level in parentheses. ***/**/* indicate significance at the 1% /5% /10% level. Data source: Progress in International Reading Literacy Study (IGLU-E) 2001 and 2006, National Assessment Study (NAS) 2011 and 2016. Program for International Student Assessment (PISA-E) 2000, 2003, and 2006 and National Assessment Study (NAS) 2009, 2012, and 2015.

Table A4.8: Reform Effects on Students' Age

	Age in Months			
	(1)	(2)	(3)	(4)
Panel A: Fourth Grade Students				
Binding Recommendation	-0.352	-0.531	-0.779**	-0.765**
	(0.395)	(0.394)	(0.342)	(0.320)
	[0.489]	[0.480]	[0.139]	[0.159]
Control Mean	124.833	124.833	124.833	124.833
Observations	68,135	68,135	68,135	68,135
R-squared	0.102	0.102	0.104	0.148
Panel B: Ninth Grade Students				
Binding Recommendation	0.188	0.159	0.155	0.307
	(0.182)	(0.174)	(0.279)	(0.278)
	[0.220]	[0.221]	[0.219]	[0.268]
Control Mean	187.680	187.680	187.680	187.680
Observations	208,247	208,247	207,812	207,812
R-squared	0.220	0.221	0.219	0.268
State FEs	Yes	Yes	Yes	Yes
School Year FEs	Yes	Yes	Yes	Yes
Reform Controls	No	Yes	Yes	Yes
Economic Controls	No	No	Yes	Yes
Individual Controls	No	No	No	Yes

Notes: Differences-in-differences regressions weighted by students' sampling probability, including state and school year of transition fixed effects. Binding recommendation: teacher recommendation was binding in the school year of transition from primary to secondary school. Dependent variables: age in months. Control variables: see Tables 4.3 and 4.6 for included reform, education and individual controls. Inference: standard errors clustered at the state level in parentheses. ***/**/* indicate significance at the 1% /5% /10% level. Square brackets additionally present p-values from wild cluster bootstrap by Roodman et al. (2019). Data source: Progress in International Reading Literacy Study (IGLU-E) 2001 and 2006, National Assessment Study (NAS) 2011 and 2016. Program for International Student Assessment (PISA-E) 2000, 2003, and 2006 and National Assessment Study (NAS) 2009, 2012, and 2015.

Table A4.9: Reform Effects on Ninth Grade Reading Achievemnet (Including PISA 2012 Scores)

	Reading Achievement in Ninth Grade			
	(1)	(2)	(3)	(4)
Binding Recommendation	0.041	0.068	0.027	0.020
	(0.037)	(0.041)	(0.036)	(0.027)
	[0.292]	[0.173]	[0.553]	[0.526]
Control Mean	-0.055	-0.055	-0.055	-0.055
Observations	172,686	172,686	172,279	172,279
R-squared	0.016	0.017	0.017	0.234
State FEs	Yes	Yes	Yes	Yes
School Year FEs	Yes	Yes	Yes	Yes
Reform Controls	No	Yes	Yes	Yes
Economic Controls	No	No	Yes	Yes
Individual Controls	No	No	No	Yes

Notes: Differences-in-differences regressions weighted by students' sampling probability, including state and school year of transition fixed effects. Binding recommendation: teacher recommendation was binding in the school year of transition from primary to secondary school. Dependent variables: standardized test scores in reading, ninth grade. Control variables: see Table 4.6 for included reform, education and indivudal controls. Inference: standard errors clustered at the state level in parentheses. ***/**/* indicate significance at the 1% /5% /10% level. Data source: Program for International Student Assessment (PISA-E) 2000, 2003, 2006 and 2012 and National Assessment Study (NAS) 2009 and 2015.

Table A4.10: de Chaisemartin and d'Haultfoeuille (2020): Differences-in-Differences Diagnostics

	# ATTs	# of ATTs with Negative Weight	Sum of Negative Weights
	(1)	(2)	(3)
Panel A: Reading Achievement (Fourth Grade)			
	22	9	-0.218
Panel B: Actual Academic School Attendance (Ninth Grade)			
	84	12	-0.153
Panel C: Reading Achievement (Ninth Grade)			
	65	16	-0.109
Panel D: Grade Point Average (Ninth Grade)			
	84	12	-0.160

Notes: Results from de Chaisemartin and D'Haultfœuille (2020) diagnostics test implemented using STATA *twowayfeweights* command. Estimated weights of all group-period clusters in the basic model, controlling for state and school-year fixed effects.

Table A4.11: de Chaisemartin and d'Haultfoeuille (2020): Differences-in-Differences Diagnostics without 'Always-Treated'

	# ATTs	# of ATTs with negative weight	Sum of negative weights
	(1)	(2)	(3)
Panel A: Reading Achievement (Fourth Grade)			
	10	0	–
Panel B: Actual Academic School Attendance (Ninth Grade)			
	36	3	-0.050
Panel C: Reading Achievement (Ninth Grade)			
	35	3	-0.064
Panel D: Grade Point Average (Ninth Grade)			
	36	3	-0.053

Notes: Results from de Chaisemartin and D'Haultfœuille (2020) diagnostics test implemented using STATA *twowayfeweights* command. Estimated weights of all group-period clusters in the simple model, controlling for state and school-year fixed effects. Sample, excluding the states of Bavaria, Thuringia and Saxony.

Table A4.12: Reform Effects on Students' Academic Performance without 'Always-Treated'

	Fourth Grade Students	Ninth Grade Students		
	Read. Achievement	Acad. School	Achievem.	GPA
	(1)	(2)	(3)	(4)
Binding Recommendation	0.056	0.015*	0.045**	0.025
	(0.033)	(0.007)	(0.016)	(0.026)
	[0.407]	[0.205]	[0.015]	[0.497]
Observations	58,206	171,036	133,760	161,346
R-squared	0.224	0.198	0.234	0.095
State Fes	Yes	Yes	Yes	Yes
School Year FEs	Yes	Yes	Yes	Yes
Reform & Economic & Indivdual Controls	Yes	Yes	Yes	Yes

Notes: Differences-in-differences regressions weighted by students' sampling probability, including state and school year of transition fixed effects. Binding recommendation: teacher recommendation was binding in the school year of transition from primary to secondary school. Dependent variables. (1) Standardized test scores in reading, fourth grade. (2) Academic school attendance, ninth grade. (3) Standardized test scores in reading, ninth grade. (4) Standardized grade point average, the higher the value the better the GPA, ninth grade. Control variables: see Tables 4.3 and 4.6 for included reform, education and individual controls. Inference: standard errors clustered at the state level in parentheses. ***/**/* indicate significance at the 1% /5% /10% level. Square brackets additionally present p-values from wild cluster bootstrap by Roodman et al. (2019). Data source: Progress in International Reading Literacy Study (IGLU-E) 2001 and 2006, National Assessment Study (NAS) 2011 and 2016. Program for International Student Assessment (PISA-E) 2000, 2003, and 2006 and National Assessment Study (NAS) 2009, 2012, and 2015—excluding the states of Bavaria, Thuringia and Saxony.

5 Do Party Positions Affect the Public's Policy Preferences? Experimental Evidence on Support for Family Policies*

5.1 Introduction

A standard assumption in the political economy literature is that the electorate's policy preferences are exogenous to the political process (e.g., Downs, 1957; Alesina, 1988; Besley and Coate, 1997; Lee et al., 2004). Consistently, political parties have been shown to choose policy positions that cater to their voters' preferences in order to win elections (e.g., Levitt, 1996; Strömberg, 2004; Fujiwara, 2015). Conversely, theories of endogenous preferences suggest that institutions can shape public preferences (e.g., Bowles, 1998). To test the empirical relevance of this reverse relationship, this chapter studies whether and how party positions shape public preferences toward family policies. Family policies are particularly important when studying endogeneity of policy preferences from an economics viewpoint, as they touch on a fundamental topic in political economy, namely the extent to which governments intervene in the private (family) sphere. We thus focus on two specific family policies recently disputed in Germany: the child care subsidy *(Betreuungsgeld)* and universal student aid *(Elternunabhängiges BAfoeG)*. A better understanding of the feedback between party positions and public policy preferences is relevant for policy as well as for modelling political behavior.

Empirical analysis of the causal effect of party positions on public preferences is challenging because natural experiments that induce exogenous variation in party positions are rare. Therefore, we conduct two survey experiments in a representative sample of the German voting-age population that includes an oversample of parents with school-aged children, an important special-interest group for family policies (N>4,000). For each of the two considered policies, we inform a randomly selected treatment group about the positions of the six main parties before eliciting respondents' preferences towards the respective policy. The control group answers the same policy question without receiving information on party positions. If the electorates' policy preferences were exogenously given, the experimental treatments should not affect respondents' stated policy preferences. If, however, political preferences for family policies are in fact endogenous to party positions, partisans' preferences should align more closely with their preferred party's position in the treatment group.

* This chapter is joint work with Philipp Lergetporer, Katharina Werner, and Ludger Woessmann. It is based on the paper 'Do Party Positions Affect the Public's Policy Preferences? Experimental Evidence on Support for Family Policies', *Journal of Economic Behavior and Organization*, 2020.

Our first analysis investigates preferences towards the child care subsidy. This controversially debated subsidy entitled parents who do not send their young children to publicly subsidized childcare facilities to receive a monthly payment of 150 Euro. In the control group, a majority of the German population opposes the child care subsidy (34 percent in favor, 56 percent opposed, 10 percent neither favor nor opposed). On average, providing information about current party positions does not significantly change public preferences towards the policy, which is not surprising because the treatment informs some partisans that their preferred party supports the policy and others that their preferred party opposes it.

Importantly, treatment effects vary strongly by individual partisanship. In the uninformed control group, a majority of supporters of the party which favors the child care subsidy actually *opposes* it (32 percent in favor, 58 percent opposed), which implies that party positions and uninformed partisans' preferences are largely misaligned on this policy. When provided with information about current party positions, these partisans are significantly more likely to support the policy (43 percent in favor, 47 percent opposed). Thus, the information treatment shifts these partisans' policy support and opposition by over 10 percentage points, overturning their majority against the child care subsidy by shifting their preferences towards their favored party's position. Analysis of the intensity of policy preferences indicates that the treatment shifts both weakly and strongly held preferences. We do not find significant treatment effects among partisans of other parties and non-partisans on this policy.

In our second analysis, we show that party positions also matter for preferences towards a proposal to reform federal student aid, a topic that was less debated in public. According to the proposal, the status quo of needs-based student aid should be replaced by universal student aid for all students. Overall, the German public is divided on the proposal of universal student aid: in the control group, 48 percent are in favor and 45 percent opposed.

Again, while informing about current party positions does not change average public preferences towards universal student aid, there are effect heterogeneities with respect to the respondents' party affiliation. Among partisans of parties that oppose universal student aid, a slight majority in the uninformed control group *supports* the policy (51 percent in favor, 43 percent opposed). This again reveals fundamental misalignment between party positions and uninformed partisans' preferences. The information treatment significantly ($p<0.1$) shifts these partisans' preferences towards their parties' position, with 43 percent of them favoring the policy and 48 percent opposing it. In contrast, there is no significant treatment effect for respondents attached to parties that favor universal student aid.

Subgroup analyses reveal that providing information about party positions does not only affect preferences of partisans, but also of the highly relevant subgroup of non-partisan swing voters. Specifically, swing voters—defined as respondents who regularly turn out to vote but do not identify as long-term supporters of any specific party—are more likely to support universal student aid in the treatment group than in the control group. We also investigate the prevalence of treatment effects in different sociodemographic subgroups of the affected partisan groups. While often shy of statistical power, this explorative analysis suggests, for instance, that parents show a particularly strong reaction to party-position information on universal student aid.

Overall, our results show that the electorate's preferences for family policies can be endogenous to party positions. Notably, these endogeneities seem not to be driven by the salience of the specific policy. While the child care subsidy was the subject of heated public and political debates, universal student aid received much less attention, possibly because the former but not the latter was actually implemented. Thus, the endogeneity of policy preferences is prevalent both for a relatively salient policy and for a policy that entered the public discourse less prominently. While our experiments on two policy initiatives do not allow us to draw general conclusions about the conditions under which party-position information does or does not affect partisans' preferences, it is interesting to note that treatment effects are apparent when the group of parties supporting or opposing the respective policy is rather homogeneous, but not when it includes parties from across the political spectrum. Party-position information may thus fail to align partisans with their preferred party's position if they learn that parties from the other side of the political spectrum take the same stance as their preferred party.

While differentiating the possible mechanisms that underlie our general finding is beyond the scope of our experimental design, two possible interpretations are the use of heuristics and the priming of identity. Heuristics are a common explanation for endogenous voter preferences in the political science literature. Individuals who are largely uninformed about specific political issues (Bartels, 1996; Lau and Redlawsk, 2001) may rely on simple cues when forming opinions in order to minimize the cognitive costs of their preference formation process (Tversky and Kahneman, 1974). If individuals generally vote for a party because they feel that it represents their preferences relatively well, they might also expect that this party's position will align with their values if they encounter a new policy field for which they did not yet form strong opinions. Thus, party cues can have instrumental value by helping ignorant respondents to form policy preferences with little cognitive effort that are reasonably close to the preferences they would hold after gathering and processing all relevant information about the policies.[1] While such use of heuristics is one possible explanation for why respondents align their preferences with party positions, some aspects of our results are hard to reconcile with this interpretation—e.g., the presence of treatment effects on policies with high and low salience, effects in directly involved subgroups like parents, and effects on strong as well as weakly held preferences. A complementary, but distinct channel is the one of partisan-identity priming. Since party identification can be a form of social identity (e.g., Green et al., 2002), respondents may derive direct utility from aligning their policy preferences with their in-group, i.e., their preferred party (e.g., Akerlof and Kranton, 2000).[2] Note that, while both potential treatment-effect

[1] As in Gentzkow and Shapiro (2006) model of media bias, party reputation may also play a role in our context since respondents may assign higher reputation to parties that they agreed with in the past, and hence perceive these parties' positions as superior in terms of competency.

[2] Since our treatment informs about the entire party spectrum, this channel could operate by providing in-group cues (i.e., respondents align their preferences with their preferred party's position in order to reinforce their

mechanisms rely on some form of imperfect information, imperfect information *about the policies and about party positions* is necessary for the heuristics channel to operate, but not for the partisan-identity channel, which only requires imperfect information *about party positions*.

Our study contributes to several strands of existing research. While there is surprisingly little economic literature investigating the extent to which voter preferences are endogenous to political party positions, there is a longer research tradition in political science studying the relationship between the political system and the views of the public. Following up on Campbell (1980)'s idea that party positions can change how people think about policy issues, non-experimental studies have looked at how political values and attitudes of the public relate to elite discourse and polarization (e.g., Bartels, 2002; Zaller, 2004; Gabel and Scheve, 2007), knowledge of party positions (Lenz, 2009, 2012), and changes in parties' views (Slothuus, 2010). Experimental studies have investigated, for instance, the effects of partisan stereotypes of political candidates (e.g., Rahn, 1993), lobbies' vote recommendations (e.g., Lupia, 1994), and elite communication (e.g., Broockman and Butler, 2017) on political perceptions and preferences. Further experimental studies have focused on the effects of different kinds of information about party positions on various outcomes including political values (e.g., Goren et al., 2009), 'obscure' left-wing or right-wing policies (e.g., Samuels and Zucco, 2014), attitudes towards the European Union (e.g., Pannico, 2017), and perceptions of the public budget (e.g., Bisgaard and Slothuus, 2018).

The chapter adds to the experimental literature on effects of party positions on voter preferences in three important ways. First, we are the first to study family policies, a policy domain that gets to the core of government intervention in the private sphere and thus to the heart of the extent of governmental reach as studied by the political economy of public economics. By showing that voters' preferences related to the outreach of government are endogenous to party positions, this perspective contributes to the political-economy literature on the role and size of government and government spending more generally (e.g., Persson and Tabellini, 2013). Second, the fact that we administer the information treatments to all respondents—independent of their partisanship or voting behavior—enables us to analyze effects not only on partisans, but also on swing voters. Swing voters are a highly relevant and often pivotal group in the political process (e.g., Zaller, 2004), but their preference-formation processes are generally not well understood (e.g., Ryan, 2017). We contribute to the research on swing voters by showing that their preferences are indeed responsive to party-position information.[3] Third, whereas most existing studies focus on two-party systems prevalent in countries such as the United States and the United Kingdom, we extend the small evidence base on party-position information in mature multi-party settings. Relatedly, we provide information on *all* parties

in-group membership) or out-group cues (i.e., respondents align their preferences with their preferred party's position in order to maximize the distance to out-group parties with other positions).

[3] In a Brazilian setting, Samuels and Zucco (2014) do not find effects of party-position information on non-partisans' policy preferences.

currently constituting the German parliament, which sets us apart from many studies that provide partial information on a subset of parties' positions (e.g., Goren et al., 2009; Samuels and Zucco, 2014; Bisgaard and Slothuus, 2018).[4]

In the economics literature, Carlsson et al. (2020) provide evidence that increased representation of radical populist parties in Swedish municipality elections decreases public support for their signature policies.[5] They suggest higher politician turnover and negative media coverage as potential channels for their effects. Our survey experiments complement this analysis by directly testing the effect of party positions on public preferences net of such mediating channels and by indicating that the effect is at work beyond extreme parties.

Methodologically, the chapter contributes to the growing literature in economics that employs survey experiments to examine the effects of information provision on public preferences (e.g., Cruces et al., 2013; Kuziemko et al., 2015; Bursztyn, 2016; Lergetporer et al., 2018a; Grigorieff et al., 2020; Haaland and Roth, 2020; Roth et al., 2020). We extend this literature by showing that the effects of information treatments on public preferences are not limited to information about facts which underlie the respective policy itself (e.g., the effect of informing about current public spending levels on preferences for spending increases), but that they are also prevalent in the context of information about party positions.

The remainder of the chapter is structured as follows. Section 5.2 provides background information on the investigated policies. Section 5.3 describes the opinion survey and the experimental design. Section 5.4 presents our main results. Section 5.5 analyzes effect heterogeneities. Section 5.6 concludes.

5.2 Background Information

This section provides background on the two German family policy initiatives whose policy preferences we investigate in this chapter, the child care subsidy and universal student aid.

5.2.1 Two Family Policies with Differing Public Salience

Family policies have important repercussions in many areas of life, including fertility, parent's labor-market participation, and children's educational opportunities. Since family policies entail direct government interventions in family life and decision making, it is particularly interesting to study whether the public's views on them are affected by the positions of political

[4] According to social identity theory, providing partisans with information on the positions of other parties can trigger out-group pressure to conform to their preferred party's position (e.g., Samuels and Zucco, 2014). Thus, holistic information treatments as ours guarantee to capture the gross effect of such out-group cues across the whole party spectrum, which would be missed in partial-information frameworks.

[5] Related research in economics studies the determinants of citizens' voting behavior, as opposed to their policy preferences (e.g., DellaVigna et al., 2017).

parties. In addition, political parties differ widely in their views about different policies to support and influence families.

The two specific family policies on which we focus are the child care subsidy and universal student aid, both of which offer universal financial support for their respective target group irrespective of income. Before the national German elections in 2013, each party took a position on these two policy initiatives.[6] This allows us to use the same uniform information treatment for all respondents, irrespective of their state of residence. We expected that at least some partisans are uninformed about their party's position on the two policies, which constitutes a necessary condition for the information treatment to affect policy preferences (e.g., Samuels and Zucco, 2014; Slothuus, 2016). While we abstained from eliciting respondents' prior beliefs about party positions directly (see section 5.3.2), the significant information effects reported below suggest that the selected policies fulfill this criterion.

A crucial difference between the policies is that the child care subsidy was actually implemented, whereas universal student aid was discussed but never introduced. As a result, both policies differed greatly in their public visibility. While the introduction of the child care subsidy was accompanied by a controversial public debate, discussions of universal student aid have never taken center stage. Focusing on two policy initiatives with different public attention allows us to study whether endogeneity of public policy preferences hinges on the extent of public salience of the policy under consideration.

The difference in public attention is nicely illustrated by the relative search frequencies of the two policies on the internet. Appendix Figure A5.1 depicts the relative frequency of Google search requests in Germany from January to June 2015 (i.e., the half-year before the end of our survey's three-month field phase; see section 5.3.1). The figure clearly shows that the relative number of Google search requests for the child care subsidy ('Betreuungsgeld') dwarfed search requests for universal student aid ('Elternunabhängiges BAfoeG'). To put these numbers into perspective, the figure also shows searches for the German chancellor ('Angela Merkel'). As it turns out, search requests for the child care subsidy are comparable to, and sometimes exceed, searches for the chancellor, whereas searches for universal student aid were much lower throughout the entire period. Accordingly, one might expect that the share of citizens who are aware of the different parties' positions might also have been higher for the child care subsidy than for universal student aid.

[6] In general, it is not self-evident that the different federal-state branches of a given party agree on a common position on a family policy, because family policies might vary across federal states. For the case of the two policies investigated here, however, party positions were internally consistent. Our main source of information on the different party positions is the voting advice application Wahl-O-Mat for the German federal elections 2013 (see `www.bpb.de/politik/wahlen/wahl-o-mat`). Created by the German Federal Agency for Political Education, the Wahl-O-Mat elicits party positions on various policies directly from the parties using a three-point scale ('favor', 'oppose', or 'neutral'). Throughout the chapter, we use this three-point scale to inform about parties' policy positions. We cross-checked party-position information using Wahl-O-Mat content for previous state elections, the state parties' programs, and, where necessary, direct inquiries to federal parties' bureaus.

5.2.2 Child Care Subsidy

The child care subsidy (*Betreuungsgeld*) was introduced Germany-wide in August 2013. The law entitled parents to receive a monthly payment of 150 Euro for each child in the second and third year of age if the respective child did not attend a publicly subsidized childcare program. Eligibility for the subsidy was independent of income. The subsidy was paid in addition to other family support programs. After our survey, the German Federal Constitutional Court abolished the existing policy in July 2015 because it interfered with the legislative autonomy of the federal states in the area of family policies that is guaranteed in the German constitution.

Proponents of the subsidy argue that the payments improve freedom of choice for young families between private childcare (including care by a stay-at-home parent) and public childcare programs. According to proponents, incentives for families are distorted by public financing of public childcare facilities, and the child care subsidy increases efficiency of family's choices and hence children's well-being (Fichtl et al., 2012).

Opponents primarily criticize that the policy would decrease public childcare enrollment rates among children from low socioeconomic backgrounds, as their parents have lower opportunity costs of parental childcare. At the same time, these children would likely benefit most from public childcare (e.g., Elango et al., 2015). Another point of criticism is that the policy caters to traditional gender roles by providing disincentives for young mothers to re-enter the labor market (Schuler-Harms, 2010).

The six major political parties in Germany, CDU/CSU (the federal Christian Democratic Union and its Bavarian sister party Christian Social Union), SPD (Social Democratic Party), Linke (Left Party), Gruene (Green Party), AfD (Alternative for Germany), and FDP (Free Democratic Party), adopted clear positions on the issue. While the major conservative party (CDU/CSU) was in favor of the policy, the more left-leaning parties Linke, Gruene, and SPD opposed the policy and demanded its abolishment. The right-wing AfD also opposed the child care subsidy. The liberal FDP was rather neutral.[7]

5.2.3 Universal Student Aid

Governmental financial support for university students is based on the Federal Training Assistance Act (*Bundesausbildungsförderungsgesetz*, commonly known as *BAfoeG*), which currently provides students from low-income households with direct transfers of up to 735 Euro a month. Half of the payment is a grant, and the other half is an interest-free loan. Eligibility and the

[7] Many major nationwide newspapers in Germany commented on the ongoing discussion among the different parties (see, for instance, Ulrike Meyer Timpe in *Die Zeit* Nr. 46/2007 `www.zeit.de/2007/46/Argument-Kinderbetreuung` [accessed 12 December 2018] or Rudzio Kolja in *Die Zeit* Nr. 32/2013 `www.zeit.de/2013/32/sozialpolitik-betreuungsgeld` [accessed 12 December 2018]). Interestingly, the German parliament had passed the child care subsidy in 2012 with the votes of delegates from CDU/CSU and FDP.

amount paid depend on parental income, students' own income, their partners' income (if applicable), and the number of siblings.[8]

A proposal to extend this system to lump-sum payments for all students (regardless of individual financial need and family resources) was discussed as part of the campaign for the German federal election in 2013. According to proponents, this change in policy would reduce the administrative burden of student aid management, reduce uncertainty about available financial support for prospective students, and encourage university enrollment. Opponents emphasize increased fiscal costs and argue that a change to universal student aid would be highly regressive since a disproportionate share of university students is from high socioeconomic backgrounds.

In the political debate, the FDP, as well as Gruene and Linke, support the proposal, whereas CDU/CSU and the AfD oppose it. The SPD is relatively neutral. Although the parties disagree in their stance on the introduction of universal student aid, the reform proposal was not as controversially discussed in the public as the child care subsidy.[9]

5.3 Data and Empirical Strategy

This section describes the opinion survey, the design of the experiments, and the empirical model.

5.3.1 The Opinion Survey

We implemented our experiments in an opinion survey with 4,105 respondents that we conducted in Germany in 2015, the ifo Education Survey. The sample consists of a baseline sample (N=3,063) that is representative for the German voting-age population (18 years and older) and an oversample of parents with children aged between 6 and 15 years (N=1,042). The oversample allows us to study preferences of those who are potentially affected by the reforms. The survey encompassed 33 questions related to educational topics as well as questions about respondents' sociodemographic characteristics.[10] Median completion time was 18 minutes. The sampling and polling were carried out by the survey company TNS Infratest (now called Kantar Public) between April and June 2015.

8 See `www.bafög.de` [accessed 12 December 2018] for details.

9 Media coverage was far less than for child care subsidy. Heike Schmoll's article in *FAZ* 2013 (`www.faz.net/aktuell/politik/bundestagswahl/die-plaene-der-parteien/plaene-der-parteien-8-der-bund-soll-mitreden-und-zahlen-12538721.html` [accessed 12 December 2018]) is one of the few contributions in major newspapers which address universal student aid.

10 The questions and wordings can be found at `www.cesifo-group.de/ifo-bildungsbarometer`.

While rare in experimental analyses, survey representativeness constitutes a key requirement for studying determinants of the electorate's policy preferences (e.g., in the framework of median voter models). Since computerized surveys likely produce non-participation bias for people who are less familiar with digital technologies, TNS Infratest collected the data in two strata. First, persons who use the internet (80 percent) were drawn from an online panel and answered all questions autonomously on their devices. Second, persons who reported not to use the internet (20 percent) were surveyed at their homes by trained interviewers. These respondents were provided with a tablet computer for completing the survey. This mixed-mode design allows us to draw general conclusions for the German electorate.

All analyses presented in this chapter use survey weights that were designed to match official statistics with respect to age, gender, parental status, school degree, federal state, and municipality size. In our main analysis, oversampled parents are weighted down accordingly to assure representativeness of the German electorate.

5.3.2 The Survey Experiments

Experimental Design

Our aim is to investigate whether information provision on competing party positions changes public support for the two family policies. In both experiments, we provide information to a randomly selected group of respondents before eliciting their preferences for the respective policy in the same way as in the uninformed control group. Our information treatments informed respondents about the official positions held by the six main German parties.

In our first experiment, we test the impact of information provision on preferences for the child care subsidy. The main question was worded as follows: *'The government pays parents who do not enroll their children aged 2 to 3 years in a childcare facility, but instead provide private home care, a child care subsidy in addition to the child benefits. Do you favor or oppose that parents receive a child care subsidy in addition to the child benefits?'*[11] Respondents were asked to select one of the following five answer categories: strongly favor, somewhat favor, neither favor nor oppose, somewhat oppose, strongly oppose.[12] In contrast to the uninformed control group, respondents in the information treatment group received the following information

[11] Child benefits (*Kindergeld*) refer to the financial support paid by the German government to parents. In 2015, the amount paid per child was 188 Euro for the first and second child, 194 Euro for the third child, and 219 Euro for the fourth child and any additional children.
[12] Appendix Table A5.1 presents the question wordings of the experiments in this chapter, and Appendix Figure A5.2 provides screenshots of the survey questions as they appeared on respondents' devices. To prompt people to give a considered answer and to minimize the error of central tendency, the category 'neither favor nor oppose' was placed below the other answer categories for both questions. We implemented a methodological experiment on another survey question (on granting teachers civil service protections) and found that the position of the neutral category does not change relative support and opposition towards the policy proposal (not shown).

when stating their policy preferences: *'CDU/CSU tend to favor the child care subsidy, SPD, Linke, Gruene, and AfD tend to oppose it, the FDP is rather neutral.'*

Our second experiment assesses the impact of party-position information on public preferences for universal student aid. The question was worded as follows: *'BAfoeG is federal financial aid for students which is paid contingent on parents' income. Do you favor or oppose that all students should generally receive BAfoeG by the government irrespective of parents' income?'* Respondents in the treatment group were additionally informed about the following party positions when stating their preferences for universal student aid: *'Linke, Gruene, and FDP tend to favor paying BAfoeG irrespective of parents' income, CDU/CSU and AfD tend to oppose it, the position of the SPD is rather neutral.'*[13]

We list party positions on the whole spectrum of the political landscape rather than just providing information on the position of the party the respondent supports. This practice follows previous studies that provided information on more than one party's position (e.g., Cohen, 2003; Levendusky, 2010; Druckman et al., 2013; Samuels and Zucco, 2014). Since all treated respondents receive the exact same information, we can directly compare the effects of information provision across respondents with different party preferences. This design feature also allows us to elicit party preferences after respondents stated their policy preferences, which lowers the risk that stated policy preferences are influenced by priming party identities or by respondents' preferences for consistency (e.g., Falk and Zimmermann, 2013).[14]

Providing information on the whole spectrum of party positions yields gross treatment effects of providing full versus no party information. These gross effects may stem from respondents who (i) adopt the position of the party they support or (ii) react to the other parties' positions (e.g., by taking the opposite stance from parties which they dislike). Additionally, respondents might respond to the distribution of parties across the spectrum, for example, by basing own policy preferences on the number of parties in support of a certain policy. While we consider the gross effects particularly relevant for policy, disentangling these and other treatment-effect components is an interesting avenue for future research.

[13] Like many other recent papers using survey experiments, our outcomes of interest are self-reported policy preferences (e.g., Cruces et al., 2013; Kuziemko et al., 2015; Karadja et al., 2016). Recent evidence corroborates the relevance of survey-based outcome measures since they closely correspond to actual political behavior, such as signing petitions or donating to charity (e.g., Alesina et al., 2018b; Haaland and Roth, 2020; Roth et al., 2020).

[14] To study belief-updating about party positions, it would have been interesting to elicit respondents' prior beliefs about each party's position on the two policies. We abstained from measuring these priors to minimize the risks of priming control-group respondents with their party identity (e.g., Alesina et al., 2018a) and of invoking experimenter-demand effects (Zizzo, 2010).

Eliciting Partisanship for Political Parties

In eliciting party preferences, we focus on partisanship as a long-term tendency towards a certain party, rather than short-term voting intentions. In the context of investigating the endogeneity of policy preferences with respect to party positions, long-term party attachment is particularly relevant because it reflects fundamental political values instead of short-term considerations guiding intended voting behavior. Furthermore, combining information on long-term party attachment with information on voter participation allows us to identify swing voters, which are a particularly interesting focus group (e.g., Zaller, 2004). The measure of partisanship should not be influenced by information on the parties' positions on just two issues within the whole policy spectrum. To corroborate that this is not the case, Appendix Table A5.2 regresses stated partisanship on the two treatments and their interaction. None of the coefficients on any of our treatments is statistically significant, individually or jointly, for any party. The fact that the treatment does not predict stated long-term party preferences underlines the validity of our measure of partisanship and also suggests that experimenter demand effects are not important for this question.[15]

Partisanship extends over a wide spectrum of political parties. In our sample, 31.0 percent of respondents identify as non-partisans. 23.0 percent state to generally support CDU/CSU, 21.3 percent support SPD, 8.6 percent Linke, 7.5 percent Gruene, and 4.4 percent AfD. Only 2.2 percent support FDP, and 1.7 percent support other parties.

Partisans of the different parties differ substantially in their sociodemographic characteristics. Table 5.1 provides descriptive results from a multinomial logit regression of partisanship on sociodemographic characteristics. It is reassuring that stylized facts about partisans are replicated in our data. For instance, respondents with higher income are more likely to support the conservative party CDU/CSU or the liberal party FDP, and are less likely to sympathize with the left party Linke. Respondents with low educational attainment are more likely to support SPD and less likely to support Gruene, and respondents living in East Germany are more likely to support Linke and less likely to support SPD and Gruene.

5.3.3 The Econometric Model

We estimate the effects of the experimental information treatment on policy preferences with the following regression model:

$$y_i = \alpha_0 + \alpha_1 T_i + \delta^{'} + \epsilon_i \tag{5.1}$$

where y_i is the outcome variable of interest for respondent i, T_i indicates whether respondent i received the information treatment, X_i is a vector of control variables, and ϵ_i is an error term.

[15] Consistently, Mummolo and Peterson (2019) show that experimenter demand effects are likely absent in survey experiments, and de Quidt et al. (2018) show that they hardly affect results of experiments on economic preferences.

Throughout the chapter, we measure the outcome variable both as the probability to support and the probability to oppose the respective policy, but we also analyze effects on each of the five underlying answer categories separately to investigate preference intensity. Since ϵ_i is uncorrelated with treatment status through randomization, the parameter α_1 provides an unbiased estimate of the causal treatment effect of information provision even without including further covariates. However, since the inclusion of covariates can increase the precision of estimates, we show results both with and without covariates in our main analyses.

Since we expect the information treatment to operate through respondents' party preferences, we are particularly interested in heterogeneous treatment effects by partisanship. In our preferred specification, we group individuals into four categories according to their partisanship: supporters of parties that favor the respective policy, supporters of parties that are neutral towards the policy, supporters of parties that oppose the policy, and non-partisans. Hence, we additionally employ the following regression model:

$$y_i = \beta_0 + \beta_1(T_i \sum_j P_i^j) + \beta_2 \sum_j P_i^j + \delta^{'} X_i + \eta_i, \; with \; j \; \in \{f, n, o, non\} \tag{5.2}$$

where P_i^f equals 1 if respondent i supports a party that favors the respective policy (0 otherwise), P_i^n refers to supporters of parties that are neutral towards the policy, P_i^o to supporters of parties that oppose the policy, and P_i^{non} to respondents who report no particular long-term party preference. The information treatment effects on those who support a party that favors, is neutral towards, or opposes the policy and for non-partisans are given by the coefficients β_1. In additional analyses, we also show disaggregated results for partisans of each of the six major parties.

5.3.4 Test of Randomization

To check whether randomization in our two experiments successfully balanced respondents' observable characteristics c_i between control and treatment groups, we estimate the following model for each covariate and both experiments:

$$c_i = \gamma_0 + \gamma_1 T_i + \xi_i \tag{5.3}$$

Table 5.2 reports the γ_1-coefficients from these regressions along with the corresponding means of the covariates. Sociodemographic characteristics are well balanced across experimental groups: There are small but significant differences (p<0.1) in only 3 out of 64 pairwise comparisons. Thus, the share of significant differences does not exceed the share that would be expected by pure chance. In addition, regressing treatment status simultaneously on all covariates, partisanship, and an indicator for item non-response yields *p*-values for joint significance of 0.478 in the child care subsidy experiment and 0.140 in the universal student aid experiment.

While item non-response is very low at around one percent, there is a statistically significant difference in non-response between treatment and control group in the experiment on child care subsidy (see bottom part of Table 5.2). In order to rule out that non-random item non-response drives our results, we run the following bounding exercise: For the 13 respondents (3 in the control group and 10 in the treatment group) who did not state their preferences, we assign policy preferences that deviate as much as possible from the position of their preferred party. While this imputation makes it less likely to detect treatment effects, it leaves our qualitative results unchanged (results available upon request).[16]

5.4 Party-Position Information and Public Policy Preferences: Main Results

This section presents our main results on how the experimental provision of information about the positions of political parties affects the public's policy preferences for the child care subsidy and for universal student aid.

5.4.1 Treatment Effects on Policy Preferences for Child Care Subsidy

We start our analysis by investigating whether the information treatment changes average public support for the child care subsidy across all respondents. The top panel of Table 5.3 depicts the results from regressions based on equation 5.1, investigating both support for and opposition against the child care subsidy.[17] Odd-numbered columns present estimates without controls, even-numbered columns include a rich set of sociodemographic control variables.[18]

The results indicate that a majority of respondents in the control group (56 percent) opposes the child care subsidy (see control mean). Only a minority (34 percent) supports it. The remainder is neither in favor nor opposed. As the small and insignificant coefficients on the treatment indicator show, the provision of information about the different parties' positions does not affect average support for, or opposition against, the child care subsidy. This average null effect is not surprising, given that some respondents learn that their preferred party supports the policy, whereas other respondents learn that their preferred party opposes it.

[16] We ran the bounding exercise twice: While partisans who did not provide an answer were always assigned the preference furthest away from their favored party's position, non-partisans were first assigned answer category 'strongly favor' and then 'strongly oppose'. Note that no partisan of parties with a neutral position skipped this question.

[17] We use linear probability models throughout the chapter. (Ordered) probit models lead to the same qualitative results (available upon request).

[18] The controls are essentially those listed in Table 5.2; see notes to Table 5.3 for details. The slightly reduced number of observations in the specifications with controls is due to item non-response on control variables; imputing missing observations on the control variables and running the model with control variables for the full sample provides qualitatively identical results (available upon request).

Since we are primarily interested in whether partisans align their policy preferences with their preferred party's position, we next investigate preferences by partisanship. Figure 5.1 illustrates our main results on preferences for the child care subsidy. In the uninformed control group, the majority of partisans of the party which favors the child care subsidy actually opposes the policy (32 percent in favor, 58 percent opposed). Thus, partisans' uninformed preferences for the child care subsidy do not reflect their favored parties' position. Interestingly, these shares are statistically indistinguishable from partisans of parties which oppose the policy (31 percent in favor, 63 percent opposed). By contrast, non-partisans' preferences are significantly more favorable towards the child care subsidy (40 percent in favor, 47 percent opposed).

Importantly, respondents' reactions to the information treatment depend on partisanship. Table 5.4 presents estimates of heterogeneous treatment effects with respect to respondents' partisanship based on equation 5.2. There are no statistically significant treatment effects on policy preferences of respondents who support parties that oppose or are neutral to the child care subsidy and of non-partisans.[19]

In contrast, the information treatment significantly shifts preferences of partisans of the party that favors the child care subsidy, aligning their preferences more closely with their preferred party's position (see also Figure 5.1). Among this group, the treatment increases support for the policy from 32 percent to 43 percent and turns a majority opposing the policy (58 percent opposed) into a minority (47 percent remain opposed). Both the increase in support by 10.2 percentage points and the decrease in opposition by 10.9 percentage points are statistically and quantitatively significant.[20] Thus, these partisans' preferences are endogenous with respect to their preferred party's position.

The party-position information does not only affect weakly held preferences, but also strongly held preferences. The coefficients on the interaction term between the treatment indicator and partisans of the party that favors the policy in Appendix Table A5.3 shows that the treatment effects reported in Table 5.4 stem from shifts both in strongly held preferences (columns 1 and 5) and weakly held preferences (column 2).[21]

[19] While imprecisely estimated, the coefficients for respondents whose parties are neutral are quantitatively quite large.

[20] Since we test for treatment effects in four subgroups (partisans of parties favoring, opposing, and neutral to the policy, as well as non-partisans), false positives due to multiple hypothesis testing are a potential concern (e.g., List et al., 2019). The most conservative correction to account for multiplicity is to multiply the unadjusted *p*-values with the number of hypotheses tested (Bonferroni, 1935). The *p*-value of the Wald test for treatment effects for partisans of parties that favor the policy is $p=0.004$ in the preferred specification of column 2 of Table 5.4. That is, the treatment coefficient remains significant ($p<0.02$) after correcting for the four hypotheses tested.

[21] Eichenberger and Serna (1996) argue that information can alter voting behavior by changing the distribution of voters' assessments of a given policy, even if information does not affect the mean assessment. We do not find evidence for such an effect in our data: The variance of policy preferences (on a five-point scale) among partisans of the party that favors the child care subsidy does not differ significantly between the control group and the treatment group ($p=0.64$, unweighted F-test for homogeneity in variances).

5.4.2 Treatment Effects on Policy Preferences for Universal Student Aid

The bottom panel of Table 5.3 reports results from estimating equation 5.1 for the experiment on universal student aid. Overall, the German public is divided on the issue of whether student aid should be paid independently of parental income (48 percent in favor, 45 opposed, remainder neither in favor nor opposed). Again, the information treatment has no average effect on the preferences of the public as a whole.

In the uninformed control group, partisan preferences again do not align well with their preferred party's position (Figure 5.2). The shares of respondents supporting universal student aid are statistically indistinguishable between partisans of parties favoring the policy (52 percent in favor, 43 percent opposed) and partisans of parties opposing the policy (51 percent favor, 43 percent opposed). At the same time, non-partisans are significantly ($p<0.1$) less likely to support the proposal (43 percent in favor, 46 percent opposed).[22]

Estimates of heterogeneous treatment effects by partisanship in Table 5.5 again indicate that partisans are susceptible to the information treatment. While a narrow majority of partisans whose parties oppose universal student aid supports this policy in the control group, this support decreases significantly ($p<0.1$) to 43 percent when information about the parties' positions is provided (see also Figure 5.2).[23] Similarly, information provision aligns the preferences of partisans of parties that favor universal student aid more closely to their preferred parties' position, but this effect is not statistically significant at conventional levels.

Interestingly, in this case also policy preferences of non-partisans are affected by the information provision. Among those respondents who indicate that they do not lean towards a particular political party in the long term, providing information about the different parties' positions on the policy significantly decreases opposition from 46 percent to 38 percent ($p<0.05$) and increases support from 43 percent to 47 percent.

The bottom panel of Appendix Table A5.3 presents treatment effect estimates on each of the five answer categories. In this case, the treatment primarily affects weak preferences among partisans of parties that oppose universal student aid.

[22] Previous evidence showed positive associations between party positions and partisans' preferences towards classic redistributive policies (e.g., Kuziemko et al., 2015). The lack of such associations for the child care subsidy and universal student aid is consistent with the notion that party positions are less salient in the domain of the family policies investigated here.

[23] While the treatment effect of a 7 percentage-point decrease in policy support for these partisans is quantitatively significant, the power of our statistical tests is limited: Given the p-value of the corresponding coefficient in column 2 of Table 5.5 ($p=0.098$), the significance of this effect is not robust to Bonferroni-type adjustment for multiple hypothesis testing.

5.4.3 Discussion

We designed our experiments to test the basic question of whether the public's preferences towards family policies can be endogenous to the position that their preferred political party takes on these policies. As such, our result that some partisan groups significantly change their policy preferences in response to party-position information provides proof of concept that party positions can indeed be important for shaping public preferences on family policies.

At the same time, our results also show that the relevance of this mechanism is not universal in all partisan groups.[24] While the experimental design is not rich enough to provide general answers as to when party positions do and do not have significant effects, the results provide some indications that may warrant further analysis in future research aimed at going beyond proof of concept into investigating cognitive and behavioral mechanisms of voters more closely.

First, our results suggest that the effect of party positions on public policy preferences does not hinge on the policy's salience. We find significant treatment effects of party-position information both in the rather salient policy topic of the child care subsidy which was subject to heated and highly visible public debates and in the less salient policy topic of universal student aid that had limited public visibility. Relatedly, treatment effects might depend on respondents' prior knowledge about parties' positions on the two policies. While we abstained from directly eliciting these priors to minimize the risk of priming and experimenter-demand effects (see section 5.3.2), additional analyses based on proxies for respondents' prior knowledge suggest that treatment effects tend to be stronger for less-informed respondents (not shown), in line with some role for genuine belief-updating about party positions.[25]

Second, our experiments provide evidence that party-position information can align partisans' preferences to their preferred party's position if the respective party takes a clear stance for or against the policy under consideration. This result speaks to the literature on political polarization (e.g., Druckman et al., 2013) by showing that information about party positions can increase polarization in the electorates' policy preferences. Thus, they provide a complementary explanation for why voters become more polarized with increased availability of information (e.g., Ortoleva and Snowberg, 2015).

At the same time, the reason why we find effects for partisans of policy-favoring but not policy-opposing parties in the first experiment, and vice versa (at least in terms of statistical

[24] In fact, in a related analysis we do not find significant effects of party-position information on preferences for abolishing the constitutional prohibition for the federal government to interfere in states' education policies (*Kooperationsverbot*; results available upon request).

[25] As proxies for respondents' prior information about party positions, the additional analyses use information on (i) respondents' media-consumption behavior and (ii) whether partisan respondents' preferred party provides the state's education minister. Respondents who are classified as better informed based on these proxies tend to react less strongly to the information treatments, although statistical power does not allow to distinguish the subgroups' treatment effects significantly (results available upon request).

significance) in the second experiment, is less clear. One possible interpretation relates to the homogeneity of the respective party groups about which respondents are informed. In the first experiment, only one party favors the policy. By contrast, the group of parties opposing it is very heterogeneous, including parties both from the left and the right side of the political spectrum. Conversely, the group of parties favoring the policy in the second experiment is very heterogeneous, including both left-wing parties and the right-of-center liberal party, whereas two right-of-center parties oppose the policy. Thus, being informed about parties' positions may fail to align partisans' preferences with their preferred party's position if they learn that parties on the other side of the political spectrum take the same position on the policy under consideration.[26] While this is one possible interpretation, degrees of freedom in comparing two separate experiments do not allow for a solid test of this proposition, and alternative interpretations are certainly possible. For example, the extent to which partisans align their preferences with their parties' positions may ultimately be policy-specific, suggesting the usefulness of future analyses beyond the two family policies studied here.

Third, in a similar way non-partisans may in fact be affected by party-position information even though they do not have a long-term tendency to follow one specific party. Learning how different political parties—both ones they may like and ones they may dislike—stand towards the policy may be relevant for their preference formation process, as well. In particular, the group of non-partisans combines two distinct subgroups: on the one hand, swing voters who show political interest but are not bound to one party; and on the other hand, people with limited interest in politics and voting in general. We turn to this distinction next.

5.5 Heterogeneous Treatment Effects

This section tests for heterogeneities in treatment effects, focusing on swing voters among the non-partisans and on different sociodemographic subgroups among partisans.

5.5.1 Swing Voters: Effect Heterogeneity by Voting Behavior among Non-Partisans

Results in the previous section showed that non-partisans' preferences towards the child care subsidy are unaffected by the information treatment, whereas the treatment does affect their preferences towards universal student aid. Hence, also non-partisans' preferences can be endogenous with respect to party positions. To provide a more detailed analysis of non-partisans, we examine effect heterogeneities with respect to their voting behavior. A virtue of measuring partisanship as long-term party attachment is that it allows us to identify a highly relevant subgroup of the electorate: swing voters. We define swing voters as individuals who

[26] The same type of reasoning may in fact explain why, although treatment effects go in the same direction for partisans of the two policy-opposing parties in the second experiment, they are much larger and reach statistical significance only for AfD partisans (see Appendix Table A5.4 for disaggregated results for each of the six major parties). If partisans of the center-right CDU/CSU dislike the fact that the far-right AfD takes the same position, they may be less triggered to change their preferences to align with their preferred party.

(i) do not have a long-term tendency towards any particular party *and* (ii) regularly turn out to vote. Among non-partisans, 52 percent report to vote regularly.[27]

We restrict our sample to non-partisans and estimate regression models similar to equation 5.2, but interact the treatment dummy with an indicator for regular voting. This allows us to estimate treatment effect heterogeneities between swing voters and non-partisans who usually do not vote.

Table 5.6 reports heterogeneous treatment effects on preferences towards the child care subsidy (upper panel) and universal student aid (lower panel). Non-partisans who do not vote regularly are significantly less likely to support child care subsidy if they are informed about the parties' positions. The treatment effect for swing voters is insignificant.

In the experiment on universal student aid, by contrast, the overall treatment effect of non-partisans is driven by swing voters. While non-partisans who vote infrequently do not exhibit significant treatment effects, the treatment significantly increases swing voters' support by 10 percentage points and decreases opposition by a similar amount.

This analysis shows that information about party positions can impact policy preferences of swing voters, a highly relevant group within the political process. While it is likely that the effects of party positions on partisans' policy preferences reported in the previous section are due to partisans actively aligning their preferences with their favored party's position, the channel through which non-partisans incorporate party positions in their preferences is less clear. For instance, it might be that these respondents orient themselves towards the positions of the governing parties (CDU/CSU and SPD), towards the positions of the majorities of parties, or that they use out-group cues in order to choose a position that is different from their least-preferred party (e.g., Samuels and Zucco, 2014). In any case, our results highlight the policy relevance of party positions, since they not only influence partisans but also persons without long-term party attachment.

5.5.2 Effect Heterogeneities by Sociodemographic Characteristics

Our main results show that the effects of party positions vary with respondents' partisanship. In this section, we investigate the extent to which treatment effects are prevalent in different sociodemographic subgroups. For this explorative analysis, we focus on partisan groups that have been identified as susceptible to the information treatments in Section 5.4—i.e., partisans of the party that favors the child care subsidy and partisans of parties that oppose universal student aid. We concentrate on the following sociodemographic characteristics: gender, age, income, employment status, educational attainment, parental status, and importance of

[27] We use the following question to elicit voting frequency: *'Do you usually vote in federal and state elections (including postal voting)? I vote…'* The answer categories are: always, mostly, sometimes, rarely, and never. We define regular voters as those who either 'always' or 'mostly' turn out to vote. Among all respondents, 79 percent vote regularly. Among partisans, the share is 90 percent.

education policy for the personal voting decision.[28] For each characteristic, we divide our sample into two subgroups and estimate regression models based on equation 5.1.

Tables 5.7 and 5.8 report results on preferences towards the child care subsidy and universal student aid, respectively. The first column presents the coefficient for the whole sample of respondents who support a party that favors the child care subsidy respectively opposes universal student aid. The subsequent columns show the coefficients for the respective subgroups.

Table 5.7 shows that the positive effect of party position information on support for the child care subsidy captures statistical significance in eight out of 14 subgroups: females, respondents above median age, respondents below median income, non-working respondents, respondents without high school degree, respondents without children aged below 19 years, and respondents who do and do not state that education policy is important for their voting decision. In the remaining subsamples, the coefficients are statistically insignificant, although with one exception (actively employed) they all point in the same direction. Across the different subgroups, the coefficient of interest ranges from -0.011 to 0.185.

In addition to this subsample analysis, we also estimated triple interaction models using the whole sample of partisans of the party that favors the child care subsidy. This allows us to test whether heterogeneous treatment effects across sociodemographic subgroups are statistically significant. Except for employment status, where treatment effects differ significantly between working and non-working respondents ($p<0.05$), effect heterogeneities across subgroups do not capture statistical significance (results available upon request).

Table 5.8 depicts the results on support for universal student aid. We find significant negative treatment effects for respondents below median age, respondents above median income, working respondents, those without a high school degree, parents with children aged below 19 years, and respondents for whom education policy is not important for their voting decision. Again, coefficients exhibit a wide range from -0.033 to -0.143, although all point in the same direction. The triple interaction models reveal no statistically significant effect heterogeneities across sociodemographic subgroups (results available upon request).

This heterogeneity analysis yields some noteworthy patterns. Most intriguingly, it shows that even parents—i.e., those who are directly affected by the reform—decrease their support for universal student aid in response to information on party positions. It is important to note, however, that these results need to be interpreted with caution because the relatively small sample sizes in the subgroups imply that not all quantitatively sizeable effects can be detected with sufficient precision.

[28] To obtain reasonable statistical power, we focus on characteristics where there are at least 100 observations both in the treatment group and in the control group.

5.6 Conclusions

According to classical political economy theories, political parties adapt their policy positions to exogenous voter preferences in order to maximize vote shares. Focusing on preferences for family policies, this chapter studies the reverse relationship, namely the extent to which voters' preferences are amenable to party positions. We devise two survey experiments among a representative sample of the German voting-age population with more than 4,000 respondents to shed light on the causal effect of party positions on the electorates' policy preferences.

Our results suggest that voters' preferences for family policies can be endogenous with respect to party positions. In our first experiment, information about party positions on the child care subsidy induces partisans of the policy-favoring party to align their preferences more closely with their preferred party. In our second experiment on preferences for universal student aid, the treatment aligns preferences of partisans of policy-opposing parties more closely with their preferred parties' position. Information effects are also prevalent among swing voters—non-partisans who regularly turn out to vote—and among various sociodemographic subgroups.

The result that policy preferences are not exogenous to the political process is consistent with survey responses on the importance of party positions as a source of information. In a separate survey, we asked another representative sample of voting-age Germans to rate the importance of different sources of information for forming their policy preferences.[29] About one third (32 percent) considers 'positions of the political parties' important. Notably, even 23 percent of non-partisan respondents perceive party positions important, corroborating our significant findings for non-partisans. While other opinion-formation aspects, such as own experience, expert opinions, or news reports, are rated as even more important for forming policy preferences, party positions are particularly important from a policy perspective because they are within the immediate action space of policy makers. Furthermore, the fact that party-position information affects some partisans' preferences for two policies with largely different salience suggests that the information effect may apply for a considerable range of policies.

Survey experiments are certainly subject to some artificiality, but they provide the rare opportunity to isolate the causal net effect of information about party positions on the public's policy

[29] The question was included in the ifo Education Survey 2017 and was worded as follows: *'How important are the following aspects for forming your opinion on education policy issues?'* For each potential source of information, respondents were asked to rate importance on a five-point scale from very important to very unimportant. The information sources (and their corresponding shares of respondents stating 'very important' or 'somewhat important') were: discussions with acquaintances (60 percent), experience from own school days (67 percent), experience from own children or grandchildren (69 percent), expert opinions (52 percent), positions of the political parties (32 percent), news reports (51 percent), gut instinct (54 percent), and other aspects (16 percent). Note that the ifo Education Survey 2017 did not comprise experiments on party positions.

preferences in a representative sample. Furthermore, several pieces of evidence underline the relevance of survey experiments for our understanding of political decision-making processes in the field. First, Barabas and Jerit (2010) show that the information effects in their survey experiment are also found, with somewhat smaller magnitude, in a natural experiment based on exposure to news covering the same information. Second, it has been shown that survey responses are good proxies for actual (voting) behavior (e.g., Kemp, 2003; Falk et al., 2016) and are important for shaping public policy (e.g., Engelhardt and Wagener, 2014). Consistently, Blinder and Krueger (2004) argue that the fact that politicians spend large amounts of resources on assessing public opinion shows the relevance of opinion surveys for the political process.

While our experimental design allows us to ascertain the baseline impact of party-position information on public policy preferences, it does not allow for a clean identification of the specific cognitive or behavioral mechanisms that give rise to the effect. One possibility is that uninformed voters may use party positions as heuristics to reduce the costs of preference formation (Tversky and Kahneman, 1974). While such an interpretation could account for some of our results, other aspects are harder to reconcile with a heuristics interpretation. In particular, people should be more likely to use heuristics when they have no information and only weak preferences on an issue. However, we find significant treatment effects (i) for two policies which differ in their salience in the public debate, (ii) among parents (who are presumably better informed about the issues), and (iii) on strong as well as weak preferences. An alternative possible interpretation is that partisans view their preferred party as their in-group, and that they derive direct utility from adjusting their preferences towards their in-group's position (Akerlof and Kranton, 2000). We consider deeper analysis of the mechanisms underlying the endogeneity of policy preferences to party positions an important direction for future research.

Irrespective of whether these or other potential mechanisms drive our results, our findings have implications for economic and political theory. The common assumption of the exogeneity of public policy preferences does not hold for the family policies studied here. Therefore, our results highlight the need for a more extensive consideration of potential endogeneities of preferences in the political economy literature. Relatedly, the finding that public preferences can be endogenous to party positions implies the risk of increased polarization among the public if parties take extreme positions (e.g., Glaeser et al., 2005). which may reduce overall welfare (e.g., by hampering public-goods provision in fragmented societies; see Alesina, 1988).[30]

[30] The normative implications of the two potential channels that we discuss are not clear-cut. On the one hand, using heuristics may be beneficial in terms of cognitive efficiency since they allow poorly informed voters to form policy preferences that are reasonably close to their well-informed preferences (which they would hold after gathering and processing all relevant information about the policy). Similarly, identity-based reactions to party-position information may be welfare-enhancing since in this case, aligning preferences with party positions directly increases respondents' utility. On the other hand, heuristics and identity motives may also

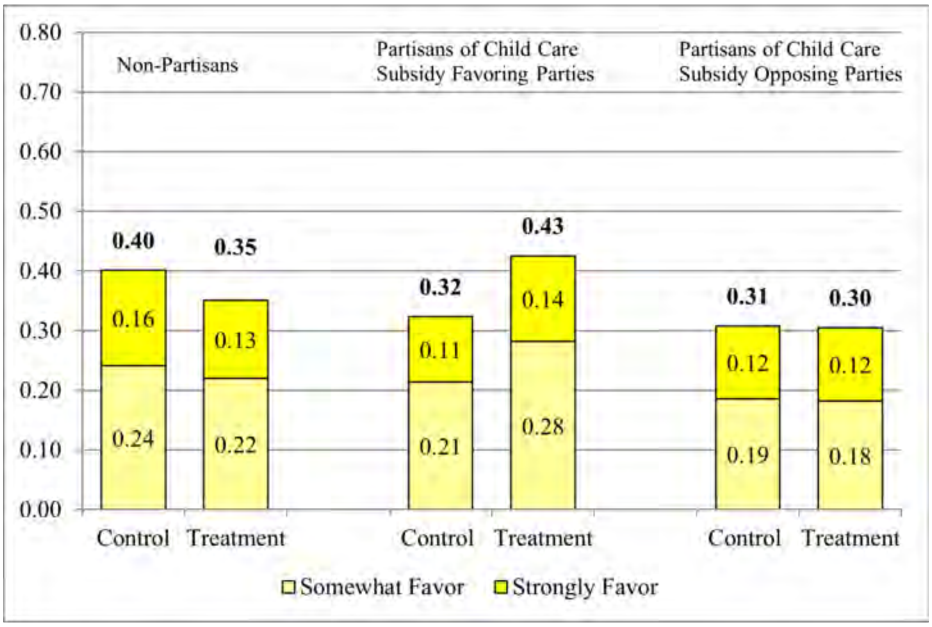
Our results also have implications for policy making and politics. Since broad public support is often decisive for successful policy implementation, the mere communication of party positions (even without putting forward any substantive arguments) can be important for the political feasibility of reform proposals. While our research does not inform about the welfare implications of this result, it shows that communicating party positions can affect public opinion.

induce systematic errors in voters' preferences by making them support policies that are not beneficial or even harmful for them.

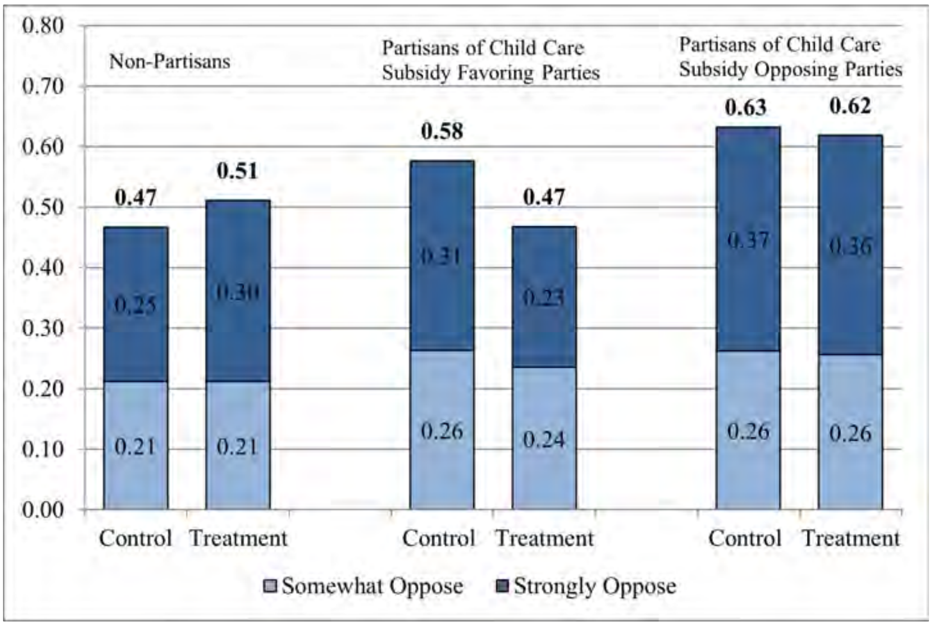
Figures and Tables

Figure 5.1: Effects of Party-Position Information Treatment on Preferences for Child Care Subsidy by Partisanship

Panel A: Support for Child Care Subsidy



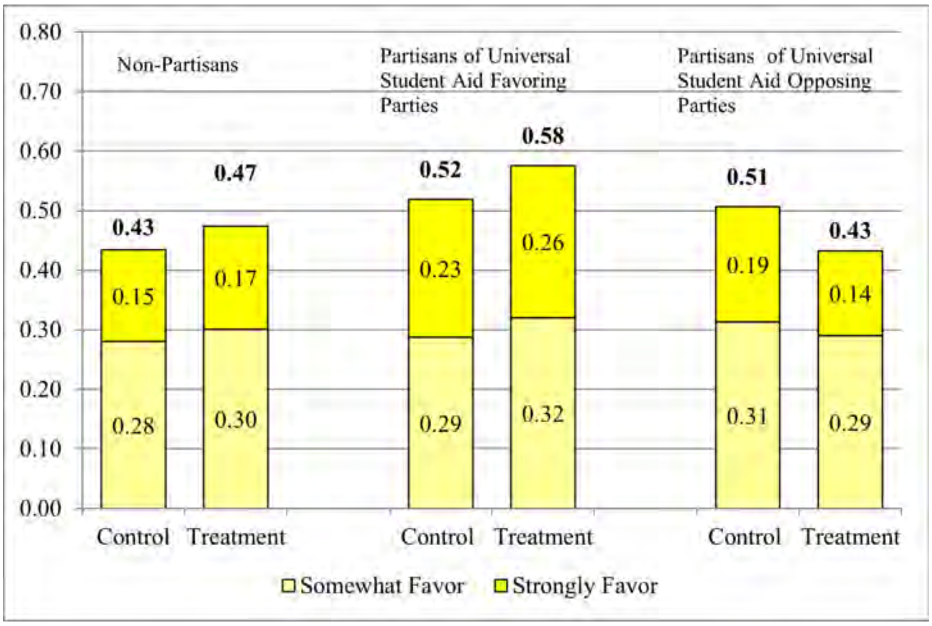
Panel B: Opposition against Child Care Subsidy



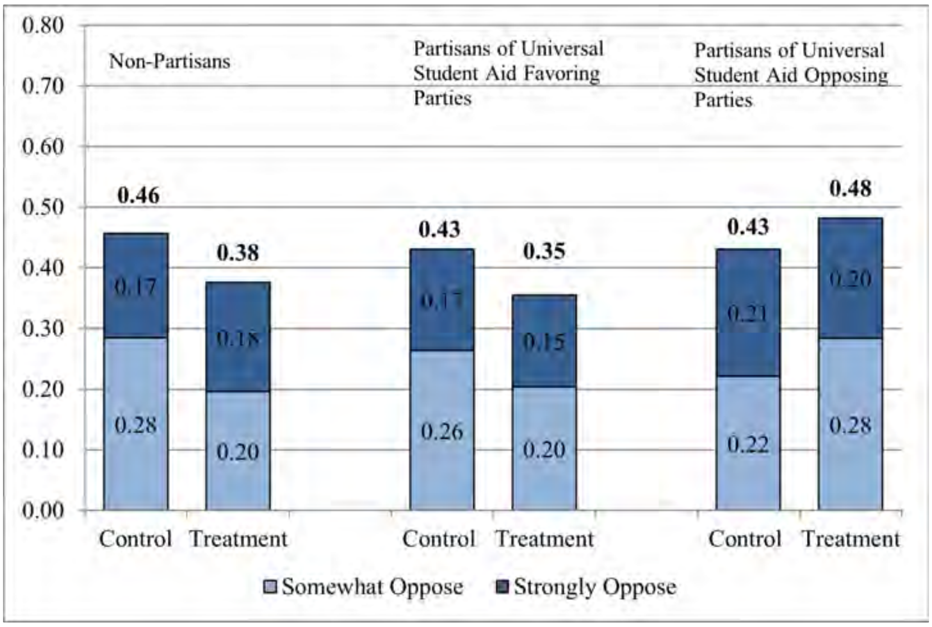
Notes: Wording of question. Control group: *'The government pays parents who do not enroll their children aged 2 to 3 years in a childcare facility, but instead provide private home care, a child care subsidy in addition to the child benefits. Do you favor or oppose that parents receive a child care subsidy in addition to the child benefits?'* Treatment: same wording as control group, with the following information included between the two sentences: *'CDU/CSU tend to favor the child care subsidy, SPD, Linke, Gruene, and AfD tend to oppose it, the FDP is rather neutral.'* ** indicates significance of difference from respective control group at p<0.05. See Table 5.4 for details and for results on partisans of parties neutral towards the policy. Data source: ifo Education Survey 2015.

Figure 5.2: Effects of Party-Position Information Treatment on Preferences for Universal Student Aid by Partisanship

Panel A: Support for Universal Student Aid



Panel B: Opposition against Universal Student Aid



Notes: Wording of question. Control group: *'BAfoeG is federal financial aid for students which is paid contingent on parents' income. Do you favor or oppose that all students should generally receive BAfoeG by the government irrespective of parents' income?'* Treatment: same wording as control group, with the following information included between the two sentences: *'Linke, Gruene, and FDP tend to favor paying BAfoeG irrespective of parents' income, CDU/CSU and AfD tend to oppose it, the position of the SPD is rather neutral.'* ** (*) indicates significance of difference from respective control group at p<0.05 (p<0.10). See Table 5.5 for details and for results on partisans of parties neutral towards the policy. Data source: ifo Education Survey 2015.

Table 5.1: Who Supports Which Party? Descriptive Evidence

	CDU/CSU (1)	SPD (2)	Gruene (3)	Linke (4)	FDP (5)	AfD (6)	Non-Partisan (7)
Highest Educational Attainment (Baseline: No/Basic Degree)							
Middle School Degree	0.014	-0.065**	0.038**	0.006	0.007	-0.010	0.024
	(0.023)	(0.022)	(0.013)	(0.016)	(0.007)	(0.013)	(0.025)
University Entrance Degree	-0.012	-0.070**	0.071**	0.024	0.025*	-0.006	-0.017
	(0.025)	(0.025)	(0.015)	(0.021)	(0.010)	(0.015)	(0.027)
Age (Baseline: Age 18-35)							
Age 36-65	0.019	-0.022	0.015	0.007	-0.011	-0.009	-0.011
	(0.036)	(0.040)	(0.023)	(0.027)	(0.015)	(0.020)	(0.040)
Age above 65	0.122*	0.006	-0.022	-0.033	0.015	-0.027	-0.077
	(0.050)	(0.053)	(0.030)	(0.031)	(0.020)	(0.022)	(0.052)
Monthly Household Income/1000	0.014*	-0.007	-0.001	-0.015*	0.005*	-0.006	0.010
	(0.007)	(0.007)	(0.004)	(0.006)	(0.002)	(0.003)	(0.009)
Female	-0.013	-0.019	0.005	-0.008	0.003	-0.039**	0.082**
	(0.020)	(0.021)	(0.013)	(0.013)	(0.006)	(0.010)	(0.021)
Born in Germany	-0.023	-0.067	0.055**	-0.060	0.016*	0.037**	0.023
	(0.044)	(0.050)	(0.013)	(0.049)	(0.007)	(0.009)	(0.041)
Lives with Partner	0.008	-0.022	0.008	-0.001	-0.013	0.018	-0.010
	(0.022)	(0.022)	(0.014)	(0.013)	(0.009)	(0.011)	(0.024)
Lives in West Germany	0.045*	0.081**	0.042**	-0.113**	0.003	-0.020	-0.042
	(0.020)	(0.020)	(0.013)	(0.021)	(0.008)	(0.012)	(0.025)
City size >= 100,000	-0.070**	0.066**	0.031*	0.004	0.005	0.014	-0.053**
	(0.019)	(0.020)	(0.014)	(0.012)	(0.007)	(0.009)	(0.020)
Parent Status (Baseline: no Children)							
At Least One Child Below 18	0.009	0.040	-0.040*	-0.004	0.000	0.020	-0.018
	(0.027)	(0.025)	(0.017)	(0.016)	(0.007)	(0.017)	(0.028)
All Children Older than 18	-0.004	0.053*	-0.017	0.037*	0.010	0.000	-0.068*
	(0.025)	(0.024)	(0.018)	(0.017)	(0.008)	(0.012)	(0.028)
At Least One Parent w. Univ. Degree	0.002	-0.038	0.041**	0.010	0.001	0.008	-0.041
	(0.020)	(0.020)	(0.014)	(0.014)	(0.008)	(0.014)	(0.021)
Labor Market Participation (Baseline: Full-time Employed)							
Part-Time Employed	-0.037	-0.014	0.010	-0.012	-0.001	0.026	0.014
	(0.024)	(0.031)	(0.016)	(0.018)	(0.012)	(0.019)	(0.027)
Self-employed	-0.065	-0.023	0.010	0.013	0.026	-0.006	0.045
	(0.033)	(0.036)	(0.022)	(0.027)	(0.017)	(0.013)	(0.038)
Unemployed	-0.091*	-0.008	-0.003	0.068	-0.019*	0.057	-0.017
	(0.042)	(0.047)	(0.023)	(0.041)	(0.009)	(0.038)	(0.040)
Housewife/Husband	0.025	-0.065*	0.028	-0.021	-0.003	0.008	-0.013
	(0.042)	(0.033)	(0.023)	(0.022)	(0.013)	(0.017)	(0.033)
Retired or Ill	0.019	0.016	0.006	-0.009	-0.007	0.007	-0.025
	(0.030)	(0.030)	(0.021)	(0.019)	(0.008)	(0.012)	(0.030)
Student, Apprentice, in Training	-0.081	-0.012	0.031	-0.052*	0.003	-0.005	0.025
	(0.058)	(0.065)	(0.038)	(0.026)	(0.022)	(0.030)	(0.065)
Works in Education Sector	0.001	-0.027	0.018	-0.018	-0.007	0.009	0.014
	(0.030)	(0.028)	(0.023)	(0.016)	(0.008)	(0.020)	(0.034)
Votes Regularly	0.141**	0.114**	0.037*	0.065**	0.011	0.038**	-0.414**
	(0.019)	(0.021)	(0.015)	(0.012)	(0.008)	(0.009)	(0.027)
Education Imp. for Voting Decision	-0.001	0.069**	0.030	0.000	-0.004	-0.006	-0.088**
	(0.021)	(0.021)	(0.013)	(0.014)	(0.008)	(0.011)	(0.023)
Percent of Respondents Belonging to Category	22.95	21.33	7.48	8.55	2.22	4.4	31.4

Notes: Multinomial logit estimation, average marginal effects. Dependent variable: stated partisanship on an eight-point scale with parties indicated in column headers, plus the category 'other party' (omitted in table). Data source: ifo Education Survey 2015. Regression weighted by survey weights. Robust standard errors in parentheses. Significance levels: *** p<0.01, ** p<0.05, * p<0.10.

Table 5.2: Summary Statistics and Balancing Tests

	Child Care Subsidy			Universal Student Aid		
	Control Mean (1)	Treatment Mean (2)	Diff. (3)	Control Mean (4)	Treatment Mean (5)	Diff. (6)
Highest Educational Attainment						
No Degree/Basic School Degree	0.405	0.390	-0.015	0.387	0.408	0.021
Middle School Degree	0.289	0.315	0.026	0.300	0.303	0.003
University Entrance Degree	0.306	0.295	-0.011	0.313	0.288	-0.025
Age	50.212	50.550	0.338	49.683	51.064	1.381
Monthly Household Income	2248	2263	14.451	2225	2286	60.796
Woman	0.519	0.511	-0.007	0.519	0.511	-0.009
Born in Germany	0.939	0.951	0.013	0.934	0.956	0.022**
Lives with Partner	0.610	0.600	-0.010	0.604	0.606	0.002
Lives in West Germany	0.801	0.795	-0.007	0.804	0.793	-0.011
City size >= 100,000	0.308	0.321	0.014	0.322	0.307	-0.015
Parent Status						
No Children	0.317	0.325	0.009	0.331	0.310	-0.021
Children of Age <= 18	0.293	0.298	0.005	0.280	0.311	0.031*
Children of Age > 18	0.391	0.377	-0.014	0.389	0.379	-0.010
At Least one Parent w. Univ. Degree	0.274	0.269	-0.005	0.261	0.283	0.022
Labor Market Participation						
Full-Time Employed	0.336	0.388	0.052**	0.371	0.351	-0.020
Part-Time Employed	0.125	0.128	0.003	0.119	0.134	0.015
Self-Employed	0.034	0.045	0.011	0.036	0.043	0.007
Unemployed	0.056	0.046	-0.010	0.052	0.050	-0.002
Housewife/Husband	0.073	0.062	-0.011	0.063	0.072	0.009
Retired or Ill	0.306	0.282	-0.025	0.295	0.293	-0.002
Student, Apprentice, In Training	0.070	0.049	-0.021	0.064	0.056	-0.008
Working in Education	0.108	0.107	-0.001	0.111	0.104	-0.006
Elections						
Votes Regularly	0.787	0.784	-0.003	0.784	0.788	0.004
Education Imp. for Voting Decision	0.731	0.721	-0.009	0.725	0.727	0.002
Partisanship						
CDU/CSU	0.233	0.226	-0.006	0.217	0.242	0.025
SPD	0.218	0.209	-0.009	0.212	0.215	0.003
Gruene	0.070	0.080	0.010	0.079	0.070	-0.009
Linke	0.090	0.080	-0.010	0.086	0.085	-0.001
FDP	0.025	0.020	-0.005	0.022	0.023	0.001
AfD	0.039	0.049	0.010	0.047	0.041	-0.005
No Partisanship	0.307	0.321	0.013	0.325	0.303	-0.021
Item Non-Response	0.002	0.011	0.009**	0.007	0.011	0.003
Observations	2,072	2,033		2,027	2,078	

Notes: Weighted group means. Diff.: difference in means between the control group and the respective treatment group. Significance levels of the difference stem from linear regressions of the background variables on the respective treatment dummy. Data source: ifo Education Survey 2015. Regressions weighted by survey weights. Significance levels: *** p<0.01, ** p<0.05, * p<0.10.

Table 5.3: Average Treatment Effects of Party-Position Information on Policy Preferences

	Support		Opposition	
	(1)	(2)	(3)	(4)
Experiment 1: Preferences for Child Care Subsidy				
Treatment	0.004	0.010	-0.009	-0.019
	(0.021)	(0.020)	(0.022)	(0.021)
Covariates	No	Yes	No	Yes
Control Mean	0.340	0.340	0.561	0.561
Observations	4092	3908	4092	3908
R-squared	0.000	0.061	0.000	0.074
Experiment 2: Preferences for Universal Student Aid				
Treatment	-0.001	-0.003	-0.032	-0.030
	(0.022)	(0.022)	(0.021)	(0.021)
Covariates	No	Yes	No	Yes
Control Mean	0.475	0.475	0.451	0.451
Observations	4,083	3,907	4,083	3,907
R-squared	0.000	0.040	0.001	0.037

Notes: OLS regressions. Treatment: Information on party positions. Control: No information. Dependent variable: (1)-(2): Dummy variable (1='strongly favor' or 'somewhat favor' the respective policy, 0 else); (3)-(4): Dummy variable (1='strongly oppose' or 'somewhat oppose' the respective policy, 0 else). Control mean: mean of the dummy variable for the control group. Covariates include: age, gender, born in Germany, living with partner, education, employment status, working in education sector, parent status, household income, West Germany, living in large city, parental education level. Data source: ifo Education Survey 2015. Regressions weighted by survey weights. Robust standard errors in parentheses. Significance levels: *** p<0.01, ** p<0.05, * p<0.10.

Table 5.4: Partisan-Specific Treatment Effects of Party-Position Information on Preferences for Child Care Subsidy

	Support for Child Care Subsidy		Opposition ag. Child Care Subsidy	
	(1)	(2)	(3)	(4)
Treatment x Party Favors	0.102**	0.120***	-0.109**	-0.128***
	(0.041)	(0.042)	(0.043)	(0.044)
Treatment x Party Neutral	-0.168	-0.179	0.150	0.157
	(0.120)	(0.120)	(0.124)	(0.129)
Treatment x Party Opposes	-0.003	0.005	-0.013	-0.028
	(0.032)	(0.031)	(0.034)	(0.032)
Treatment x Non-Partisan	-0.050	-0.051	0.044	0.049
	(0.039)	(0.038)	(0.039)	(0.038)
Party Favors	-0.078*	-0.059	0.109***	0.091**
	(0.040)	(0.041)	(0.041)	(0.041)
Party Neutral	-0.020	0.066	0.134	0.047
	(0.100)	(0.103)	(0.101)	(0.107)
Party Opposes	-0.093***	-0.068**	0.165***	0.146***
	(0.036)	(0.035)	(0.035)	(0.034)
Covariates	No	Yes	No	Yes
Control Mean	0.401	0.401	0.467	0.467
Observations	3,993	3,830	3,993	3,830
R-squared	0.009	0.071	0.020	0.089

Notes: OLS regressions. Treatment: information on party positions. Control: no information. Dependent variable: (1)–(2): dummy variable (1='strongly favor' or 'somewhat favor' child care subsidy, 0 else); (3)–(4): dummy variable (1='strongly oppose' or 'somewhat oppose' child care subsidy, 0 else). Non-partisan: respondents without long-term party partisanship; party favors (neutral, opposes): respondent supports party that favors (is neutral towards, opposes) child care subsidy. Control mean: mean of the dummy variable for non-partisans in the control group. Covariates include: age, gender, born in Germany, living with partner, education, employment status, working in education sector, parent status, household income, West Germany, living in large city, parental education level. Data source: ifo Education Survey 2015. Regressions weighted by survey weights. Robust standard errors in parentheses. Significance levels: *** p<0.01, ** p<0.05, * p<0.10.

Table 5.5: Partisan-Specific Treatment Effects of Party-Position Information on Preferences for Universal Student Aid

	Support for Universal Student Aid		Opposition against Universal Student Aid	
	(1)	(2)	(3)	(4)
Treatment x Party Favors	0.057	0.053	-0.076	-0.069
	(0.049)	(0.048)	(0.047)	(0.046)
Treatment x Party Neutral	-0.016	-0.011	-0.038	-0.037
	(0.048)	(0.049)	(0.048)	(0.049)
Treatment x Party Opposes	-0.074*	-0.067*	0.051	0.046
	(0.040)	(0.041)	(0.040)	(0.041)
Treatment x Non-Partisan	0.040	0.032	-0.081**	-0.075*
	(0.039)	(0.038)	(0.038)	(0.038)
Party Favors	0.085*	0.078*	-0.026	-0.028
	(0.045)	(0.045)	(0.045)	(0.045)
Party Neutral	0.040	0.072	0.026	-0.001
	(0.044)	(0.045)	(0.044)	(0.045)
Party Opposes	0.073*	0.086**	-0.026	-0.042
	(0.040)	(0.040)	(0.040)	(0.040)
Covariates	No	Yes	No	Yes
Control Mean	0.434	0.434	0.457	0.457
Observations	3,987	3,829	3,987	3,829
R-squared	0.007	0.047	0.007	0.043

Notes: OLS regressions. Treatment: information on party positions. Control: no information. Dependent variable: (1)–(2): dummy variable (1='strongly favor' or 'somewhat favor' universal student aid, 0 else); (3)–(4): dummy variable (1='strongly oppose' or 'somewhat oppose' universal student aid, 0 else). Non-partisan: respondents without long-term party partisanship; party favors (neutral, opposes): respondent supports party that favors (is neutral towards, opposes) universal student aid. Control mean: mean of the dummy variable for the non-partisans in the control group. Covariates include: age, gender, born in Germany, living with partner, education, employment status, working in education sector, parent status, household income, West Germany, living in large city, parental education level. Data source: ifo Education Survey 2015. Regressions weighted by survey weights. Robust standard errors in parentheses. Significance levels: *** p<0.01, ** p<0.05, * p<0.10.

Table 5.6: Heterogeneous Treatment Effects among Non-Partisans by Voting Behavior: Swing Voters

	Support		Opposition	
	(1)		(2)	
Experiment 1: Preferences for Child Care Subsidy				
Treatment	-0.141**	(0.056)	0.163***	(0.056)
Treatment x Regular Voter	0.159**	(0.074)	-0.196***	(0.073)
Regular Voter	-0.087	(0.055)	0.122**	(0.051)
Observations	1,254		1,254	
R-squared	0.095		0.113	
Wald Test: H_0 = no treatment effect for regular voters	0.018		-0.032	
Experiment 2: Preferences for Universal Student Aid				
Treatment	-0.050	(0.054)	-0.042	(0.053)
Treatment x Regular Voter	0.154**	(0.074)	-0.060	(0.075)
Regular voter	-0.003	(0.055)	0.026	(0.055)
Observations	1,254		1,254	
R-squared	0.079		0.052	
Wald Test: H_0 = no treatment effect for regular voters	0.104**		-0.102**	

Notes: OLS regressions. Sample: non-partisans (respondents without long-term party partisanship). Treatment: iformation on party positions. Control: no information. Dependent variable: (1): dummy variable (1='strongly favor' or 'somewhat favor' the respective policy, 0 else); (2): dummy variable (1='strongly oppose' or 'somewhat oppose' the respective policy, 0 else). Covariates included: age, gender, born in Germany, living with partner, education, employment status, working in education sector, parent status, household income, West Germany, living in large city, parental education level. Data source: ifo Education Survey 2015. Regressions weighted by survey weights. Robust standard errors in parentheses. Significance levels: *** p<0.01, ** p<0.05, * p<0.10.

Table 5.7: Heterogeneous Treatment Effects among Partisans of Parties Favoring Child Care Subsidy by Sociodemographic Subgroups

	All	Gender		Age		Income	
		Male	Female	Below Median	Above Median	Below Median	Above Median
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Treatment	0.103**	0.023	0.185***	0.035	0.133**	0.132**	0.078
	(0.040)	(0.055)	(0.055)	(0.059)	(0.052)	(0.058)	(0.053)
Covariates	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Control Mean	0.323	0.353	0.285	0.324	0.322	0.29	0.325
Observations	850	411	439	431	419	361	489
R-squared	0.069	0.099	0.116	0.127	0.093	0.129	0.088

	Employment Status		High School Degree		Child below 19		Educ. Policy Import. for Vote	
	Not Active	Active	No	Yes	No	Yes	No	Yes
	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)
Treatment	0.174***	-0.011	0.109**	0.080	0.123**	0.055	0.175**	0.089*
	(0.061)	(0.051)	(0.049)	(0.065)	(0.052)	(0.054)	(0.072)	(0.047)
Covariates	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Control Mean	0.297	0.355	0.346	0.256	0.291	0.372	0.262	0.343
Observations	322	528	579	271	407	426	186	664
R-squared	0.178	0.064	0.069	0.214	0.120	0.102	0.226	0.063

Notes: OLS regressions. Sample: respondents who support a party that favors child care subsidy. Treatment: information on party positions. Control: no information. Dependent variable: dummy variable (1='strongly favor' or 'somewhat favor' child care subsidy, 0 else). Control mean: mean of the dummy variable for the control group. Covariates include: age, gender, born in Germany, living with partner, education, employment status, working in education sector, parent status, household income, West Germany, living in large city, parental education level. Data source: ifo Education Survey 2015. Regressions weighted by survey weights. Robust standard errors in parentheses. Significance levels: *** p<0.01, ** p<0.05, * p<0.10.

Table 5.8: Heterogeneous Treatment Effects among Partisans of Parties Opposing Universal Student Aid by Sociodemographic Subgroups

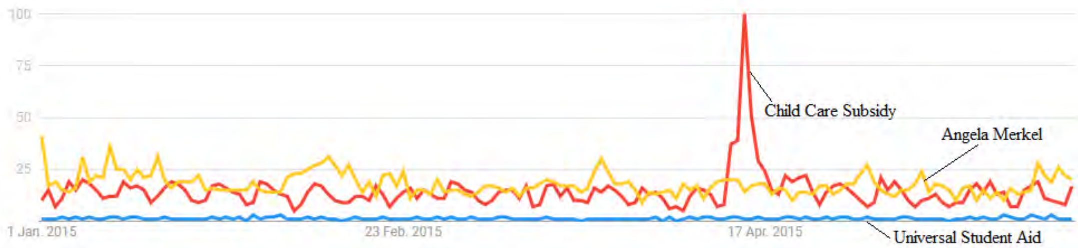
	All	Gender		Age		Income	
		Male	Female	Below Median	Above Median	Below Median	Above Median
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Treatment	-0.070*	-0.051	-0.043	-0.100*	-0.033	-0.055	-0.093*
	(0.039)	(0.051)	(0.055)	(0.057)	(0.050)	(0.058)	(0.050)
Covariates	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Control Mean	0.507	0.486	0.527	0.615	0.424	0.483	0.529
Observations	1,055	527	528	547	508	456	599
R-squared	0.070	0.121	0.095	0.115	0.034	0.091	0.104

	Employment Status		High School Degree		Child below 19		Educ. Policy Import. for Vote	
	Not Active	Active	No	Yes	No	Yes	No	Yes
	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)
Treatment	-0.050	-0.082*	-0.093**	-0.039	-0.041	-0.100**	-0.143*	-0.063
	(0.060)	(0.048)	(0.045)	(0.060)	(0.052)	(0.047)	(0.074)	(0.043)
Covariates	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Control Mean	0.430	0.587	0.537	0.437	0.445	0.637	0.479	0.516
Observations	386	669	719	336	487	549	225	829
R-squared	0.049	0.102	0.082	0.166	0.066	0.097	0.197	0.093

Notes: OLS regressions. Sample: respondents who support a party that opposes universal student aid. Treatment: information on party positions. Control: no information. Dependent variable: dummy variable (1='strongly favor' or 'somewhat favor' universal student aid, 0 else). Control mean: mean of the dummy variable for the control group. Covariates include: age, gender, born in Germany, living with partner, education, employment status, working in education sector, parent status, household income, West Germany, living in large city, parental education level. Data source: ifo Education Survey 2015. Regressions weighted by survey weights. Robust standard errors in parentheses. Significance levels: *** p<0.01, ** p<0.05, * p<0.10.

Appendix

Figure A5.1: Google Search Requests for Child Care Subsidy, Universal Student Aid, and the German Chancellor, First Half-Year 2015

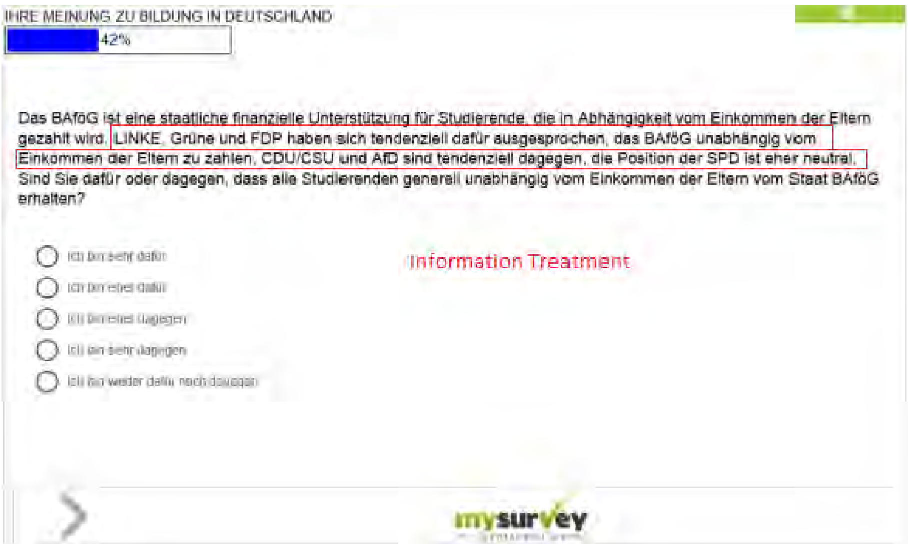


Notes: Google search requests for child care subsidy ('Betreuungsgeld' in red), universal student aid ('Elternunabhängiges BAfoeG' in blue), and the German chancellor ('Angela Merkel' in yellow) in Germany from January to June 2015. Frequencies depicted relative to the highest number of search requests (i.e., child care subsidy on 14 April 2015). The spike in search requests for child care subsidy coincides with the start of the lawsuit at the Federal Constitutional Court. Source: Google Trends (`www.google.com/trends` [accessed 12 December 2018]).

Figure A5.2: Screenshots of the Survey Questions

Panel A: Child Care Subsidy

Panel B: Universal Student Aid



Notes: Screen of respondents in the treatment group depicted. Information treatment is highlighted in red (survey respondents did not see red markings). Respondents in the control group saw a similar screen but without the respective information treatment.

Table A5.1: Wording of the Survey Questions

No.	Subgroup	Wording of Question	Type of question
Experiment 1: Preferences for Child Care Subsidy			
4	Control	The government pays parents who do not enroll their children aged 2 to 3 years in a childcare facility, but instead provide private home care, a child care subsidy in addition to the child benefits. Do you favor or oppose that parents receive a child care subsidy in addition to the child benefits?	Closed-ended with 5 answer categories: Strongly favor Somewhat Favor Neither Favor nor Oppose Somewhat Oppose
	Treatment	The government pays parents who do not enroll their children aged 2 to 3 years in a childcare facility, but instead provide private home care, a child care subsidy in addition to the child benefits. CDU/CSU tend to favor the child care subsidy, SPD, Linke, Gruene, and AfD tend to oppose it, the FDP is rather neutral. Do you favor or oppose that parents receive a child care subsidy in addition to the child benefits?	
Experiment 2: Preferences for Universal Student Aid			
28	Control	BAfoeG is federal financial aid for students which is paid contingent on parents' income. Do you favor or oppose that all students should generally receive BAfoeG by the government irrespective of parents' income?	Closed-ended with 5 answer categories: Strongly favor Somewhat Favor Neither Favor nor Oppose Somewhat Oppose
	Treatment	BAfoeG is federal financial aid for students which is paid contingent on parents' income. Linke, Gruene, and FDP tend to favor paying BAföG irrespective of parents' income, CDU/CSU and AfD tend to oppose it, the position of the SPD is rather neutral. Do you favor or oppose that all students should generally receive BAfoeG by the government irrespective of parents' income?	
Elicitation of Partisanship			
17*	All	Many people in Germany lean towards a particular political party in the long term, even if they occasionally also vote for another party. With which party do you sympathize in general?	Open-ended: CDU/CSU, SPD, Gruene, Linke, FDP, AFD, Another Party, namely…, None

Notes: Translation from the German original. No.: number of question in the ifo Education Survey 2015. Subgroup: specific control or treatment group that received the respective question. * This question was posed as the 17th question of the block on sociodemographic characteristics at the very end of the survey.

Table A5.2: Treatment Effects on Stated Partisanship

	CDU CSU (1)	SPD (2)	Gruene (3)	Linke (4)	FDP (5)	AfD (6)	Non Partisan (7)
Treatment (Child Care Subs.)	-0.012	0.011	-0.001	-0.017	0.001	0.010	0.006
	(0.024)	(0.026)	(0.018)	(0.015)	(0.010)	(0.013)	(0.029)
Treatment (Univ. Student Aid)	0.017	0.009	-0.020	-0.008	0.005	-0.006	-0.011
	(0.025)	(0.025)	(0.016)	(0.016)	(0.010)	(0.011)	(0.028)
Treatment (Child Care Subs.) x	0.024	-0.033	0.024	0.014	-0.011	-0.001	-0.009
Treatment (Univ. Student Aid)	(0.035)	(0.036)	(0.025)	(0.023)	(0.013)	(0.017)	(0.040)
Covariates	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Control Mean	0.222	0.210	0.081	0.095	0.022	0.040	0.319
Observations	3903	3903	3903	3903	3903	3903	3903
R-squared	0.048	0.038	0.049	0.050	0.017	0.020	0.060
p-value: Joint Significance All Depicted Coefficients	0.329	0.752	0.395	0.736	0.775	0.740	0.876

Notes: OLS regressions. Treatment: information on party positions (in the experiment on preferences for child care subsidy and universal student aid, respectively). Control: no information in none of the experiments. Dependent variable: (1)–(2): dummy variable (1='strongly favor' or 'somewhat favor' the respective policy, 0 else); (3)–(4): dummy variable (1='strongly oppose' or 'somewhat oppose' the respective policy, 0 else). Control mean: mean of the dummy variable for the control group in both experiments. Covariates include: age, gender, born in Germany, living with partner, education, employment status, working in education sector, parent status, household income, West Germany, living in large city, parental education level. Data source: ifo Education Survey 2015. Regressions weighted by survey weights. Robust standard errors in parentheses. Significance levels: *** p<0.01, ** p<0.05, * p<0.10.

Table A5.3: Partisan-Specific Treatment Effects: All Answer Categories

	Strongly favor		Somewhat favor		Neither favor nor oppose		Somewhat oppose		Strongly oppose	
	(1)		(2)		(3)		(4)		(5)	
Experiment 1: Preferences for Child Care Subsidy										
Treatment x Party Favors	0.051*	(0.028)	0.069*	(0.038)	0.008	(0.030)	-0.023	(0.038)	-0.105***	(0.039)
Treatment x Party Neutral	-0.058	(0.079)	-0.121	(0.100)	0.022	(0.038)	0.123	(0.146)	0.033	(0.130)
Treatment x Party Opposes	0.013	(0.022)	-0.008	(0.027)	0.024	(0.019)	-0.013	(0.030)	-0.015	(0.028)
Treatment x Non-Partisan	-0.035	(0.029)	-0.016	(0.035)	0.002	(0.027)	-0.004	(0.031)	0.053*	(0.031)
Party Favors	-0.037	(0.030)	-0.022	(0.036)	-0.032	(0.031)	0.056	(0.036)	0.035	(0.038)
Party Neutral	0.046	(0.069)	0.019	(0.089)	-0.112***	(0.028)	0.025	(0.114)	0.021	(0.095)
Party Opposes	-0.026	(0.028)	-0.042	(0.032)	-0.078***	(0.021)	0.055*	(0.029)	0.092***	(0.028)
Control Mean	0.160		0.241		0.132		0.212		0.255	
Experiment 2: Preferences for Universal Student Aid										
Treatment x Party Favors	0.027	(0.045)	0.026	(0.043)	0.017	(0.028)	-0.054	(0.041)	-0.015	(0.033)
Treatment x Party Neutral	-0.021	(0.035)	0.010	(0.045)	0.048*	(0.025)	-0.048	(0.043)	0.011	(0.040)
Treatment x Party Opposes	-0.040	(0.029)	-0.027	(0.038)	0.021	(0.023)	0.063*	(0.034)	-0.017	(0.034)
Treatment x Non-Partisan	0.016	(0.030)	0.016	(0.036)	0.043*	(0.025)	-0.081**	(0.033)	0.007	(0.029)
Party Favors	0.082**	(0.038)	-0.004	(0.040)	-0.050**	(0.024)	-0.017	(0.042)	-0.012	(0.032)
Party Neutral	0.021	(0.036)	0.051	(0.042)	-0.071***	(0.021)	-0.015	(0.041)	0.014	(0.033)
Party Opposes	0.040	(0.031)	0.046	(0.038)	-0.045**	(0.022)	-0.076**	(0.035)	0.034	(0.032)
Control Mean	0.154		0.280		0.109		0.285		0.172	
Covariates	Yes		Yes		Yes		Yes		Yes	

Notes: OLS regressions. Treatment: information on party positions. Control: no information. Dependent variable: dummy variables (1=answer category given in respective table header, 0 else). Non-partisan: respondents without long-term party partisanship; party favors (neutral, opposes): respondent supports party that favors (is neutral towards, opposes) the respective policy Control mean: mean of the dummy variable for non-partisans in the control group. Covariates include: age, gender, born in Germany, living with partner, education, employment status, working in education sector, parent status, household income, West Germany, living in large city, parental education level. Data source: ifo Education Survey 2015. Regressions weighted by survey weights. Robust standard errors in parentheses. Significance levels: *** p<0.01, ** p<0.05, * p<0.10.

Table A5.4: Heterogeneous Treatment Effects by Party-Specific Partisanship

	Support				Opposition			
	(1)		(2)		(3)		(4)	
Experiment 1: Preferences for child care subsidy								
Treatment x CDU/CSU	0.102**	0.041	0.120***	(0.042)	-0.109**	(0.043)	-0.128***	(0.044)
Treatment x SPD	0.034	(0.046)	0.042	(0.045)	-0.032	(0.048)	-0.053	(0.048)
Treatment x Gruene	-0.039	(0.076)	-0.029	(0.072)	-0.020	(0.082)	-0.037	(0.076)
Treatment x Linke	-0.099	(0.067)	-0.078	(0.063)	0.087	(0.068)	0.066	(0.063)
Treatment x FDP	-0.168	(0.120)	-0.179	(0.120)	0.150	(0.124)	0.158	(0.129)
Treatment x AfD	0.060	(0.103)	0.049	(0.082)	-0.093	(0.104)	-0.075	(0.081)
Treatment x Non-partisan	-0.050	(0.039)	-0.051	(0.038)	0.044	(0.039)	0.048	(0.038)
Control Mean	0.401		0.401		0.467		0.467	
Experiment 2: Preferences for universal student aid								
Treatment x CDU/CSU	-0.058	(0.044)	-0.049	(0.045)	0.024	(0.044)	0.018	(0.044)
Treatment x SPD	-0.016	(0.048)	-0.010	(0.049)	-0.038	(0.048)	-0.038	(0.049)
Treatment x Gruene	0.138*	(0.081)	0.140*	(0.079)	-0.174**	(0.074)	-0.177**	(0.071)
Treatment x Linke	-0.004	(0.066)	-0.012	(0.065)	-0.020	(0.065)	-0.007	(0.063)
Treatment x FDP	0.028	(0.143)	0.011	(0.142)	0.021	(0.146)	0.045	(0.149)
Treatment x AfD	-0.172*	(0.099)	-0.182*	(0.095)	0.211**	(0.099)	0.216**	(0.096)
Treatment x Non-partisan	0.040	(0.039)	0.032	(0.038)	-0.081**	(0.038)	-0.075*	(0.038)
Control Mean	0.434		0.434		0.457		0.457	
Covariates	No		Yes		No		Yes	

Notes: OLS regressions. Treatment: information on party positions. Control: no information. Dependent variable: (1)–(2): dummy variable (1=‘strongly favor’ or ‘somewhat favor’ the respective policy, 0 else); (3)–(4): dummy variable (1=‘strongly oppose’ or ‘somewhat oppose’ the respective policy, 0 else). Control mean: mean of the dummy variable for nonpartisans in the control group. Covariates included: age, gender, born in Germany, living with partner, education, employment status, working in education sector, parent status, household income, West Germany, living in large city, parental education level. Data source: ifo Education Survey 2015. Regressions weighted by survey weights. Robust standard errors in parentheses. Significance levels: *** p<0.01, ** p<0.05, * p<0.10.

6 Incentives, Search Engines, and the Elicitation of Subjective Beliefs: Evidence from Representative Online Survey Experiments*

6.1 Introduction

Measuring people's subjective beliefs about economic facts is essential for understanding economic behavior and choices. For example, beliefs about earnings returns to education shape educational decisions (e.g., Boneva and Rauh, 2017; Bleemer and Zafar, 2018; Delavande and Zafar, 2019), beliefs about the size of government affect support for public spending (e.g., Lergetporer et al., 2018a; Roth et al., 2020), and beliefs about societal inequality are closely linked to preferences for redistribution (e.g., Kuziemko et al., 2015; Alesina et al., 2018b).[1] To elicit such beliefs, economists often rely on survey questions which do not provide respondents with incentives for accurate answers. This raises concerns of systematic biases in unincentivized belief measures that might stem from lack of cognitive effort invested in truthful reporting or from socially desirable and self-serving answering behavior. Biased belief measures are not only uninformative about the target population's actual beliefs, but they can also bias estimates of the investigated relationship between subjective beliefs and economic decisions (e.g., Bertrand and Mullainathan, 2001; Benítez-Silva et al., 2004). While the effects of incentivizing belief accuracy have been studied extensively in the context of economic games conducted in the laboratory (e.g., Schlag et al., 2015; Trautmann and van de Kuilen, 2015), little is known about incentive effects on survey-based belief measures about economic facts.

In this chapter, we investigate experimentally whether incentive provision is necessary to retain unbiased measures of subjective beliefs in representative online surveys. Our two main experiments provide incentives for correct beliefs about two relevant economic facts. Results show that incentivization improves the accuracy of stated beliefs in one case but not in the other. Inspection of response spikes indicates that incentive effects may be related to the usage of search engines. Therefore, we conduct a third experiment that explicitly encourages online-search activity without providing incentives. The encouragement treatment produces very similar effects to the incentive treatment, suggesting that improved beliefs in the incen-

* This chapter is joint work with Philipp Lergetporer, Katharina Werner, and Ludger Woessmann. It is based on the paper 'Incentives, Search Engines, and the Elicitation of Subjective Beliefs: Evidence from Representative Online Survey Experiments', *Journal of Econometrics*, fourthcoming.

1 Other economically important dimensions of people's subjective beliefs include, for instance, beliefs about inflation (e.g., Armantier et al., 2013, 2016) and beliefs about the extent and consequences of immigration (e.g., Alesina et al., 2018a; Grigorieff et al., 2020; Haaland and Roth, 2020).

tive treatment mainly reflect altered online-search behavior. In a fourth experiment, we show that inducing experimenter-demand effects does not affect stated beliefs, suggesting that the incentive effect is unlikely to reflect experimenter-demand effects. Overall, our results highlight that effects on online-search activity should be carefully considered when researchers design survey instruments for belief elicitation in settings where these effects are undesirable.

We conduct our two incentive experiments in a representative online survey in Germany (N>3,600). The elicited beliefs about economic facts refer to respondents' stated beliefs about (i) average earnings by professional degree and (ii) average public school spending.[2] In both experiments, we offer respondents in a randomly selected treatment group a monetary reward if their answer to the belief question is close to the true value. The control group answers the same questions without any incentives. If unincentivized belief questions are in fact biased (e.g., by lack of effort or self-serving answering behavior), we expect incentives to improve the accuracy of stated beliefs.

The first experiment indicates that incentivizing beliefs hardly affects stated beliefs about earnings by professional degree. We elicit beliefs about net average monthly earnings of full-time employed persons without any professional degree and of persons with a university degree. In the unincentivized control group, the 10–90 percentile range of earnings beliefs without a degree is 800 Euro to 1,500 Euro (true value: 1,400 Euro). For earnings beliefs about university graduates, the range is 1,500 Euro to 4,000 Euro (true value: 2,750 Euro). The majority of respondents underestimate current earnings levels: 82 percent of unincentivized beliefs about earnings without a degree and 57 percent of those with a university degree are below the true value. We find almost no incentive effects on these earnings beliefs: While incentives marginally increase beliefs about average earnings with a university degree ($p<0.1$), they do not affect any of several measures of belief accuracy or respondents' confidence about their beliefs. Incentivized respondents take more time to answer the belief question, which suggests that they put more effort into reporting a meaningful answer. In this sense, incentives could reduce noise in the data and lead to estimates that are closer to the true subjective belief of respondents. At the same time, treatment-induced increases in time spent on the question do not translate into beliefs that are closer to the correct value. Consistently, we also find no incentive effects on beliefs re-elicited in a follow-up survey about two weeks after the main survey.

[2] Beliefs about earnings by professional degree are central to the economics literature on educational decisions (e.g., Arcidiacono, 2004; Arcidiacono et al., 2012; Hoxby and Turner, 2013; Kaufmann, 2014; Belfield et al., 2020). While this strand of research is often particularly interested in people's individual earnings expectations (which cannot be verified and therefore not incentivized; see Manski, 2004), beliefs about population averages, the focus of this chapter, have also been featured prominently in the literature (see, Wiswall and Zafar, 2015; Bleemer and Zafar, 2018; Lergetporer et al., 2018b). Beliefs about public education spending are important for shaping public budgets since they determine the electorate's preferences for the size of government (e.g., Schueler and West, 2016; Lergetporer et al., 2018a).

The second experiment shows that, by contrast, the incentive treatment significantly improves the accuracy of stated beliefs about average annual public school spending per student. In the control group, school-spending beliefs are strongly dispersed with a 10–90 percentile range of 50 Euro to 15,000 Euro (true value: 7,400 Euro). The majority of unincentivized respondents (78 percent) underestimate the actual spending level. In contrast to the experiment on earnings beliefs, incentivization improves the accuracy of school-spending beliefs in the main survey. It also raises respondents' confidence about the accuracy of their beliefs. These effects persist in the follow-up survey about two weeks after the main survey. Again, incentive provision increases response time in the main survey.

To reconcile the different effects of incentives in the two experiments, we present further analyses on potential treatment-effect mechanisms. Closer inspection of the distribution of beliefs on public school spending shows that incentivizing belief accuracy increases the frequency of stated beliefs which correspond to top-listed online-search-engine results. Since information on average public school spending (which is published regularly by the German Statistical Office) is easier to find online than information on average earnings by professional degree (which we calculated ourselves for the purposes of this chapter), we suggest that differences in incentive effects across belief domains might be due to differences in the use of online-search engines.

To further investigate the role of search-engine usage, we conduct an additional experiment in a new representative sample (N>4,000). Instead of providing monetary incentives, in this experiment we encourage a randomly selected group of respondents to use online-search engines to inform their beliefs about public school spending. Strikingly, this encouragement produces treatment effects which are very similar in magnitude and significance to the effects of incentivizing belief accuracy. This similarity strongly suggests that our incentives for accurate beliefs about public school spending improve beliefs through altering respondents' online-search activity.

The investigation yields two main conclusions. First, the fact that monetary incentives do not improve belief accuracy about average earnings (i.e., information not easily found on the internet) suggests that the lack of incentives in commonly used survey-based belief measures does not yield reporting bias in our context.[3] Second, incentivizing belief accuracy in online surveys can induce respondents to use online-search engines for the purpose of improving their answers. While incentive effects on reported beliefs have been studied extensively in closely controlled laboratory environments, this result highlights potential limitations of applying incentives to less-controlled contexts such as online surveys. Since researchers are usually interested in unbiased measures of prior subjective beliefs about economic facts and not in beliefs updated after consulting external sources such as online-search engines, our

[3] Of course, the extent to which unincentivized belief measures about economic facts are biased or not may rest on the properties of the elicited beliefs, e.g., on whether objective information is readily available online or whether respondents have well-formed priors.

results underline that researchers need to consider unintended effects of incentivizing beliefs when designing survey instruments. At the same time, using monetary incentives or encouraging online-search activity when eliciting beliefs might be interesting for researchers who seek to shift survey respondents' information sets. We provide a discussion of the advantages and disadvantages of incentivizing beliefs in the conclusion of the chapter.

Finally, we present results of an additional experiment to assess whether the effects of the incentive treatment reflect experimenter-demand effects, as opposed to genuine incentive effects.[4] In a randomly selected treatment group, we emphasize that it is important that respondents provide a correct answer to the belief question about public school spending. While this wording explicitly states the experimenter demand, it does not affect stated beliefs or response time. This result suggests that the reported incentive effects are due to respondents adapting their answering behavior to earn the incentives, and not because the offer of incentives contains information about the importance of accurate beliefs to the experimenter. This finding is consistent with the fact that incentive-treatment effects in the school-spending experiment persist over a two-week period, which is often interpreted as evidence for the absence of reporting biases such as priming effects or experimenter-demand effects (e.g., Cavallo et al., 2017; Haaland and Roth, 2020).

This chapter contributes to several strands of the existing literature. At the most basic level, it adds to the large literature in economics which studies people's subjective beliefs about economic outcomes in large-scale surveys. One strand of this literature focuses on beliefs about future events, such as inflation expectations (e.g., Armantier et al., 2016; Cavallo et al., 2017; Coibion et al., 2018; Manski, 2018), future home prices (e.g., Fuster et al., 2020), or educational expectations (e.g., Zafar, 2011; Stinebrickner and Stinebrickner, 2013; Boneva and Rauh, 2017, 2018; Delavande and Zafar, 2019; Attanasio et al., 2020). Another strand studies beliefs about realized (contemporaneous or past) economic outcomes, such as beliefs about existing inequality (e.g., Alesina and La Ferrara, 2005; Karadja et al., 2016; Alesina et al., 2018b; Lergetporer et al., 2020), immigration (e.g., Alesina et al., 2018a; Grigorieff et al., 2020; Haaland and Roth, 2020), or the size of government (e.g., Lergetporer et al., 2018a; Roth et al., 2020).[5] This literature mostly relies on unincentivized questions, often implemented in online surveys, to elicit respondents' beliefs. While some recent papers incentivize correct answers to belief questions (e.g., Roth and Wohlfart, 2019; Fuster et al., 2020; Grigorieff et al., 2020), the

[4] In our context, one concern might be that providing incentives for accurate beliefs signals to respondents that the surveyor's aim is to show that incentives in fact improve beliefs. Consequently, respondents might put more effort in providing a correct answer only because they want to please the experimenter. Similarly, the incentive treatment may induce respondents to state second-order beliefs about what they think the surveyor would want them to answer rather than their own true belief. Our experimental design to measure experimenter-demand effects is similar to Mummolo and Peterson (2019) and de Quidt et al. (2018).

[5] In addition, several studies investigate the connection between beliefs about future events and realized outcomes, e.g., by shifting beliefs about the future through providing information about contemporaneous (or past) outcomes (e.g., Armantier et al., 2013; Wiswall and Zafar, 2015; Coibion et al., 2018; Armona et al., 2019). Relatedly, Manski (2004) provides an exhaustive discussion of eliciting subjective expectations in surveys.

effects of such incentives on stated beliefs have hardly been studied experimentally. Our first contribution is therefore to provide a systematic assessment of the effects of incentivizing belief accuracy in a large-scale, representative online survey.

Relatedly, this chapter is part of the smaller literature on the effectiveness of different belief-elicitation methods (e.g., Delavande et al., 2011; Ansolabehere et al., 2013). Within this literature, a few papers compare incentivized versus unincentivized belief measures (e.g., Roth and Wohlfart, 2019).[6] In particular, Bullock et al. (2015) and Prior et al. (2015) show that monetary incentives for accurate answers can reduce partisan bias in stated beliefs about political facts.[7] In some cases, researchers seem to anticipate online-search activity and adopt measures to avoid this effect of incentivizing belief accuracy in online surveys. For instance, Bullock et al. (2015) and Prior et al. (2015) restrict response time in order to keep respondents from looking up answers using outside references.[8] In contrast to the aforementioned papers, we explicitly study the extent to which incentive provision affects the use of outside resources—in particular online-search activity. We therefore contribute to the literature on measuring subjective beliefs by directly investigating online-search activity and experimenter-demand effects as two possible mechanisms of how incentives affect stated beliefs.

Finally, this chapter contributes to the methodological literature on the opportunities and challenges of conducting experiments on the internet. Comparing experimental choices elicited in the laboratory and online, recent studies suggest that differences across these modes are modest (e.g., Clifford and Jerit, 2014; Arechar et al., 2018). Our finding that online-survey

[6] Note that Roth and Wohlfart (2019) do not focus on the methodological question whether incentives affect belief accuracy. Consequently, they do not implement their incentives treatment in their main representative survey, but rather in a complementary sample of MTurk workers to show that stated recession expectations are unaffected by providing incentives for correct answers.

[7] Relatedly, there is an extensive literature using laboratory experiments to study effects of incentives on university-student participants' beliefs in economic games (see Schotter and Trevino, 2014; Schlag et al., 2015, for a review). Broadly speaking, this literature has investigated three sets of research questions. First, similar to our research question, some studies have investigated whether incentivizing belief accuracy (versus providing no incentives) affects stated beliefs (e.g., Gächter and Renner, 2010; Wang, 2011). The evidence seems mixed and provides no clear advantage of incentivized elicitation methods over non-incentivized belief elicitation (see Trautmann and van de Kuilen, 2015, for a discussion). The second strand of research compares different incentivization methods with different theoretical properties, such as the linear scoring rule or the quadratic scoring rule. While there is no consistent evidence that one incentivization method clearly dominates the others in terms of belief accuracy, methods that rely on the assumption of risk neutrality have been shown to pose the risk of producing biased belief data when not correcting for risk attitudes (e.g., Schlag et al., 2015). A final strand of research investigates whether incentivizing beliefs changes subjects' experimental choices—for instance by using stated beliefs as a hedging device against adverse outcomes of other decisions in the experiment (e.g., Blanco et al., 2010). While hedging problems do exist in some situations, they are unlikely to produce major confounds as long as the hedging opportunities are relatively opaque.

[8] Bullock et al. (2015) restrict response time to 20 seconds and Prior et al. (2015) to between 45 and 60 seconds. While restricting response time might be a pragmatic approach to avoid respondents consulting outside sources in some cases, inducing time pressure can have additional effects on answering behavior that might be unattractive in some settings (e.g., Karagözoğlu and Kocher, 2019; Kocher et al., 2019). Testing the effects of time pressure on belief elicitation in surveys is an interesting avenue for future research.

participants likely use search engines to improve stated beliefs points to thus far underappreciated challenges when conducting experiments in environments where the experimenter has limited control over the experimental setting.

This chapter is structured as follows. Section 6.2 introduces our survey and the experimental setup of the incentive experiments. Section 6.3 shows the experimental results on belief incentivization. Section 6.4 presents the experiment on encouraging online-search activity. Section 6.5 provides experimental evidence on the role of experimenter-demand effects. Section 6.6 concludes.

6.2 Data and Empirical Strategy

This section describes the survey, the experimental design, and the econometric model and reports results of randomization tests.

6.2.1 The Survey

Our research is based on data from the ifo Education Survey, an annual representative opinion survey on education policy that we conduct in Germany. The main experiments of this chapter were implemented in the 2017 survey wave and focus on 3,696 respondents who were sampled and surveyed via an online platform.[9] The sample is weighted to match official statistics with respect to age, gender, federal state, school degree, and municipality size. The survey contains a total of 34 questions on different topics of education policy and collects information on respondents' sociodemographic characteristics (see Table 6.1). Median completion time is 17 minutes. Item non-response is very low, at less than one percent for the belief questions which measure our outcomes of interest. As we show below, treatment status does not predict item non-response on the outcome variables.

Sampling and polling was carried out by Kantar Public, a renowned survey company, in April and May 2017. As is standard for online surveys, respondents answer the survey autonomously on their own digital devices. For all respondents, survey completion is incentivized with 75 tokens (worth about 0.70 Euro in total), which they can exchange for gift vouchers of well-known online retailers.[10]

[9] The overall sample comprised 382 additional respondents who do not use the internet and were therefore polled at their homes by trained interviewers. As it was not possible to incentivize their answers, we exclude these respondents from our analysis. Grewenig et al. (2018) show that our approach to weight online-survey observations to match the characteristics of the entire population yields representativeness for the entire population. Our results are qualitatively identical when using unweighted data (results available upon request).

[10] It is worth noting that the compensation for survey participation corresponds to the standard rate that is offered by the polling firm. The hourly wage equivalent of the compensation is relatively low at about 2.50 Euro, which suggests that the monetary compensation is not the only factor that motivates respondents to participate in the survey. Instead, intrinsic motivation or 'gamification'—a phenomenon where respondents value tokens

To investigate the persistence of potential treatment effects beyond the immediate survey horizon, respondents were invited to participate in a follow-up survey about two weeks after the main survey. The follow-up survey re-elicits some outcomes, but does not contain any experimental treatments and does not incentivize correct answers. Overall, 2,535 respondents (69 percent) participated in the follow-up survey.[11] The median lag to the main survey was 12 days with a range from 5 to 41 days.

6.2.2 Experimental Design

To test whether monetary incentives affect stated beliefs about economic facts, we devise two experiments that incentivize belief accuracy in randomly selected treatment groups. One experiment focuses on beliefs about average earnings by professional degree, the other experiment focuses on beliefs about average public school spending. All respondents participated in both experiments consecutively, and randomization was independent across experiments.[12]

The Incentive Treatment

In both experiments, we randomly assigned respondents to a treatment group that is offered additional tokens for accurate answers before eliciting their beliefs in the same way as in the unincentivized control group. In the incentive treatment, respondents can earn 35 tokens—in addition to the 75 tokens for participation—if their answer is 'roughly correct'.[13] We define answers as 'roughly correct' if they fall into the intervals around the true values that contain 50 percent of respondents. To calculate these intervals, we used data from the 2016 wave of the ifo Education Survey where we asked the same (unincentivized) belief questions.[14]

more than their monetary equivalent (e.g., Puleston, 2011; Keusch and Zhang, 2015)—might also foster survey participation.

[11] The take-up rate is relatively high compared to other recent studies. For instance, take-up in follow-up surveys was 14 percent in Kuziemko et al. (2015) and 24 percent in Alesina et al. (2018b).

[12] In the questionnaire, the experiment on school-spending beliefs preceded the experiment on earnings beliefs. One potential concern with running both experiments with all respondents is that incentive provision in the first experiment might affect answering behavior in the subsequent experiment. Reassuringly, treatment effects in the earnings experiment are robust to restricting the sample to those respondents who were assigned to the control group in the school-spending experiment. In our analysis, all regressions for the second experiment (beliefs on earnings) control for treatment status in the first experiment (beliefs on school spending), but results are qualitatively unchanged if the control is excluded (results available upon request).

[13] The 35 tokens that can be earned for a correct answer correspond to a monetary reward of about 0.33 Euro. Note that this amount is at the lower bound—but still within the range—of the incentives offered in comparable studies. For instance, in one experiment, Bullock et al. (2015) vary the amount paid for a correct answer between 0.1 Dollar, 0.25 Dollar, 0.5 Dollar, 0.75 Dollar, and 1 Dollar. Importantly, they find that incentives of all these sizes decrease the partisan gap in factual beliefs (their main outcome of interest), while the difference in treatment effects between the lowest and highest incentive payment is only marginally significant ($p<0.1$). However, our results likely constitute a lower bound of the effects of monetary incentives on the consultation of external sources.

[14] While the ifo Education Survey is a repeated cross-sectional survey, some respondents participate in more than one survey wave. About 13 percent of respondents in the 2017 survey wave also participated in the 2016

From a technical viewpoint, our approach to reward beliefs that correspond to the true value with a fixed amount of tokens is equivalent to eliciting the mode of the distribution (Hurley and Shogren, 2005). Note that this method is 'truth-telling' since it is optimal for the respondent to report the highest-probability value as her belief, independent of her risk preferences (Schlag et al., 2015). Note, however, that we elicit beliefs about continuous variables, which is why we reward answers within an interval around the true value. Theoretically, this opens up the possibility that truth-telling is no longer the dominant strategy for certain belief-preference constellations.[15] Laboratory experiments with university students usually use more sophisticated methods to incentivize beliefs, such as the quadratic scoring rule (e.g., Trautmann and van de Kuilen, 2015). Since our representative sample comprises respondents from various educational backgrounds, including 36 percent with basic school degree or less (see Table 6.1), we deliberately opted for a much simpler incentive scheme to minimize the possibility of comprehension problems.[16]

In the incentive treatment, we use the following wording to inform about the possibility to earn additional tokens for accurate answers: *'On the following screen, you will be asked another guess question. If your guess is roughly correct, you will receive an additional reward worth about half of the reward you get for participating in this survey.'*[17] To maximize salience of the treatment, treated respondents are also reminded of the incentives on the subsequent belief-elicitation screen. We kept the information about the incentive short to convey the idea that belief accuracy is rewarded as clearly and simply as possible. At the same time, treated respondents had the possibility to retrieve more information about the incentive scheme by clicking on an information button, which 28 percent (32 percent) chose to do in the earnings experiment (school-spending experiment).[18] Appendix Table A6.2 shows that older

survey wave. Excluding those respondents in the 2017 survey who also participated in the 2016 survey wave from the analysis yields virtually identical results (available upon request).

[15] For instance, a risk-averse respondent who thinks that the true value is X with probability ρ>50% and Y with probability (1-ρ) would report X as her belief if the incentive scheme only pays for reporting exactly the true value. Instead, incentivizing beliefs within an interval around the true value might make it optimal for the respondent to report a value between X and Y in order to maximize the chance of being covered in the interval of both possible realizations of the true value. While this strategic reporting is theoretically possible, note that it requires a high degree of sophistication and specific constellations of beliefs about the likely true values and the size of the incentivized range. While we consider it unlikely that this sort of reporting bias is a concern in our representative sample, extending our experimental design towards investigating how alternative incentive schemes with different theoretical properties affect beliefs and online-search activity is certainly an interesting area for future research.

[16] The extent to which scoring rules make truthful reporting optimal usually depends on respondents' risk preferences (e.g., Trautmann and van de Kuilen, 2015). Interestingly, our heterogeneity analysis in section 6.3 reveals that incentive-treatment effects are not heterogenous with respect to respondents' general willingness to take risk.

[17] Appendix Table A6.1 presents the wording of all questionnaire items used in this chapter.

[18] The following text appeared upon clicking on the information button: *'You will receive an additional reward worth about half of the reward you get for participating in this survey if your answer lies within the best 50 percent of answers from the previous year.'*

respondents are more likely to acquire this information, and that information acquisition varies with parental education and employment status.

Our data suggest that the incentive treatments successfully increased respondents' effort to give a correct answer. This is not only apparent from the positive treatment effects on response time (see section 6.3), but also from descriptive survey evidence. We asked respondents in the treatment group whether the incentive increased their effort to give a correct answer. As Figure 6.1 shows, 58 percent 'strongly' or 'somewhat' agree with the statement that the prospect of earning more tokens encouraged them to put more effort in their answers. Appendix Table A6.3 depicts the relationship between these survey answers and respondents' sociodemographic characteristics. Younger respondents, males, those living together with a partner, those who do not have children, those without university-educated parents, and more risk-tolerant respondents are more likely to report that they are motivated by the incentives. Interestingly, respondents who acquired additional information about the scoring rule and those who were randomized into the incentive treatment in both (rather than just one) experiments do not state higher levels of incentive-induced motivation (result available upon request).

While the existing literature usually motivates the usage of incentives by the notion that incentives increase respondents' effort to report a belief truthfully, what is meant by effort is not always well defined.[19] In our context, it is important to distinguish between two types of effort: the *internal* cognitive effort that respondents invest in retrieving information that they stored in their memory (e.g., Zimmermann, 2020) and the *external* effort to consult other sources to inform stated beliefs. When interested in prior subjective beliefs about specific facts, researchers in many applications would like to see high levels of internal effort but would like to prevent external effort. That is, depending on the research question, some forms of effort may be desirable whereas others may be not. While investigating the exact way in which incentives affect different forms of effort is beyond the scope of this chapter, our analysis of response spikes that correspond to Google results in section 6.3.3 allows us to identify online-search activity as one specific—and empirically relevant—form of consulting external sources to inform beliefs.

Eliciting Beliefs

In the earnings-beliefs experiment, we use the following question to elicit beliefs: *'Persons with a professional degree (apprenticeship) currently earn on average 1,850 Euro net per month (full-time position). What is your best guess, how much do the following groups with lower resp. higher educational attainment earn on average?'* Respondents are asked to enter their earnings beliefs about persons without a professional degree and about persons with a university degree into open numeric fields. Based on calculations using the German Microcensus,

[19] The second key motivation for incentivizing beliefs in the literature is that incentives might mitigate biases in reporting beliefs, which might stem from belief-based utility (Zimmermann, 2020), political-identity considerations (Prior et al., 2015), or social-desirability bias.

the median net monthly earnings of full-time employed persons in these groups are 1,400 Euro and 2,750 Euro, respectively.[20] Following our above definition of 'roughly correct' answers, respondents in the treatment group whose stated earnings beliefs for those without a professional degree was between 1,008 Euro and 1,792 Euro *and* whose earnings beliefs for those with a university degree was between 1,980 Euro and 3,520 Euro received the incentive payment. This applied to 46 percent of respondents in the treatment group.

The experiment on beliefs about average public school spending follows the same structure. The question reads as follows: *'Based on your best guess, what is the average amount of money spent each year for a child in public schools in Germany?'* According to official statistics of the Federal Statistical Office, the current spending level is 7,400 Euro (Statistisches Bundesamt, 2016). 50 percent of respondents in the treatment group state a belief within the correct range (between 1,332 Euro and 13,468 Euro) and are therefore paid the incentive.

After each belief question, we elicit respondents' confidence that the stated belief is close to correct on a seven-point scale (from 1 = 'very unsure' to 7 = 'very sure'). To avoid potential treatment-effect spillovers across experiments, respondents are informed about whether their respective belief falls into the incentivized range only at the very end of the survey.[21]

6.2.3 Econometric Model

We estimate the effects of the incentive treatment on stated beliefs using the following regression model:

$$y_i = \alpha + \beta_1 Incentive_i + \epsilon_i \tag{6.1}$$

where y_i is the outcome variable of interest, $Incentive_i$ is a dummy variable indicating whether respondent i received the incentive treatment, and ϵ_i is the error term. Randomization ensures that the average treatment effect, captured by coefficient β_1, is identified.

As expected, adding the rich set of sociodemographic characteristics (as shown in Table 6.1) as control variables to the analyses does not substantively alter the estimated treatment coefficients. For ease of exposition, we therefore only report models without covariates throughout the chapter. Detailed results of models with control variables are available upon request.

[20] Research Data Centres of the Federal Statistical Office and the statistical offices of the Laender, Microcensus, census year 2012.

[21] Note that respondents learned about whether they receive a payment for accurate beliefs after the main survey, but before the follow-up survey. Therefore, while treatment effects in the main survey can be interpreted as the pure causal effect of providing incentives on belief accuracy, treatment effects in the follow-up survey can reflect a combination of the pure effect of the incentives provided in the main survey and the additional effect of receiving a payment, which informs treatment-group respondents whether their stated belief was in the incentivized range.

Throughout, we report treatment effects on seven different outcome variables. (i) To test whether the incentive treatment affects average values of stated beliefs, we report effects on respondents' beliefs relative to the correct value. We construct four different variables to measure our main outcome of interest, the accuracy of stated beliefs: (ii) the absolute distance between the belief and the true value; and indicators of whether (iii) the belief is in the incentivized range; (iv) the belief is in the 10-percent interval around the true value; and (5) the belief takes a very low value of below 100.

Apart from these belief measures, we are interested in treatment effects on (vi) respondents' confidence about their beliefs and on (vii) their response time. Since the instruction text on the screens is longer for the treatment group than for the control group, we apply the following two-step procedure to approximate response time.[22] First, we regress response time in all other questions of the ifo Education Survey on a function of question characteristics and individual fixed effects.[23] We use this model to predict each individual's expected response times for both belief-elicitation questions in the absence of incentive provision. Second, we take the difference between actual and predicted response time, interpreting this difference as the time the respondent invests in belief formation.

To analyze whether the effects of providing incentives are heterogeneous across different subgroups of respondents, we extend our basic regression model to:

$$y_i = \alpha + \beta_1 Incentive_i + \beta_2 Subgroup_i + \beta_3 Incentive_i * Subgroup_i + \epsilon_i \tag{6.2}$$

where $Subgroup_i$ equals one if respondent i belongs to the respective subgroup and zero otherwise. Accordingly, β_1 indicates the incentive-treatment effect for the omitted baseline group of respondents and β_3 measures the additional incentive effect for the subgroup of interest.

6.2.4 Test of Randomization

To test whether the randomization successfully balanced respondents' observable characteristics between treatment and control groups, we run the following regression for each

[22] An ideal experiment to investigate incentive effects on response times would have had the exact same instruction texts on the screens of the control group and the treatment group. While we are interested in response time, our main outcome variable of interest is belief accuracy. Therefore, we opted for reminding treated respondents of the incentive on the belief-elicitation screen to minimize the risk that inattention or imperfect recall attenuates our results. This design decision came at the cost that the response times are not directly comparable across experimental groups without the indicated prediction model.

[23] Question characteristics include the number of words and characters, the number of screens on which the question is presented, and the type of question. We run an individual fixed-effects lasso including fourth-order polynomials of words, characters, and their interactions to determine the optimal inputs (details available upon request).

characteristic and both experiments:

$$y_i = \alpha + \gamma Treatment_i + \epsilon_i \quad (6.3)$$

Table 6.1 reports the coefficients γ from equation 6.3 for the earnings experiment and the school-spending experiment (columns 3 and 6, respectively), along with the corresponding means of the control and treatment groups (columns 1–2 and 4–5). Reassuringly, only one out of 42 estimated differences turns out statistically significant ($p<0.05$), which would be expected by pure change. Thus, the balancing tests suggest that randomization worked as intended. The tests also indicate that the incentive treatment does not affect item non-response, which is very low (below 0.5 percent) on both belief questions.

6.3 The Effects of Incentives on Belief Accuracy

This section presents our main results in three steps. First, we analyze incentive effects on earnings beliefs. Second, we analyze incentive effects on school-spending beliefs. Third, we provide an initial exploration of how to reconcile different incentive effects in the two experiments.

6.3.1 Beliefs about Earnings by Professional Degree

We start our analysis by investigating whether the incentive treatment affects stated beliefs about earnings by professional degree. Panel A of Figure 6.2 depicts the distribution of respondents' beliefs about earnings without a professional degree (left Panel) and with a university degree (right Panel) in the main survey. Solid bars and transparent bars represent beliefs in the control group and in the incentive-treatment group, respectively. In general, beliefs are quite dispersed with a 10–90 percentile range of 800 Euro to 1,500 Euro for earnings without a degree (true value: 1,400 Euro) and 1,500 Euro to 4,000 Euro for earnings with a university degree (true value: 2,750 Euro) in the control group. Moreover, the majority of respondents, 82 percent and 57 percent, respectively, underestimate current earnings of those without a degree and those with a university degree. Comparing beliefs between the control and the treatment groups, graphical inspection does not reveal obvious differences in the distributions of beliefs by treatment status.

Table 6.2 presents treatment-effect estimates of incentive provision based on equation 6.1. Panel A depicts treatment effects on earnings beliefs without a degree and Panel B on earnings beliefs with a university degree.[24] Results indicate that coefficients on the treatment-group indicator are mostly insignificant, suggesting that the incentive treatment hardly affects earnings beliefs. In particular, the treatment does not affect any of the four measures of belief

[24] Since we elicited both earnings beliefs on the same screen, confidence and response time were recorded only once for earnings beliefs in general.

accuracy in Panels A or B. The effects on the absolute distance between respondents' average reported belief and the true value are small at 97.46 Euros and 80.94 Euros, respectively, and statistically insignificant (see column 2). The same is true for treatment effects on the probability of reporting a belief within the incentivized range of 1.5 and -1.1 percentage points (column 3), on the probability to report a belief within the 10-percent interval around the true value of -0.1 and -1.1 percentage points (column 4), and on the probability of stating a very low belief (below 100 Euro) of -0.3 and -0.5 percentage points (column 5). Similarly, the treatment does not affect respondents' confidence that their belief is close to correct (column 6).[25] While we find that the treatment marginally significantly ($p<0.1$) increases respondents' beliefs about university graduates' earnings relative to the true value (2,750 Euro) by 5.7 percentage points (from 101 percent in the control group, see control mean), the overall pattern suggests no incentive effects on earnings beliefs. However, the treatment significantly increases our response-time measure by 19.6 seconds, which suggests that respondents think more carefully about the question when incentives are provided. This interpretation is supported by additional analyses that reveal a positive association between our measure of response time and belief accuracy, both in the treatment group and in the control group (not shown).[26] Furthermore, closer inspection of the distribution of response times in fact shows that the treatment shifts the entire distribution of response times upward—including very short response times (results available upon request). Thus, even though increased response times do not translate into higher belief accuracy, the positive incentive effects on response times suggest positive effects on respondent effort, which could improve data quality.

Another interesting outcome dimension is the overall belief dispersion, describing respondents' disagreement in beliefs.[27] We do not find that incentive provision affects the dispersion of beliefs: The standard deviation of earnings beliefs without a degree (with a university degree) amounts to 351 Euro (1052 Euro) in the control group and is not statistically significantly different from the respective value of 351 Euro (1064 Euro) in the treatment group (results available upon request).

While incentives have no overall effect on the entire population's earnings beliefs, they might improve beliefs in certain subgroups of respondents. To explore this possibility, Appendix Table A6.4 estimates heterogeneous treatment effects on the accuracy of beliefs about earnings of those without a degree (column 1) and those with a university degree (column 2). This heterogeneity analysis is based on equation 6.2 and focuses on subgroups defined by the sociodemographic characteristics in Table 6.1. We find no effect heterogeneities on the accuracy of earnings beliefs for those without a degree. Similarly, treatment effects on the accuracy of earnings beliefs for those with a university degree are largely homogeneous across sociodemographic subgroups. The only exceptions are respondents with a university entrance

[25] As confidence is elicited on a seven-point scale, the incentive coefficient shows that the treatment increases confidence by an insignificant 0.033 points on this scale.

[26] We thank an anonymous referee for suggesting this additional analysis.

[27] For instance, Mankiw et al. (2003) show that disagreement about inflation is associated with several macroeconomic variables.

degree (respondents aged over 65), who exhibit a larger (smaller) treatment effect than persons with basic degree or less (persons below age 45). Overall, the heterogeneity analysis shows that the population-wide null effect of incentives on the accuracy of earnings beliefs reflects null effects in different sociodemographic subgroups, rather than opposing subgroup effects that cancel each other out.

To test for persistence of any treatment effects, we examine whether incentive provision affects respondents' stated beliefs in a follow-up survey about two weeks after the main survey. The follow-up survey is designed to evaluate the persistence of potential treatment effects beyond the immediate horizon of the main survey by eliciting beliefs without incentive provision. While participants in the follow-up differ from participants in the representative main survey in some sociodemographic characteristics, it is reassuring that follow-up participation does not relate to treatment status in the main survey (see Appendix Table A6.5). As a consequence, follow-up survey respondents' characteristics remain well balanced between control and treatment groups (see Appendix Table A6.6). This mitigates concerns about non-random selection into the follow-up and facilitates identification of persistent incentive effects.

Panel B of Figure 6.2 (as well as Appendix Table A6.7) confirm our results from the main survey by showing that incentive provision does not induce noteworthy improvements in the accuracy of stated earnings beliefs in the follow-up survey. Interestingly, the treatment effect on response time turns small and insignificant in the follow-up survey, suggesting that incentive provision in the main survey, which prolongs response time initially, does not affect how much time respondents invest in answering the same question in the follow-up survey where no incentives are provided.

6.3.2 Beliefs about Public School Spending

The second experiment analyzes incentive effects on beliefs about average annual public school spending per student. Panel A of Figure 6.3 depicts respondents' beliefs in the main survey, separately for respondents in the unincentivized control group (solid bars) and respondents in the incentivized treatment group (transparent bars). Beliefs about current spending levels are very dispersed, with a 10–90 percentile range of 50 Euro to 15,000 Euro in the control group (true value: 7,400 Euro). Again, respondents tend to underestimate current spending levels on average, with 78 percent stating a belief below the true value. Visual inspection suggests that the treatment group's belief distribution has less density at the very left part of the distribution (representing very low belief values) and more density close to the true value, which suggests that incentive provision improves school-spending beliefs.

Panel A of Table 6.3 presents regression results based on equation 6.1 that confirm the graphical inspection that incentivization improves the accuracy of school-spending beliefs. The incentive treatment strongly and significantly increases the probability of reporting a belief within the incentivized range from 37 percent to 50 percent (column 3) and within the 10-percent interval around the true value from 3 to 7 percent (column 4). Likewise, the treatment significantly decreases the probability of reporting an implausibly low value of below 100 Euro by 6 percentage points (column 5). While treatment-effect estimates on the relative belief (column 1) and on the absolute distance between belief and true value (column 2) are not statistically significant, the signs of the coefficients indicate that beliefs in the incentivized treatment group are closer to the true value. Furthermore, the treatment significantly increases respondents' confidence about their beliefs (column 6) and doubles the time respondents take to answer the question (column 7). The latter effect on response time is again due to a positive shift throughout the distribution of response times (results available upon request).[28]

These findings are in line with the previous literature. Prior et al. (2015) compare the share of correct answers between treatment and control groups and report an average effect of 4.2 percentage points. Our incentive effect on reporting a belief within the 10-percent range around the true value is comparable in magnitude with 4.4 percentage points. Note, however, that the share of correct answers in the control group is much higher in Prior et al. (2015) than in this chapter (32.3 percent versus 3 percent), which might limit the comparability of the studies.

Investigating the dispersion of stated beliefs, once again we do not find significant incentive effects on respondents' disagreement: The standard deviation of school-spending beliefs amounts to 9,451 Euro in the control group and 9,667 Euro in the treatment group (p-value=0.33) (results available upon request).

The specification in column 3 of Appendix Table A6.4 estimates heterogeneous incentive-treatment effects on school-spending beliefs across sociodemographic subgroups. Incentive-treatment effects are significantly larger for respondents with higher educational attainment than for those with lower education, and for those living in West Germany compared to those living in East Germany. They are also significantly smaller for older respondents, those living in large cities, and parents.[29] Given that respondents' risk preferences have been theorized to affect the ability of incentive schemes to foster truthful reporting (e.g., Trautmann and van de Kuilen, 2015), it is interesting to note that treatment effects do not vary by answers to the general risk question.

Results of the follow-up survey reveal that incentivizing belief accuracy improves school-spending beliefs persistently (Panels B of Table 6.3 and Figure 6.3).[30] Similar to the immediate

[28] Again, additional analyses show that response time is positively associated with belief accuracy in both experimental groups of the school-spending experiment (not shown), which highlights the potential implications increased response time can have for data quality.

[29] Focusing on the subgroup of parents with school-aged children (i.e., those 14 percent of respondents who have children aged between 6 and 18), it is interesting to note that their treatment effects are not statistically significantly different from respondents without school-aged children.

[30] Again, Appendix Tables A6.5 and A6.6 show that selection into the follow-up survey is random with respect to treatment status in the main survey.

effects in the main survey, incentive provision increases the probability of reporting a belief within the incentivized range by 11 percentage points and within the 10-percent interval around the true value by 6 percentage points in the follow-up survey. Furthermore, the negative incentive effect on the absolute distance between belief and true value is highly significant and even larger in the follow-up survey than in the main survey. In contrast to the main survey, incentive provision does not affect the probability of reporting very low beliefs of below 100 Euro, presumably because the control-group probability for reporting such low values is only 5 percent in the follow-up survey (compared to 16 percent in the main survey).

In the follow-up survey, the treatment group also continues to be more confident about their beliefs than the control group. By contrast, there is no persistent treatment effect on response time. This suggests that incentive provision in the main survey improves beliefs immediately, and respondents remember their improved beliefs when being resurveyed later (as opposed to permanent shifts in the time invested in forming beliefs).

To explore potential reasons for why treatment effects differ between the main survey and the follow-up survey, we also estimate incentive effects in the main survey for the subset of respondents who participated in the follow-up survey (results available upon request). As it turns out, incentive effects on the distance to the true value are significantly larger in the follow-up survey than in the main survey within follow-up survey respondents, which reveals that larger treatment effects in Panel B than in Panel A of Table 6.3 are not fully explained by selective participation in the follow-up survey. Instead, results suggest that respondents update their beliefs between the main survey and the follow-up survey.[31] Understanding these belief-updating patterns is an interesting avenue for future research.[32]

6.3.3 Exploring Effect Differences: Online-Search Behavior as Potential Channel

Our results thus far show that incentive provision improves the accuracy of stated beliefs about average school spending, but not about earnings by professional degree. Still, incentives significantly increase the time respondents take to state their beliefs in both cases, which

[31] For instance, comparing responses of control-group members that participated in both surveys across surveys shows that their relative belief increases from 69 percent in the main survey to 104 percent in the follow-up survey. A potential reason for why treatment group respondents may update their beliefs between surveys is the incentive payment itself: Treated respondents learned whether or not they received a payment for belief accuracy after the main survey. Since the payment reveals information about the accuracy of respondents' stated belief, it might be that those who learn that their belief was not correct update their belief before re-stating it in the follow-up survey, even though this re-stated belief is no longer relevant for their payment. Note, however, that payment-induced belief updating cannot explain why control group members also seem to update their beliefs between surveys.

[32] Since respondents seem to be somewhat better informed in the follow-up survey than in the main survey (see, for instance, control-group responses in the main survey and in the follow-up survey), we can investigate the extent to which incentives can reduce individual 'recall noise' as measured by the absolute difference between beliefs stated in the main survey and in the follow-up survey. In additional analyses, we find no evidence that incentives have a causal effect on this measure of recall noise (results available upon request).

is consistent with an increase in respondents' effort to provide a correct answer in both experiments. In this section, we investigate increased use of search engines as a potential mechanism for the pattern of results described above.

Scrutiny of the stated school-spending beliefs in the treatment group reveals that the density of treatment-group beliefs spikes at 6,000, 6,300, 6,500, and 6,700 Euro. Quite strikingly, it turns out that these values are among the top results of online-search requests at Google of key phrases of our question wording.[33] Building on this observation, the first column of Table 6.4 regresses a dummy variable coded one if respondents state one of the online-search-engine results as their belief, and zero otherwise, on an incentive-treatment indicator. The treatment significantly increases the likelihood of stating one of the Google results as beliefs by about 12 percentage points from a control-group mean of 2 percent. That is, the incidence of reporting such values increased from 50 respondents in the control group to 277 respondents in the treatment group. While the probability of stating a belief corresponding to Google search results is certainly an imperfect measure of actual online-search activity, the analysis provides suggestive evidence that the incentive effect on school-spending beliefs may operate through respondents' increased use of online-search engines.[34]

There are at least two potential reasons for why online-search activities might only improve school-spending beliefs, but not earnings beliefs. First, beliefs about average earnings by professional degree are less dispersed than school-spending beliefs, which limits the leeway for online-search-engine use to improve beliefs.[35] Relatedly, the fact that respondents have better-informed priors about average earnings than about school spending might make it more likely that they engage in online-search activity in the latter experiment.[36] Second,

[33] Appendix Figure A6.1 shows Google results from search requests of our question wording, namely *'Ausgaben pro Schüler pro Jahr an öffentlichen, allgemeinbildenden Schulen'* and *'Bildungsausgaben pro Schüler pro Jahr'*. We searched for these values shortly after the implementation of the ifo Education Survey 2017 to obtain results close to those which were available to the respondents and found the values 6,000, 6,300, 6,500, 6,700, and 7,400 Euro in the top search results.

[34] Although we do not know which values the respondents ultimately found on the internet, we are confident that our approximation of online-search activities works reasonably well: In a subsequent experiment that explicitly encourages respondents to search for the correct value on the internet, the probability that a stated belief corresponds to the top Google-search results is highly correlated (correlation=0.64) with self-reported search engine usage (see section 6.4.2 below for details).

[35] We remain agnostic about the reasons for why earnings beliefs are relatively less dispersed. One reason might be that own experience makes monthly earnings a more tangible concept than public spending on schools. A complementary reason might be that the survey question that elicits beliefs about earnings of persons without a professional degree and persons with a university degree provides respondents with the anchor of earnings of persons with an apprenticeship degree (see section 6.2.2).

[36] To study the relationship between prior beliefs and the probability of online-search activity more systematically, we performed the following additional analysis of the school-spending experiment: We first imputed prior beliefs by regressing belief accuracy (defined as the probability to state a belief within the 10-percent range around the true value) on a set of explanatory variables in the control group, and predict counterfactual priors using the resulting coefficients. Next, we estimate heterogeneous treatment effects on our proxy of online-search activity—whether a stated posterior belief corresponds to a Google result. While statistically insignificant, the negative

accurate information on net average monthly earnings by professional degree seems to be relatively hard to find online.[37] Thus, the finding that incentive provision only improves school-spending beliefs is consistent with our proposition that incentive effects mainly operate through increased search-engine usage.

This descriptive analysis suggests that incentivizing belief accuracy in online surveys can have the (potentially unintended) effect that respondents resort to online-search engines to improve their stated beliefs—as opposed to increased effort to retrieve truthful beliefs from memory. To scrutinize this possibility further, the next section presents an additional experiment in which we encourage a randomly selected treatment group to use online-search engines before stating school-spending beliefs.

6.4 Encouraging Online-Search Activity

The results discussed in the previous section suggest that incentivized respondents may be more likely to search for correct answers online. To better understand how the use of online-search engines changes stated beliefs, we conduct an additional experiment in which we explicitly encourage respondents to search the internet for the correct answer. This experiment allows us to investigate whether the incentive-treatment effect described above can be reproduced by exogenously induced online-search activity. In what follows, we describe the experimental setup, report results of the encouragement treatment on online-search activity and on belief accuracy, and compare the results to the above effects of incentive provision.

6.4.1 Experimental Design

We conduct the encouragement experiment in a new representative sample of 4,046 respondents in the 2018 wave of the ifo Education Survey.[38] To investigate the impact of online-search activity on stated beliefs, we randomly assigned respondents to a control group and to an encouragement-treatment group before eliciting school-spending beliefs. In both groups, beliefs about average public school spending are elicited using the same wording as in the

point estimate on the interaction term suggests that those with better-informed priors are less likely to react to the incentive by engaging in online-search activity. In fact, the incentive treatment effect is close to zero in the better-informed subgroup (results available upon request). We also estimate heterogeneous incentive effects on our outcomes of interest and find that those with more accurate priors react significantly less strongly to the incentive treatment.

[37] As official statistics on earnings by professional degree are not published, we obtained the values from own calculations based on the German Microcensus. Appendix Figure A6.2 presents results from Google search requests of key phrases of the earnings-beliefs question. These Google results seem to correspond to minimum-wage earnings or to the earnings of workers in specific occupations rather than the German averages by professional degree. Further analyses show that 3 respondents in the control group and 10 respondents in the treatment group report a belief that corresponds to one of the results listed on the first page of Google results.

[38] See Appendix for additional information about the 2018 wave of the ifo Education Survey.

2017 survey wave.[39] Instead of being offered incentives for belief accuracy, members of the treatment group were encouraged to use online-search engines before stating their beliefs. The wording of the encouragement treatment is as follows: *'As an exception for this question, you can search the internet for the right answer to improve your guess, for example by using an internet-search engine.'*[40]

After eliciting school-spending beliefs, we asked all respondents whether they searched for the correct answer on the internet in order to check whether encouraging online-search activity worked as intended. This question is worded as follows: *'To answer the preceding guess question, did you search for the correct value on the internet?'*

Columns 1–3 of Appendix Table A6.8 present sociodemographic characteristics for the control group and the encouragement-treatment group, indicating that the randomization successfully balanced respondents' observable characteristics across experimental groups.

6.4.2 The Effect of the Encouragement Treatment on Online-Search Activity

We start the analysis by investigating whether the encouragement treatment successfully increased the use of online-search engines. Column 1 of Table 6.5 reports results of regressing self-reported online-search activity on the treatment indicator. While search-engine use is very low in the control group at 1 percent, respondents in the treatment group are 14 percentage points more likely to state that they searched for the correct value on the internet. This manipulation check suggests that our encouragement treatment did in fact increase the usage of online-search engines.

To allow for direct comparison with the incentive experiment, we next construct the proxy for online-search activity introduced in section 6.3.3 for the search-engine experiment. Again, we code a dummy variable equal to one if respondents' stated beliefs correspond to one of the top Google search results for average public school spending.[41] As it turns out, the correlation of this dummy variable with the dummy variable for self-reported search-engine use is high (correlation=0.64), confirming its validity as a proxy for online-search activity.

[39] The true value in the 2018 survey wave is 7,500 Euro (Statistisches Bundesamt, 2018), which differs slightly from the previous year's spending level.

[40] To keep the structure of this experiment as similar as possible to the incentive experiment, the question text also informed respondents in the treatment group that they will receive feedback about the accuracy of their belief at the end of the survey.

[41] In 2018, the top results of Google searches for key phrases of our question wording (*'Ausgaben pro Schüler pro Jahr an oeffenlichen, allgemeinbildenten Schulen'* and *'Bildungsausgaben pro Schueler pro Jahr'*) were 6,300, 6,500, 6,700, 6,900, 7,400, 7,500, and 8,900 Euro. Again, we searched for these values shortly after the implementation of the ifo Education Survey 2018 to obtain results close to those which the respondents would have found. Note that these values differ slightly from those found one year earlier. Screenshots are available upon request.

Column 2 of Table 6.4 displays the encouragement-treatment effect on this proxy for online-search activity. The treatment significantly increases proxied search-engine usage by 9.1 percentage points, which is close to the 11.6 percentage-point treatment effect of incentive provision in the main experiment (see column 1). Thus, incentive provision and encouraging online-search activity seem to produce very similar effects on the probability to search the internet for the correct answer on the school-spending belief question.

6.4.3 The Effect of Encouraging Online-Search Activity on Belief Accuracy

The encouragement treatment strongly affects stated beliefs about average public school spending. Panel A of Figure 6.4 depicts the distribution of respondents' belief in the control (solid bars) and treatment groups (transparent bars). In the control group, beliefs about spending levels are again very dispersed, with a 10–90 percentile range of 170 Euro to 15,000 Euro (true value: 7,500 Euro). The majority of 79 percent states a belief below the true value. This distribution of beliefs is remarkably similar to the unincentivized control group in the main experiment in the 2017 survey wave, where the 10–90 percentile range was 50 Euro to 15,000 Euro, and 78 percent of beliefs were below the true value. Comparing the distribution of beliefs between the control and the encouragement-treatment groups, the patterns are again very similar to the incentive experiment: Treated respondents are less likely to report very low belief values and more likely to report beliefs close to the true value.

Columns 2–8 of Table 6.5 present treatment-effect estimates based on equation 6.1. For most outcomes, the effects of encouraging search-engine usage on respondents' stated beliefs are very similar to the incentive effects in the main experiment (see Panel A of Table 6.3 for comparison). The encouragement treatment increases average school-spending beliefs by 11 percent ($p<0.1$) (column 2). While the incentive-treatment effect on this outcome was not significant, the sign and magnitude of the two treatment effects are virtually identical. The negative sign on the treatment indicator in column 3 suggests that encouraging search-engine usage decreases the distance between stated belief and true value by 489 Euro. While this effect is shy of statistical significance, it again goes in the same direction as the incentive-treatment effect on this variable (which is larger and statistically significant). The encouragement treatment significantly increases the probability of reporting a belief within the incentivized range by 9 percentage points (column 4) and the probability of reporting a belief within the 10-percent interval around the true value by 5 percentage points (column 5).[42] Again, these effects are similar, in significance and magnitude, to the incentive-treatment effects.

The encouragement treatment also significantly increases respondents' confidence about their beliefs (column 7) and the response time (column 8). While treatment effects on confi-

[42] The incentivized interval in the search-engine experiment ranges from 1,332 Euro to 13,468 Euro. Even though no incentives were provided in this experiment, we report treatment effects on this measure to allow comparison to the findings from the main experiment. The 10-percent interval around the true value ranges from 6,750 Euro to 8,250 Euro.

dence are very similar across the two experiments, the encouragement effect on response time is somewhat smaller than the incentive effect. The only outcome for which we observe fairly different treatment effects between the encouragement treatment and the incentive treatment is the probability to report very low belief values below 100 Euro (column 6). In contrast to the incentive treatment, encouraging search-engine usage does not affect reporting of implausibly low belief values, although (as in the follow-up survey of the incentive experiment) this is driven by a lower incidence in the control group. In general, the results confirm the visual impression that the effects of incentives and encouraging online-search activity are remarkably similar.

Appendix Table A6.9 shows results of a model that stacks the data of the incentive experiment and the encouragement experiment. Results confirm that the treatment effects differ significantly only on value estimates below 100 Euro. For all other outcome measures, the treatment effects of the encouragement experiment do not differ significantly from the treatment effects of the incentive experiment.

In sum, the similarity between the effects of the incentive treatment and the encouragement treatment, together with the inspection of response spikes in the incentive treatment in section 6.3.3, strongly suggest that at least part of the effect of incentive provision on the accuracy of school-spending beliefs is due to incentives increasing respondents' online-search activity to improve their stated beliefs.

6.5 Inducing Experimenter-Demand Effects

One major concern with the interpretation of treatment effects in experimental work is bias due to experimenter-demand effects, which refer to 'changes in behavior by experimental subjects due to cues about what constitutes appropriate behavior' (Zizzo, 2010, p. 75). For the case of incentive provision, the experimenter's willingness to pay out monetary incentives for accurate beliefs may signal to respondents that the incentivized belief question is of particular importance to the experimenter. If so, respondents might adapt their answering behavior not only to earn the incentive, but also to please the experimenter by being a 'good' respondent. To address this concern, we report results from an additional experiment where we aim to induce experimenter-demand effects in order to assess their potential effect on stated school-spending beliefs. In what follows, we first describe the experimental design and then present the results.

6.5.1 Experimental Design

We conducted the demand experiment in another representative sample of 3,124 respondents in the 2016 wave of the ifo Education Survey.[43] Respondents were randomized into a control

[43] See Appendix for additional information about the 2016 wave of the ifo Education Survey.

group or a demand-treatment group. For respondents in the demand treatment, we used the following wording to induce experimenter-demand effects before eliciting their beliefs about average school spending: *'As you might know, government institutions collect a variety of key statistics about schools. We are interested in discovering whether the public is familiar with these key statistics. On the next screen, we will ask you a question about such a key statistic, to which there are correct and incorrect answers. In order for your response to be informative for us, it is very important that you answer this question as accurately as possible.'*[44] On the next screen, beliefs about school spending were elicited in the same way as in the control group, using the same wording as in the 2017 and 2018 waves.[45]

Columns 4–6 of Appendix Table A6.8 shows that respondents' sociodemographic characteristics are again well balanced across experimental groups, indicating that randomization was successful.

6.5.2 The Effect of Inducing Experimenter-Demand Effects on Belief Accuracy

Panel B of Figure 6.4 depicts the results of the demand experiment on respondents' stated beliefs about average public school spending. The distribution of beliefs in the control group is similar to the previously reported survey waves, with a 10–90 percentile range of 150 Euro to 10,000, and 82 percent of respondents stating a belief below the true value (true value: 7,100 Euro; Statistisches Bundesamt, 2016). However, in contrast to the previously reported experiments on school-spending beliefs, graphical inspection does not yield any obvious difference in the distribution of beliefs between control group (solid bars) and demand-treatment group (transparent bars).

The regression analysis in Table 6.6 confirms this impression. There is no indication whatsoever that inducing experimenter demand improves the accuracy of stated beliefs. The small and insignificant coefficients on the treatment indicator in most regressions suggest that the experimenter-demand treatment does not affect stated beliefs about average school spending. The only marginally significant treatment effect ($p<0.1$) is that the treatment *increases* the absolute distance between stated belief and true value.

Consistently, column 3 of Table 6.4 shows that the demand treatment also does not affect our proxy for online-search activity.[46] Taken together, the evidence suggests that the effects of

[44] Again, the question text also informed treated respondents that they will receive feedback about the accuracy of their belief at the end of the survey.

[45] In spirit, this demand treatment is very similar to recent papers by de Quidt et al. (2018) and Mummolo and Peterson (2019) which measure and bound experimenter-demand effects in the context of economic games and survey experiments, respectively.

[46] For this purpose, we again code a dummy variable equal to one if respondents report one of the top Google search results for average public school spending (2,200, 4,900, 5,600, 6,000, 6,200, 6,300, 6,500, 7,300, 8,000, and 8,100 Euro). Unlike in the 2017 and 2018 waves, we did not record the search results immediately after the implementation of the 2016 survey wave. However, when looking up the Google search results in autumn 2018,

incentive provision on answering behavior reflect genuine incentive effects as opposed to experimenter-demand effects.

6.6 Conclusion

Subjective beliefs about economic facts are a key concept for explaining economically relevant behavior and choices. The fact that researchers usually measure these beliefs using unincentivized survey questions raises concerns about systematic biases in reporting that might undermine meaningful analysis of subjective beliefs. We conduct two experiments in a representative German online survey to study whether incentivizing belief accuracy affects stated beliefs about two important economic facts: average earnings by professional degree and average public school spending. We find that incentivization increases response time in both experiments but that this translates into more accurate beliefs only for average school spending. To reconcile the heterogeneity of incentive effects on stated beliefs in the two experiments, we present evidence from comparisons of response spikes with search-engine results and from an additional experiment that encourages online-search activity which suggests that respondents resort to online-search engines in response to incentive provision, in particular for beliefs for which they do not have strong priors. In another experiment that induces experimenter-demand effects, we show that these results are unlikely to reflect experimenter-demand effects in our setting.

We draw two main conclusions from our analyses. First, the finding that monetary incentives fail to improve respondents' earnings beliefs suggests that unincentivized belief measures, which are heavily used in the literature, do not necessarily suffer from systematic reporting bias due to self-serving answering behavior. Of course, the extent to which this finding is generalizable to other belief domains is an open question and likely depends on the properties of the elicited belief (e.g., whether objective information is readily available online or whether respondents' priors are well-formed).

Second, providing monetary incentives in online surveys might increase respondents' use of external resources such as online-search engines to improve the accuracy of their stated beliefs. The extent to which respondents engage in this behavior will likely depend on a number of factors, such as the question at hand, the specific survey setting (e.g., whether it is possible to restrict or track respondents' online-search activity), or the design of the incentive scheme. Also, whether or not researchers want respondents to consult external sources to update beliefs might depend on the specific research question. In any case, our results constitute a cautionary note that incentivizing belief accuracy might trigger unintended behavioral responses in online surveys and other contexts where experimenters' control over the experimental setting is limited.

we restrict the search to display only results published until end of June 2016, the end of the field phase of the 2016 survey. Screenshots are available upon request.

In sum, our results point to a trade-off that researchers face when deciding upon whether to incentivize beliefs about economic facts in online surveys. On the one hand, it might be undesirable for many research questions related to subjective beliefs that respondents consult external resources in response to incentive provision. On the other hand, there can be clear advantages when incentivizing beliefs: First, as our response-time analysis suggests, respondents think more carefully about their answers when incentives are provided—which is desirable from a data-quality perspective. Second, the undesirable consequences of incentive provision in surveys (i.e., encouraging online-search activity) are likely muted when incentives are provided for predicting future events. While incentivizing such expectations certainly has its own challenges (e.g., administering delayed payments after the variable of interest has been realized), providing such incentives is certainly possible in the context of online surveys. Third, incentivizing beliefs is a key feature for several important research questions, for instance when researchers are interested in costly information-acquisition behavior (e.g., Fuster et al., 2020), in motivating respondents to encode information for later retrieval (e.g., Zimmermann, 2020), or in mitigating the reporting of motivated beliefs (e.g., Bullock et al., 2015; Prior et al., 2015; Zimmermann, 2020). Also, encouraging online-search activity might be an interesting research approach to induce belief updating in an unobtrusive way. Thus, while the effects of incentives on online-search activity that we discuss in this chapter might limit the usefulness of incentives in some settings, researchers should consider all these aspects carefully when designing survey instruments.

The fact that the incentives we provide for a correct answer are relatively small (at least in absolute monetary terms) at about 0.33 Euro needs to be kept in mind when interpreting our results. While we like to think of our findings as lower-bound effects of what treatment effects can be expected when incentives are higher, the sensitivity of incentive effects on belief accuracy—and on the propensity to use online-search engines or other external sources—requires further research. In particular, it would be interesting to test experimentally if incentives of different sizes lead to differences in incentive effects. While the current literature gives some indication that larger incentives indeed yield larger effects on belief accuracy (e.g., Bullock et al., 2015; Zimmermann, 2020), the extent to which incentive size influences online-search activity is an open empirical question. At the same time, our findings that incentives increase response time in both experiments, belief accuracy in the school-spending experiment, and self-reported effort to give a correct answer, make us confident that the provided incentives were meaningful for respondents.

An important concern when conducting experiments online (as opposed to in a physical laboratory) is that the lack of direct interactions between the experimenter and the study participants might undermine participants' trust in the researcher. For our study, lacking trust of respondents that they will actually receive the promised payments for belief accuracy would attenuate the incentive treatment effect estimates. However, lacking trust towards the experimenter is unlikely to be a major concern in our setting. First, Horton et al. (2011) investigate the extent to which participants in an online experiment (MTurk workers) and participants in traditional laboratory experiments differ in their trust that they will be paid as described in the experimental instructions. Their findings indicate that trust levels are fairly high and hardly differ across subject pools.[47] Second, assuming that those participants in our incentive experiments who had already participated in earlier waves of the ifo Education Survey are more likely to trust that payments will be carried out as advertised, we use earlier survey participation as a proxy for trust.[48] In additional analyses, we estimate heterogeneous incentive effects by earlier survey participation and find no evidence that respondents who participated in earlier waves react more strongly to the incentive treatment (results available upon request). This indirect evidence suggests that respondents' lack of trust is not a major concern in our setting. We consider studying the role of trust towards the experimenter on subjects' behavior an interesting avenue for future experiments, in particular in the context of online surveys.

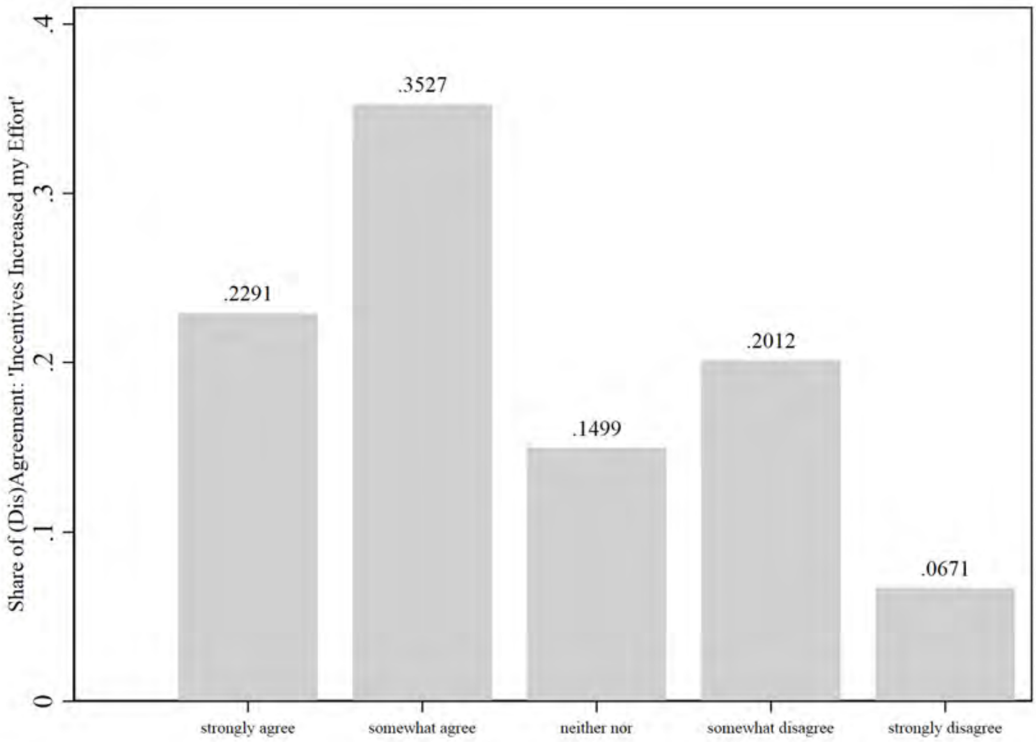
For future research, it would also be interesting to investigate which attributes of a belief question mediate the effects of incentive provision. As the two belief questions scrutinized in this chapter—beliefs about average earnings and school spending levels—differ in various dimensions such as respondents' confidence about their beliefs, question complexity, respondents' familiarity with the elicited concepts, and the online availability of belief-improving information, further investigating the interplay between question attributes and incentive effects would be insightful.

[47] On a 7-point Likert scale where higher numbers indicate higher trust, the average response was 5.41 among MTurk workers and 5.74 among lab subjects (Horton et al., 2011, p. 419) .

[48] In our 2017 data, 24 percent of respondents had participated in at least one earlier survey wave. Note that these respondents differ from new respondents along various dimensions: they are older, more likely to be full-time employed, less likely to have a middle school degree, and more risk averse.

Figures and Tables

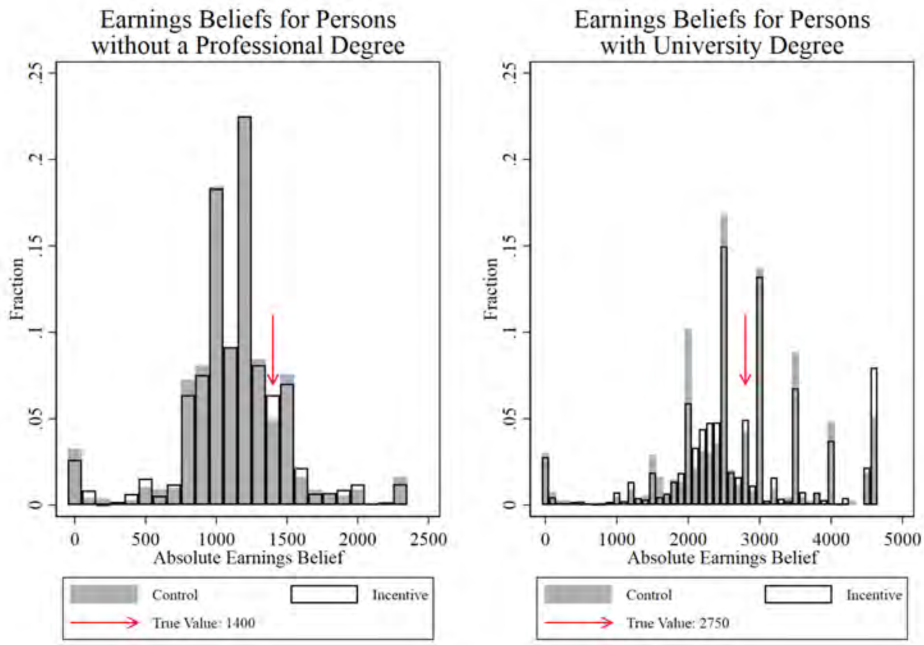
Figure 6.1: Survey Responses on Whether Incentives Increased Effort



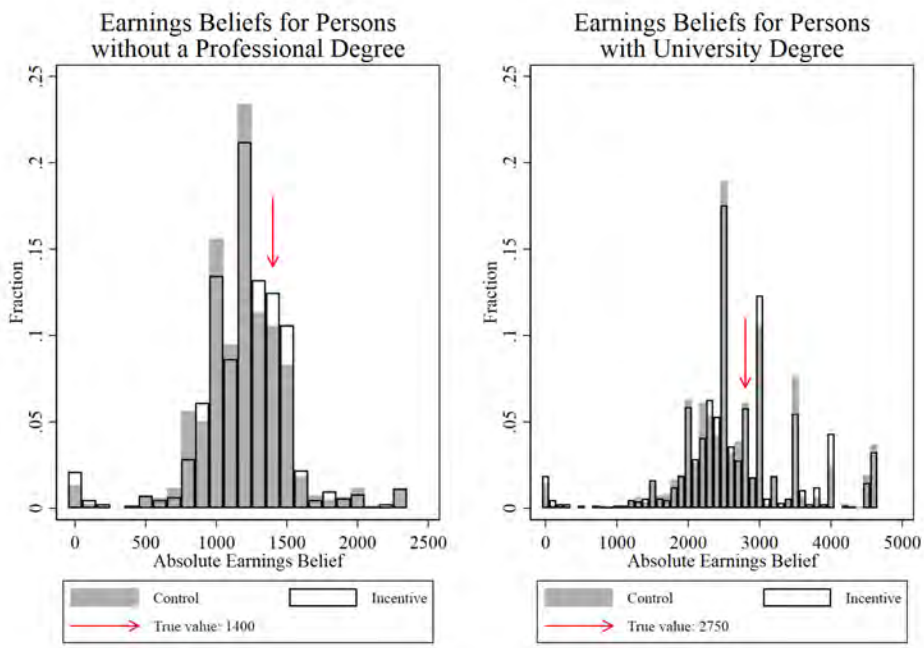
Notes: Wording of the question: *'In the previous questions, you had the opportunity to earn additional tokens by providing a good guess. To what extent do you agree with the following statement (your answer will not affect your probability of winning): The prospect of receiving additional tokens has provided an incentive for me to put more effort in my guess.'* Sample: respondents incentivized in at least one belief question. Responses weighted by survey weights. Source: ifo Education Survey 2017.

Figure 6.2: Distribution of Earnings Beliefs with and without Incentive Provision

Panel A: Main Survey



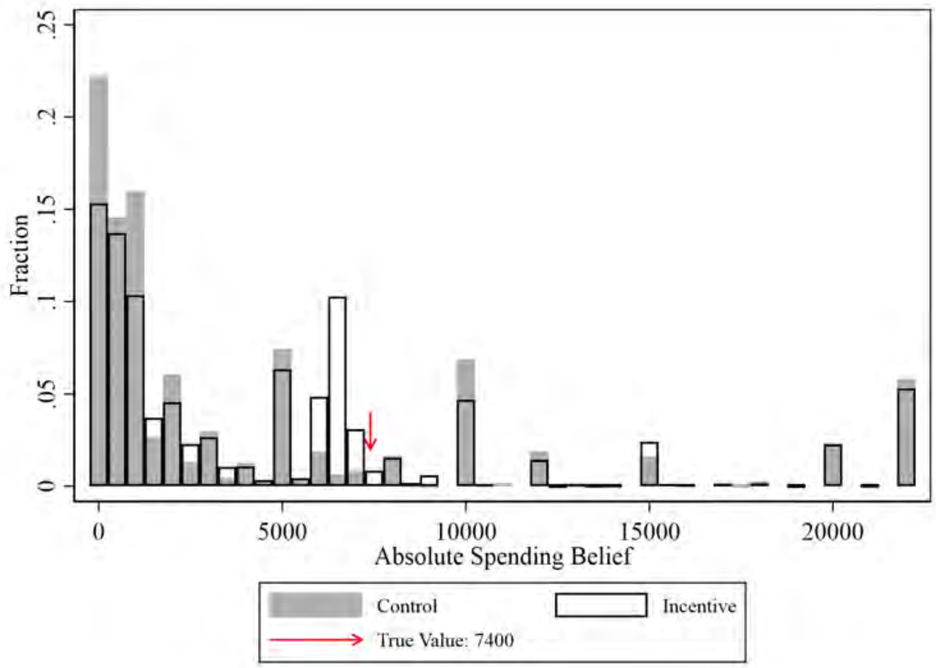
Panel B: Follow-up Survey



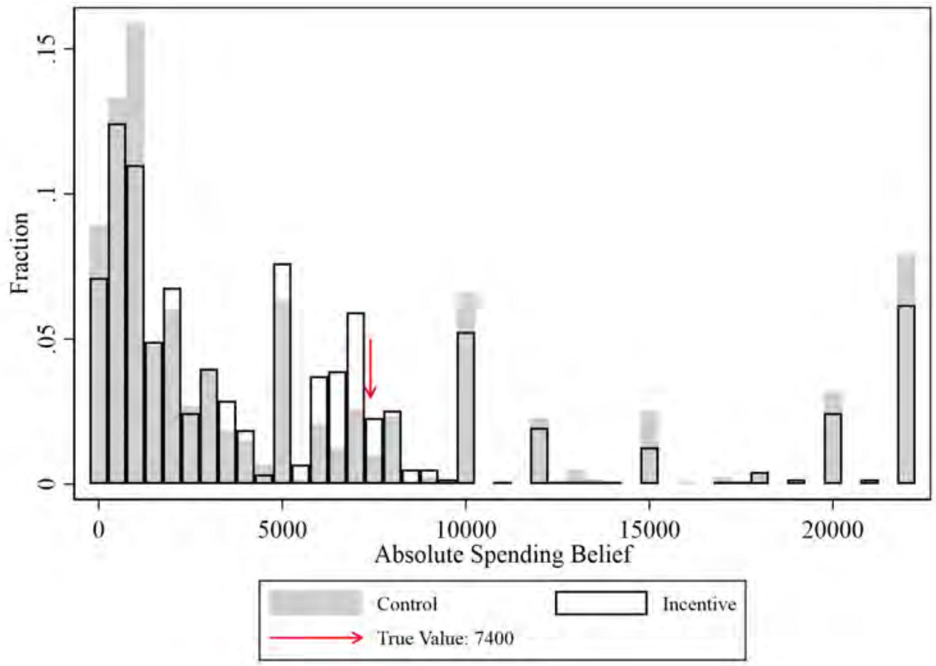
Notes: Histograms of respondents' stated beliefs about average earnings of persons without a degree (left panels) and with a university degree (right panels). Beliefs are winsorized at 2,380 Euro respectively 4,675 Euro to deal with outliers. Incentive: belief accuracy incentivized. Control: belief accuracy not incentivized. Panel A: beliefs elicited in main survey. Panel B: beliefs elicited in follow-up survey about two weeks later. Responses weighted by survey weights. Source: ifo Education Survey 2017.

Figure 6.3: Distribution of School-Spending Beliefs with and without Incentive Provision

Panel A: Main Survey



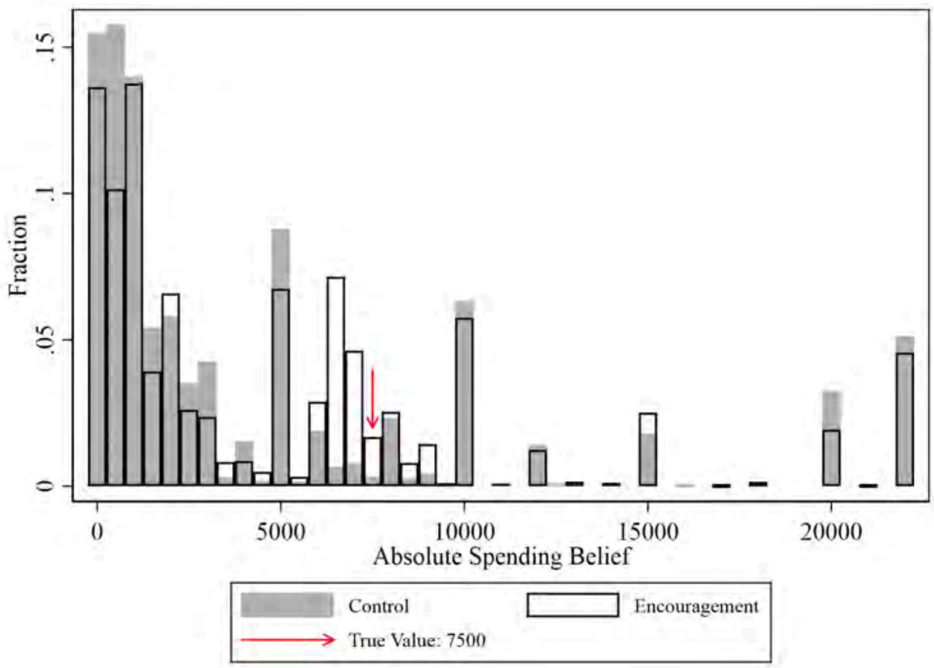
Panel B: Follow-up Survey



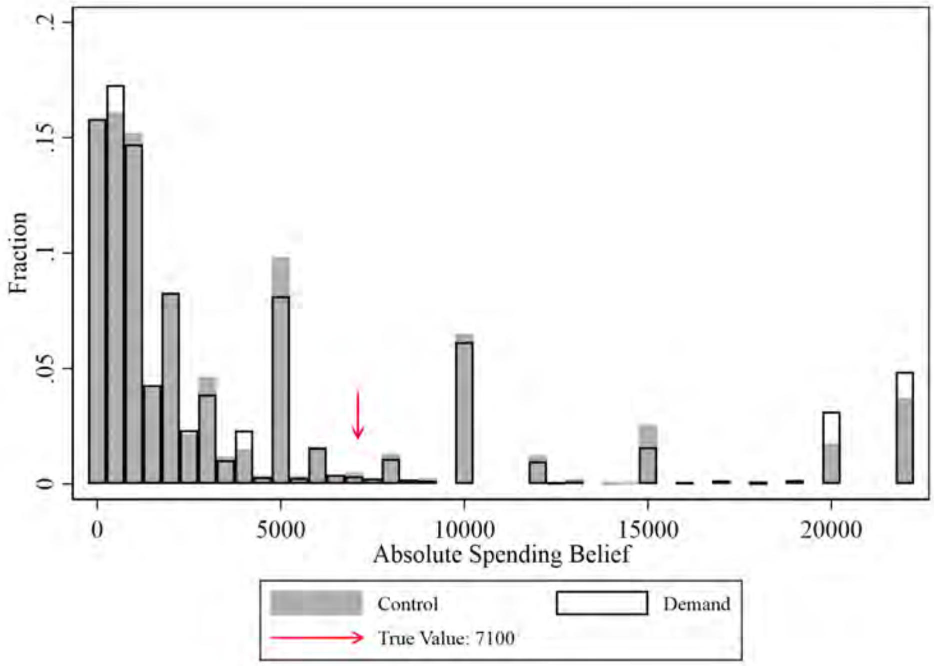
Notes: Histograms of respondents' stated beliefs about average annual public school spending per student. Beliefs are winsorized at 22,200 Euro to deal with outliers. Incentive: belief accuracy incentivized. Control: Belief accuracy not incentivized. Panel A: beliefs elicited in main survey. Panel B: beliefs elicited in follow-up survey about two weeks later. Responses weighted by survey weights. Source: ifo Education Survey 2017.

Figure 6.4: Distribution of School-Spending Beliefs in Additional Experiments

Panel A: Experiment on Encouraging Online-Search Activity



Panel B: Experiment on Inducing Experimenter-Demand Effects



Notes: Histograms of respondents' stated beliefs about average annual public school spending per student. Panel A: beliefs in the experiment on encouraging online-search activity. Panel B: Beliefs in the experiment on inducing experimenter-demand effects. Beliefs are winsorized at 22,200 Euro to deal with outliers. Encouragement/Demand: belief in the respective treatment group. Control: belief in the control group. Responses weighted by survey weights. Source: Panel A: ifo Education Survey 2018; Panel B: ifo Education Survey 2016.

Table 6.1: Summary Statistics and Balancing Tests of Incentive Experiments

	Earnings Experiment			School-Spending Experiment		
	Control Mean (1)	Treatment Mean (2)	Difference (3)	Control Mean (4)	Treatment Mean (5)	Difference (6)
Highest Educational Attainment						
No Degree/Basic Degree	0.360	0.403	0.043	0.386	0.378	-0.007
Middle School Degree	0.307	0.297	-0.009	0.302	0.302	0.000
University Entrance Degree	0.333	0.300	-0.033	0.313	0.320	0.007
Age	49.010	49.855	0.845	49.235	49.663	0.428
Monthly Household Income (€)	2324	2282	-42	2327	2276	-51
Female	0.495	0.527	0.032	0.505	0.518	0.013
Born in Germany	0.948	0.966	0.018**	0.956	0.959	0.003
Partner in Household	0.593	0.563	-0.030	0.576	0.580	0.004
Lives in West Germany	0.795	0.806	0.011	0.805	0.797	-0.008
City Size = 100,000	0.323	0.312	-0.011	0.307	0.328	0.020
Has Children	0.579	0.576	-0.003	0.581	0.572	-0.009
Parent(s) with University Degree	0.279	0.268	-0.011	0.277	0.269	-0.008
Employment Status						
Full-time Employed	0.348	0.327	-0.021	0.347	0.328	-0.019
Part-time Employed	0.134	0.114	-0.020	0.126	0.121	-0.005
Self-Employed	0.039	0.043	0.004	0.046	0.035	-0.011
Unemployed	0.045	0.041	-0.004	0.037	0.049	0.013
House Wife/Husband/Retired/Ill/Student	0.053	0.072	0.019	0.058	0.067	0.009
Works in Education Sector	0.088	0.079	-0.010	0.089	0.077	-0.012
Risk Tolerance (11-Point Scale)	4.389	4.258	-0.131	4.337	4.306	-0.031
Patience (11-Point Scale)	6.089	6.217	0.128	6.118	6.194	0.076
Item Non-response: Beliefs	0.001	0.001	-0.001	0.003	0.003	0.000
Observations	1,852	1,844		1,856	1,840	

Notes: Group means. 'Difference' displays the difference in means between the respective control and treatment groups. Significance levels of 'Difference' stem from linear regressions of the background variables on the respective treatment dummies. Significance levels: *** p<0.01. ** p<0.05. * p<0.10. Data source: ifo Education Survey 2017.

Table 6.2: Incentive Effects on Earnings Beliefs

	Belief Relative to True Value	Distance to True Value	Belief within Incentivized Range	Belief within 10-percent Range	Extreme Values Below 100	Confidence	Response Time
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Panel A: Beliefs on Earnings without a Degree							
Incentive	0.080	97.463	0.015	-0.001	-0.003		
	(0.064)	(87.020)	(0.021)	(0.017)	(0.007)		
Control Mean	0.840	429.440	0.550	0.190	0.040		
Observations	3,694	3,694	3,694	3,694	3,694		
R-squared	0.001	0.001	0.000	0.001	0.001		
Panel B: Beliefs on Earnings with University Degree							
Incentive	0.057*	80.94	-0.011	-0.011	-0.005	0.033	19.609***
	(0.034)	(83.067)	(0.02)	(0.021)	(0.007)	(0.067)	(5.886)
Control Mean	1.01	900	0.71	0.39	0.04	3.39	31.17
Observations	3,693	3,693	3,693	3,693	3,693	3,692	3,696
R-squared	0.002	0.001	0.000	0.001	0.001	0.001	0.004

Notes: OLS regressions. Dependent variable: column (1): stated belief divided by true value; column (2): absolute distance between stated belief and true value; column (3): dummy variable coded one if stated belief is in the incentivized range; column (4): dummy variable coded one if stated belief is in the 10-percent interval around the true value; column (5): dummy variable coded one if stated belief is below 100; column (6): confidence about belief on seven-point Likert scale (1='very unsure', 7='very sure'); column (7): difference between actual and predicted response time. Randomized experimental treatment 'Incentive': respondents offered monetary incentive for belief accuracy. Control mean: mean of the outcome variable for the control group. Panel A: beliefs about net average monthly earnings of full-time employed persons without any professional degree. Panel B: beliefs about net average monthly earnings of full-time employed persons with a university degree. Since both earnings beliefs were elicited on the same screen, confidence and response time were recorded only once for earnings beliefs in general. Regressions control for treatment status in the incentive experiment on school-spending beliefs. Regressions weighted by survey weights. Robust standard errors in parentheses. Significance levels: *** p<0.01. ** p<0.05. * p<0.10. Data source: ifo Education Survey 2017.

Table 6.3: Incentive Effects on School-Spending Beliefs

	Belief Relative to True Value	Distance to True Value	Belief within Incentivized Range	Belief within 10-percent Range	Extreme Values Below 100	Confidence	Response Time
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Panel A: Main Survey							
Incentive	0.112	-720.657	0.132***	0.044***	-0.064***	0.768***	86.068**
	(0.078)	(470.283)	(0.021)	(0.009)	(0.015)	(0.068)	(38.013)
Control Mean	0.770	7590.930	0.370	0.030	0.160	2.420	84.640
Observations	3,682	3,682	3,682	3,682	3,682	3,680	3,696
R-squared	0.000	0.000	0.020	0.010	0.010	0.060	0.000
Panel B: Follow-Up Survey							
Incentive	-0.094	-1354.549***	0.110***	0.056***	-0.008	0.355***	18.012
	(0.072)	(441.580)	(0.020)	(0.012)	(0.008)	(0.064)	(15.376)
Control Mean	1.040	8008.430	0.470	0.060	0.050	2.900	12.340
Observations	2,377	2,377	2,377	2,377	2,377	2,378	2,392
R-squared	0.001	0.004	0.012	0.010	0.000	0.013	0.001

Notes: OLS regressions. Panel A: beliefs recorded in the main survey; panel B: beliefs recorded in follow-up survey conducted about two weeks after the main survey. Dependent variable: beliefs about average annual public school spending per student; column (1): stated belief divided by true value; column (2): absolute distance between stated belief and true value; column (3): dummy variable coded one if stated belief is in the incentivized range; column (4): dummy variable coded one if stated belief is in the 10-percent interval around the true value; column (5): dummy variable coded one if stated belief is below 100; column (6): confidence about belief on seven-point Likert scale (1='very unsure', 7='very sure'); column (7): difference between actual and predicted response time. Randomized experimental treatment 'Incentive': respondents offered monetary incentive for belief accuracy. Control mean: mean of the outcome variable for the control group. Regressions weighted by survey weights. Robust standard errors in parentheses. Significance levels: *** p<0.01. ** p<0.05. * p<0.10. Data source: ifo Education Survey 2017.

Table 6.4: Treatment Effects on Proxy for Online-Search Activity

	Dependent Variable: Proxy for Online-Search Engine Usage		
	Experiment on Incentive Provision	Experiment on Encouraging Online-Search Activity	Experiment on Inducing Experimenter Demand
	(1)	(2)	(3)
Treatment	0.116***	0.091***	-0.001
	(0.011)	(0.010)	(0.007)
Control Mean	0.020	0.010	0.030
Observations	3,682	4,031	3,124
R-squared	0.050	0.040	0.000

Notes: OLS regressions. Dependent variable: dummy variable coded one if respondents' belief corresponds to one of the top search-engine results. Randomized experimental treatment. 'Treatment': column (1): respondents offered monetary incentive for belief accuracy; column (2): respondents encouraged to search the internet for the correct answer; column (3): respondents informed that it is important for the experimenter that they provide a correct answer. Control mean: mean of the outcome variable for the control group. Regressions weighted by survey weights. Robust standard errors in parentheses. Significance levels: *** p<0.01. ** p<0.05. * p<0.10. Data source: column (1): ifo Education Survey 2017; column (2): ifo Education Survey 2018; column (3): ifo Education Survey 2016.

Table 6.5: Encouragement-Treatment Effects on Self-Reported Online-Search Activity and School-Spending Beliefs

	Self-Reported Online-Search Activity	Belief Relative to True Value	Distance to True Value	Belief within Incentivized Range	Belief within 10-percent Range	Extreme Values Below 100	Confidence	Response Time
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Encouragement	0.137***	0.113*	-488.511	0.087***	0.054***	0.002	0.615***	24.851***
	(0.013)	(0.064)	(371.087)	(0.023)	(0.010)	(0.013)	(0.074)	(7.107)
Control Mean	0.010	0.700	6778.190	0.440	0.030	0.080	2.580	6.180
Observations	4,045	4,031	4,031	4,031	4,031	4,031	4,046	4,046
R-squared	0.050	0.000	0.000	0.010	0.010	0.000	0.040	0.000

Notes: OLS regressions. Dependent variable: beliefs about average annual public school spending per student; column (1): self-reported online search activity; (2): stated belief divided by true value; column (3): absolute distance between stated belief and true value; column (4): dummy variable coded one if stated belief is in the incentivized range; column (5): dummy variable coded one if stated belief is in the 10-percent interval around the true value; column (6): dummy variable coded one if stated belief is below 100; column (7): confidence about belief on seven-point Likert scale (1='very unsure', 7='very sure'); column (8): difference between actual and predicted response time. Randomized experimental treatment 'Encouragement': respondents encouraged to search for accurate answer on the internet. Control mean: mean of the outcome variable for the control group. Regressions weighted by survey weights. Robust standard errors in parentheses. Significance levels: *** p<0.01. ** p<0.05. * p<0.10. Data source: ifo Education Survey 2018.

Table 6.6: Demand-Treatment Effects on School-Spending Beliefs

	Belief Relative to True Value	Distance to True Value	Belief within Incentivized Range	Belief within 10-percent Range	Extreme Values Below 100	Confidence	Response Time
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Demand	0.073	559.019*	-0.016	-0.001	0.002	-0.040	0.974
	(0.059)	(323.720)	(0.017)	(0.004)	(0.012)	(0.060)	(3.343)
Control Mean	0.680	6485.190	0.230	0.010	0.090	2.590	8.170
Observations	3,124	3,124	3,124	3,124	3,124	3,134	3,302
R-squared	0.001	0.001	0.000	0.000	0.000	0.000	0.000

Notes: OLS regressions. Dependent variable: beliefs about average annual public school spending per student; column (1): stated belief divided by true value; column (2): absolute distance between stated belief and true value; column (3): dummy variable coded one if stated belief is in the incentivized range; column (4): dummy variable coded one if stated belief is in the 10-percent interval around the true value; column (5): dummy variable coded one if stated belief is below 100; column (6): confidence about belief on seven-point Likert scale (1='very unsure', 7'very sure'); column (7): difference between actual and predicted response time. Randomized experimental treatment. 'Demand': respondents informed that it is important for the experimenter that they provide a correct answer. Control mean: mean of the outcome variable for the control group. Regressions weighted by survey weights. Robust standard errors in parentheses. Significance levels: *** p<0.01. ** p<0.05. * p<0.10. Data source: ifo Education Survey 2016.

Appendix

Figure A6.1: Google Search Results: School Spending

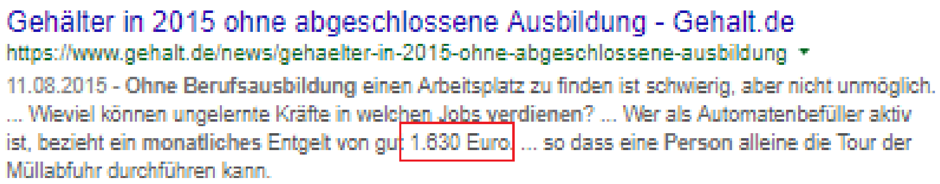
Panel A: Search for 'Bildungsausgaben pro Schueler pro Jahr'

Panel B: Search for 'Ausgaben pro Schueler pro Jahr an oeffentlichen allgemeinbildenden Schulen'

Notes: Screenshots of Google search results of the indicated keywords.

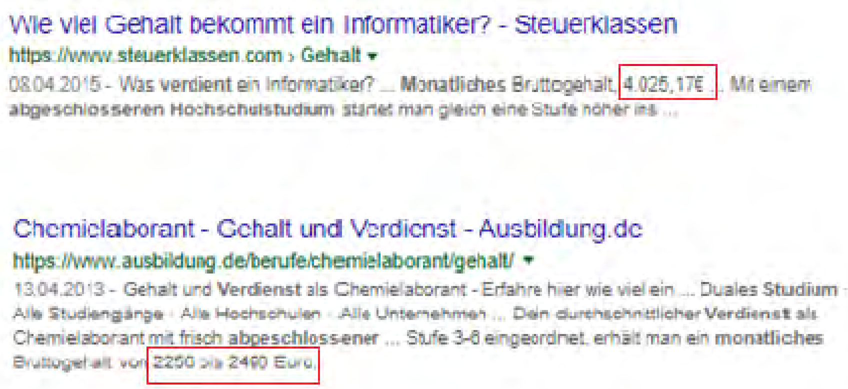
Figure A6.2: Google Search Results: Earnings by Professional Degree

Panel A: Earnings of Persons without Professional Degree: Search for 'Monatlicher Verdienst Personen ohne abgeschlossene Berufsausbildung'



Gehälter in 2015 ohne abgeschlossene Ausbildung - Gehalt.de
https://www.gehalt.de/news/gehaelter-in-2015-ohne-abgeschlossene-ausbildung
11.08.2015 - Ohne Berufsausbildung einen Arbeitsplatz zu finden ist schwierig, aber nicht unmöglich. ... Wieviel können ungelernte Kräfte in welchen Jobs verdienen? ... Wer als Automatenbefüller aktiv ist, bezieht ein monatliches Entgelt von gut 1.830 Euro ... so dass eine Person alleine die Tour der Müllabfuhr durchführen kann.

Panel B: Earnings of Persons with University Degree: Search for 'Monatlicher Verdienst Personen mit abgeschlossenem Hochschulstudium'



Wie viel Gehalt bekommt ein Informatiker? - Steuerklassen
https://www.steuerklassen.com › Gehalt
08.04.2015 - Was verdient ein Informatiker? ... Monatliches Bruttogehalt 4.025,17€ ... Mit einem abgeschlossenen Hochschulstudium startet man gleich eine Stufe höher ins ...

Chemielaborant - Gehalt und Verdienst - Ausbildung.de
https://www.ausbildung.de/berufe/chemielaborant/gehalt/
13.04.2013 - Gehalt und Verdienst als Chemielaborant - Erfahre hier wie viel ein ... Duales Studium · Alle Studiengänge · Alle Hochschulen · Alle Unternehmen ... Den durchschnittlicher Verdienst als Chemielaborant mit frisch abgeschlossener ... Stufe 3-6 eingeordnet, erhält man ein monatliches Bruttogehalt von 2250 bis 2490 Euro.

Panel C: Earnings of Persons with Different Professional Degrees: Search for 'Monatlicher Verdienst mit unterschiedlichem Bildungsabschluss'

[PDF] Verdienste auf einen Blick - Statistisches Bundesamt
https://www.destatis.de/.../VerdiensteArbeitskosten/Arbeitnehmerverdienste/Broschuer...
23.04.2017 - ... auf einen Blick. 2017. 5. Die Broschüre basiert auf unterschiedlichen Datenquellen, und .... Verdienst der Person, die genau in der Mitte steht, ist der Median- verdienst. Die eine Hälfte ... mer ohne einen beruflichen Bildungsabschluss einen Niedrig- lohn. .... gesetzlicher monatlicher Mindestlohn von 1.923 Euro. Während.

Gehälter in 2015 ohne abgeschlossene Ausbildung - Gehalt.de
https://www.gehalt.de/news/gehaelter-in-2015-ohne-abgeschlossene-ausbildung
11.08.2015 - Wieviel können ungelernte Kräfte in welchen Jobs verdienen? ... Ein /-e Taxifahrer /-in kann mit einem monatlichen Einkommen von 1.430 Euro rechnen. ... so dass eine Person alleine die Tour der Müllabfuhr durchführen kann. ... Klar ist, dass ein junger Mensch ohne qualifizierten Bildungsabschluss länger braucht, um ...

Notes: Screenshots of Google search results of the indicated keywords.

Table A6.1: Wording of the Survey Questions

No.	Subgroup	Wording of question
Incentive Experiments (2017 Survey Wave)		
10	Control	Based on your best guess, what is the average amount of money spent each year for a child in public schools in Germany?
	Treatment 'Incentive'	On the following screen, you will be asked another guess question. If your guess is roughly correct, you will receive an additional reward worth about half of the reward you get for participating in this survey. (next screen) Based on your best guess, what is the average amount of money spent each year for a child in public schools in Germany?
11	Control	Persons with a professional degree (apprenticeship) currently earn on average 1,850 Euro net per month (full-time position). What is your best guess, how much do the following groups with lower resp. higher educational attainment earn on average?
	Treatment 'Incentive'	On the following screen, you will be asked another guess question. If your guess is roughly correct, you will receive an additional reward worth about half of the reward you get for participating in this survey. (next screen) Persons with a professional degree (apprenticeship) currently earn on average 1,850 Euro net per month (full-time position). What is your best guess, how much do the following groups with lower resp. higher educational attainment earn on average?
Experiment on Encouraging Online-Search Activity (2018 Survey Wave)		
9	Control	Based on your best guess, what is the average amount of money spent each year for a child in public schools in Germany?
	Treatment 'Encouragement'	As an exception for this question, you can search the internet for the right answer to improve your guess, for example by using an internet-search engine. Based on your best guess, what is the average amount of money spent each year for a child in public schools in Germany?
Experiment on Inducing Experimenter-Demand Effects (2016 Survey Wave)		
10	Control	Based on your best guess, what is the average amount of money spent each year for a child in public schools in Germany?
	Treatment 'Demand'	As you might know, government institutions collect a variety of key statistics about schools. We are interested in discovering whether the public is familiar with these key statistics. On the next screen, we will ask you a question about such a key statistic, to which there are correct and incorrect answers. In order for your response to be informative for us, it is very important that you answer this question as accurately as possible. Based on your best guess, what is the average amount of money spent each year for a child in public schools in Germany?

Notes: Own translation from the German original. No.: consecutive ordering of the question in the ifo Education Survey of the respective wave. Subgroup: specific control or treatment group that received the respective question.

Table A6.2: Who Acquires Additional Information about the Incentive Scheme?

	Dependent Variable: Belief within 10-Percent Range			
	Earnings-Beliefs (1)		School-Spending Belief (2)	
Middle School Degree	-0.016	(0.036)	-0.021	(0.038)
University entrance Degree	-0.029	(0.040)	-0.013	(0.042)
Age	0.008***	(0.001)	0.010***	(0.001)
Monthly household Income (1000 €)	-0.005	(0.011)	0.002	(0.011)
Female	0.039	(0.029)	0.025	(0.030)
Born in Germany	-0.039	(0.057)	-0.026	(0.061)
Partner in Household	0.013	(0.033)	-0.056*	(0.034)
Lives in West Germany	0.014	(0.036)	0.012	(0.039)
City size = 100,000	-0.014	(0.031)	-0.007	(0.030)
Has Children	0.002	(0.034)	0.040	(0.034)
Parent(s) with University Degree	0.068**	(0.032)	-0.035	(0.033)
Full-time Employed	-0.060**	(0.030)	-0.078**	(0.032)
Part-time Employed	-0.029	(0.045)	0.040	(0.045)
Self-Employed	-0.101	(0.070)	-0.026	(0.070)
Unemployed	-0.136**	(0.054)	0.010	(0.065)
Works in Education Sector	-0.026	(0.042)	0.055	(0.049)
Risk Tolerance (11-Point Scale)	-0.003	(0.006)	0.004	(0.006)
Patience (11-Point Scale)	0.000	(0.007)	-0.001	(0.006)
Constant	-0.010	(0.099)	-0.073	(0.113)
Observations	1,837		1,837	
R-squared	0.12		0.17	

Notes: OLS regressions. Sample: respondents assigned to the incentive-treatment group in the respective experiment. Dependent variable: dummy variable coded one if respondent clicked on the information button. Regressions weighted by survey weights. Robust standard errors in parentheses. Significance levels: *** p<0.01. ** p<0.05. * p<0.10. Data source: ifo Education Survey 2017.

Table A6.3: Who Reports to Increase Effort in Response to Incentive Provision?

	'Incentives Increased my Effort.'			
	(Strongly) Agree (1)		(Strongly) Disgree (2)	
Middle School Degree	0.024	(0.032)	0.006	(0.029)
University entrance Degree	0.019	(0.035)	0.026	(0.032)
Age	-0.002**	(0.001)	0.000	(0.001)
Monthly household Income (1000 €)	0.002	(0.010)	-0.014*	(0.008)
Female	-0.063**	(0.026)	0.026	(0.023)
Born in Germany	-0.075	(0.058)	0.076	(0.052)
Partner in Household	0.057**	(0.029)	-0.013	(0.026)
Lives in West Germany	0.019	(0.032)	0.007	(0.029)
City size = 100,000	-0.008	(0.026)	0.004	(0.024)
Has Children	-0.045	(0.030)	0.056**	(0.027)
Parent(s) with University Degree	-0.067**	(0.027)	0.038	(0.026)
Full-time Employed	-0.025	(0.028)	0.037	(0.025)
Part-time Employed	-0.028	(0.038)	0.032	(0.036)
Self-Employed	0.074	(0.062)	0.018	(0.059)
Unemployed	-0.055	(0.053)	0.056	(0.052)
Works in Education Sector	0.044	(0.040)	-0.012	(0.034)
Risk Tolerance (11-Point Scale)	0.019***	(0.005)	-0.008*	(0.005)
Patience (11-Point Scale)	0.008	(0.006)	0.000	(0.005)
Constant	0.640***	(0.097)	0.163*	(0.089)
Observations	2,752		2,752	
R-squared	0.04		0.01	

Notes: OLS regressions. Sample: respondents assigned to the incentive-treatment group in at least one experiment. Dependent variable: column (1): dummy variable coded one if respondent 'strongly agrees' or 'somewhat agrees' with the statement that the prospect of earning more tokens encouraged them to put more effort in their answers; column (2): dummy variable coded one if respondent 'strongly disagrees' or 'somewhat disagrees' with the statement. Regressions weighted by survey weights. Robust standard errors in parentheses. Significance levels: *** p<0.01. ** p<0.05. * p<0.10. Data source: ifo Education Survey 2017.

Table A6.4: Heterogeneous Incentive Effects by Sociodemographic Subgroups

	Dependent Variable: Belief within 10-Percent Range					
	Earnings Beliefs w\o Degree (1)		Earnings Beliefs w University Degree (2)		School-Spending Belief (3)	
Educational Attainment: Baseline (no/basic degree)	-0.015	(0.032)	-0.070*	(0.039)	0.020	(0.015)
Middle School Degree	0.050	(0.042)	0.078	(0.050)	0.032	(0.021)
University Entrance Degree	-0.001	(0.042)	0.128**	(0.050)	0.047**	(0.022)
Age: Baseline (below 45)	0.023	(0.023)	0.031	(0.028)	0.086***	(0.016)
Aged 45-65	-0.055	(0.034)	-0.021	(0.041)	-0.043**	(0.020)
Aged above 65	-0.010	(0.053)	-0.155**	(0.064)	-0.106***	(0.025)
Household Income: Baseline (below median)	-0.010	(0.022)	-0.036	(0.028)	0.047***	(0.012)
Income above Median	0.022	(0.034)	0.057	(0.041)	-0.006	(0.018)
Gender: Baseline (male)	-0.012	(0.025)	0.004	(0.028)	0.052***	(0.014)
Female	0.027	(0.033)	-0.030	(0.041)	-0.015	(0.018)
Country of Birth: Baseline (not Born in Ger.)	0.001	(0.080)	0.018	(0.080)	0.003	(0.072)
Born in Germany	-0.001	(0.082)	-0.032	(0.083)	0.043	(0.073)
Partner in Household: Baseline (no)	-0.007	(0.026)	-0.017	(0.033)	0.062***	(0.013)
Partner in Household	0.011	(0.034)	0.010	(0.042)	-0.030*	(0.018)
Area of Residence: Baseline (East Germany)	0.044	(0.036)	-0.024	(0.045)	0.014	(0.015)
West Germany	-0.056	(0.041)	0.016	(0.050)	0.037**	(0.018)
City Size: Baseline (Population<100,000)	-0.003	(0.021)	-0.007	(0.025)	0.056***	(0.012)
Population >=100,000	0.004	(0.034)	-0.019	(0.043)	-0.036**	(0.018)
Children: Baseline (Has no Children)	0.005	(0.025)	0.019	(0.029)	0.075***	(0.014)
Has Children	-0.009	(0.034)	-0.051	(0.041)	-0.053***	(0.018)
School-Aged Children: Baseline (no)	-0.001	(0.018)	-0.014	(0.023)	0.039***	(0.010)
Has School-Aged Children	0.005	(0.043)	0.016	(0.052)	0.036	(0.022)
Parental Education: Baseline (Parents w/o Uni. Degree)	0.003	(0.020)	-0.007	(0.025)	0.045***	(0.010)
Parent(s) w/ University Degree	-0.013	(0.035)	-0.014	(0.044)	0.000	(0.021)
Works in Education Sector: Baseline (no)	0.005	(0.017)	-0.020	(0.022)	0.041***	(0.009)
Works in Education Sector	-0.063	(0.065)	0.095	(0.070)	0.044	(0.031)
Risk Tolerance: Baseline (below Median)	0.000	(0.021)	-0.013	(0.026)	0.039***	(0.012)
Risk Tolerance above Median	-0.001	(0.035)	0.005	(0.043)	0.016	(0.018)
Patience: Baseline (below Median)	0.000	(0.020)	-0.012	(0.024)	0.053***	(0.011)
Patience above Median	-0.005	(0.037)	-0.004	(0.045)	-0.027	(0.020)
Information Acquisition: Baseline (not Clicked)	-0.003	(0.018)	0.015	(0.023)	0.038***	(0.010)
Clicked Info Button about Incentive Scheme	0.007	(0.026)	-0.079**	(0.031)	0.016	(0.016)

Notes: OLS regressions. Dependent variable: dummy variable coded one if stated belief is in the 10-percent interval around the true value. Coefficients in baseline rows represent incentive-treatment effects for respective sociodemographic baseline group (coefficient β_1 of equation (2)). Coefficients in other rows represent interaction terms between treatment and respective subgroup indicator (coefficient β_3 of equation (2)). Earnings-beliefs regressions control for treatment status in school-spending experiment. Regressions weighted by survey weights. Robust standard errors in parentheses. Significance levels: *** p<0.01. ** p<0.05. * p<0.10. Data source: ifo Education Survey 2017.

Table A6.5: Participation in the Follow-Up Survey

	Participation in Follow-Up Survey					
	(1)		(2)		(3)	
Treatment status in main survey						
Incentive in Earnings Experiment	0.020	(0.018)			0.015	(0.027)
Incentive in Spending Experiment			0.012	(0.018)	0.006	(0.026)
Incentives in Earnings and Spending Experiment					0.012	(0.037)
Covariates						
Middle School Degree	0.000	(0.023)	-0.001	(0.023)	0.000	(0.023)
University Entrance Degree	-0.002	(0.028)	-0.004	(0.028)	-0.003	(0.028)
Age	0.006***	(0.001)	0.006***	(0.001)	0.006***	(0.001)
Monthly Household Income (1000 €)	0.001	(0.008)	0.001	(0.008)	0.001	(0.008)
Female	0.025	(0.020)	0.026	(0.020)	0.025	(0.020)
Born in Germany	0.038	(0.048)	0.040	(0.048)	0.038	(0.048)
Partner in Household	-0.021	(0.021)	-0.022	(0.021)	-0.022	(0.021)
Lives in West Germany	-0.004	(0.024)	-0.004	(0.024)	-0.004	(0.024)
City size = 100,000	-0.047**	(0.020)	-0.047**	(0.020)	-0.047**	(0.020)
Has Children	-0.001	(0.023)	-0.001	(0.023)	-0.001	(0.023)
Parent(s) with University Degree	-0.013	(0.022)	-0.013	(0.022)	-0.013	(0.022)
Full-Time Employed	0.073***	(0.023)	0.072***	(0.023)	0.073***	(0.023)
Part-Time Employed	0.055**	(0.027)	0.054**	(0.027)	0.055**	(0.027)
Self-Employed	0.019	(0.050)	0.020	(0.050)	0.020	(0.050)
Unemployed	0.091**	(0.042)	0.089**	(0.042)	0.091**	(0.042)
Works in Education Sector	-0.053	(0.035)	-0.053	(0.035)	-0.052	(0.035)
Risk Tolerance (11-Point Scale)	-0.014***	(0.004)	-0.014***	(0.004)	-0.014***	(0.004)
Patience (11-Point Scale)	0.006	(0.004)	0.006	(0.004)	0.006	(0.004)
Constant	0.373***	(0.077)	0.376***	(0.077)	0.369***	(0.077)
Observations	3,673		3,673		3,673	
R-squared	0.006		0.006		0.006	

Notes: OLS regressions. Dependent variable: dummy variable coded one if respondent participated in the follow-up survey. Treatment status in the main survey: column (1): incentive-treatment in earnings experiment; column (2): incentive-treatment in spending experiment; column (3): incentive treatment in both experiments and their interaction. Regressions weighted by survey weights. Robust standard errors in parentheses. Significance levels: *** p<0.01. ** p<0.05. * p<0.10. Data source: ifo Education Survey 2017.

Table A6.6: Summary Statistics and Balancing Tests of Incentive Experiments: Follow-up Survey

	Earnings Experiment			School-Spending Experiment		
	Control Mean (1)	Treatment Mean (2)	Difference (3)	Control Mean (4)	Treatment Mean (5)	Difference (6)
Highest educational attainment						
No Degree/Basic Degree	0.211	0.241	0.030	0.225	0.228	0.003
Middle School Degree	0.404	0.389	-0.015	0.403	0.389	-0.014
University Entrance Degree	0.385	0.370	-0.015	0.372	0.383	0.011
Age	46.568	47.272	0.704	46.936	46.913	-0.022
Monthly Household Income (€)	2403.500	2386.800	-16.700	2438.000	2351.400	-86.700
Female	0.509	0.512	0.002	0.508	0.513	0.004
Born in Germany	0.950	0.963	0.013	0.954	0.959	0.005
Partner in Household	0.600	0.586	-0.015	0.606	0.579	-0.027
Lives in West Germany	0.752	0.780	0.028	0.777	0.756	-0.022
City Size >= 100,000	0.363	0.357	-0.006	0.345	0.376	0.032
Has Children	0.553	0.540	-0.013	0.555	0.538	-0.017
Parent(s) with University Degree	0.312	0.290	-0.021	0.306	0.296	-0.010
Employment Status						
Full-Time Employed	0.406	0.401	-0.005	0.429	0.379	-0.050**
Part-Time Employed	0.138	0.125	-0.012	0.128	0.135	0.007
Self-Employed	0.046	0.037	-0.008	0.046	0.037	-0.008
Unemployed	0.054	0.045	-0.009	0.046	0.054	0.008
House Wife/Husband/Retired/Ill/Student	0.059	0.064	0.005	0.062	0.061	-0.001
Works in Education Sector	0.082	0.090	0.008	0.090	0.081	-0.009
Risk Tolerance (11-Point Scale)	4.362	4.328	-0.034	4.335	4.355	0.020
Patience (11-Point Scale)	6.216	6.283	0.067	6.199	6.301	0.102
Item Non-Response: Beliefs	0.002	0.002	0.000	0.010	0.006	-0.005
Observations	1,251	1,284		1,275	1,260	

Notes: Group means. 'Difference' displays the difference in means between the respective control and treatment groups. Significance levels of 'Difference' stem from linear regressions of the background variables on the respective treatment dummies. Significance levels: *** p<0.01. ** p<0.05. * p<0.10. Data source: ifo Education Survey 2017.

Table A6.7: Incentive Effects on Earnings Beliefs in Follow-Up Survey

	Belief Relative to True Value	Distance to True Value	Belief within Incentivized Range	Belief within 10-percent Range	Extreme Values Below 100	Confidence	Response Time
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Panel A: Beliefs on Earnings without a Degree							
Incentive	0.018	3.501	0.031	0.049***	0.010*		
	(0.036)	(49.583)	(0.019)	(0.019)	(0.006)		
Control mean	0.870	333.550	0.660	0.280	0.020		
Observations	2,418	2,418	2,418	2,418	2,418		
R-squared	0.001	0.000	0.002	0.003	0.001		
Panel B: Beliefs on Earnings with University Degree							
Incentive	0.009	53.055	-0.023	-0.006	0.008	0.078	6.278
	(0.032)	(82.548)	(0.017)	(0.020)	(0.006)	(0.058)	(0.463)
Control mean	1.010	706.830	0.790	0.440	0.020	3.770	52.500
Observations	2,418	2,418	2,418	2,418	2,418	2,417	2,419
R-squared	0.000	0.000	0.001	0.001	0.001	0.002	0.000

Notes: OLS regressions. Dependent variable (recorded in follow-up survey conducted about two weeks after the main survey): column (1): stated belief divided by true value; column (2): absolute distance between stated belief and true value; column (3): dummy variable coded one if stated belief is in the incentivized range; column (4): dummy variable coded one if stated belief is in the 10-percent interval around the true value; column (5): dummy variable coded one if stated belief is below 100; column (6): confidence about belief on seven-point Likert scale (1='very unsure', 7='very sure'); column (7): difference between actual and predicted response time. Randomized experimental treatment 'Incentive': respondents offered monetary incentive for belief accuracy. Control mean: mean of the outcome variable for the control group in the follow-up survey. Panel A: beliefs about net average monthly earnings of full-time employed persons without any professional degree. Panel B: beliefs about net average monthly earnings of full-time employed persons with a university degree. Since both earnings beliefs were elicited on the same screen, confidence and response time were recorded only once for earnings beliefs in general. Regressions control for treatment status in the incentive experiment on school-spending beliefs. Robust standard errors in parentheses. Significance levels: *** p<0.01. ** p<0.05. * p<0.10. Data source: ifo Education Survey 2017.

Table A6.8: Summary Statistics and Balancing Tests of Additional Experiments

	Experiment on Encouraging Online-Search Activity			Experiment on Inducing Experimenter-Demand		
	Control Mean (1)	Treatment Mean (2)	Difference (3)	Control Mean (4)	Treatment Mean (5)	Difference (6)
Highest Educational Attainment						
No Degree/Basic Degree	0.368	0.381	0.014	0.401	0.380	-0.021
Middle School Degree	0.300	0.308	0.008	0.293	0.315	0.022
University Entrance Degree	0.332	0.311	-0.022	0.306	0.305	-0.001
Age	50.202	50.908	0.706	50.702	50.321	-0.381
Monthly Household Income (€)	2452	2467	15	2120	2201	81
Female	0.505	0.517	0.011	0.530	0.497	-0.033
Born in Germany	0.958	0.962	0.004	0.944	0.954	0.010
Partner in Household	0.589	0.575	-0.014	0.556	0.570	0.014
Lives in West Germany	0.684	0.651	-0.033	0.803	0.796	-0.008
City Size = 100,000	0.307	0.323	0.016	0.328	0.315	-0.013
Has Children	0.586	0.592	0.007	0.592	0.601	0.009
Parent(s) with University Degree	0.286	0.291	0.005	0.243	0.243	0.000
Employment Status						
Full-Time Employed	0.390	0.376	-0.013	0.343	0.334	-0.009
Part-Time Employed	0.160	0.156	-0.003	0.122	0.128	0.006
Self-Employed	0.053	0.050	-0.002	0.030	0.031	0.001
Unemployed	0.027	0.032	0.005	0.051	0.060	0.009
House Wife/Husband/Retired/Ill/Student	0.371	0.385	0.014	0.454	0.448	-0.006
Works in Education Sector	0.107	0.108	0.001	0.078	0.079	0.001
Risk tolerance (11-Point Scale)	4.391	4.238	-0.152	4.312	4.252	-0.060
Patience (11-Point Scale)	6.215	6.336	0.121	5.961	5.941	-0.020
Item Non-Response: Beliefs	0.005	0.003	-0.002	0.081	0.073	-0.008
Observations	1,979	2,067		1,642	1,660	

Notes: Group means. 'Difference' displays the difference in means between the respective control and treatment groups. Significance levels of 'Difference' stem from linear regressions of the background variables on the respective treatment dummies. Significance levels: *** p<0.01. ** p<0.05. * p<0.10. Data source: columns (1)-(3): ifo Education Survey 2018; columns (4)-(6): ifo Education Survey 2016.

Table A6.9: Stacked Estimation of Incentive and Encouragement Effects on School-Spending Beliefs

	Belief Relative to True Value	Distance to True Value	Belief within Incentivized Range	Belief within 10-percent Range	Extreme Values Below 100	Confidence	Response Time
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Treatment	0.112	-720.657	0.132***	0.044***	-0.064***	0.768***	86.068**
	(0.078)	(470.255)	(0.021)	(0.009)	(0.015)	(0.068)	(38.010)
Treatment x Wave 2018	0.001	232.146	-0.045	0.010	0.066***	-0.153	-61.217
	(0.101)	(595.175)	(0.031)	(0.014)	(0.019)	(0.101)	(38.662)
Wave 2018	-0.074	-812.739**	0.077***	0.007	-0.082***	0.162**	-78.467***
	(0.068)	(384.815)	(0.022)	(0.008)	(0.014)	(0.064)	(5.495)
Control Mean	0.740	7175.060	0.410	0.030	0.120	2.500	44.460
Observations	7,713	7,713	7,713	7,713	7,713	7,726	7,742
R-squared	0.000	0.000	0.020	0.010	0.010	0.050	0.020

Notes: OLS regressions. Dependent variable: beliefs about average annual public school spending per student; column (1): stated belief divided by true value; column (2): absolute distance between stated belief and true value; column (3): dummy variable coded one if stated belief is in the incentivized range; column (4): dummy variable coded one if stated belief is in the 10-percent interval around the true value; column (5): dummy variable coded one if stated belief is below 100; column (6): confidence about belief on seven-point Likert scale (1='very unsure', 7='very sure'); column (7): difference between actual and predicted response time. Randomized experimental treatment 'Treatment': wave 2017: respondents offered monetary incentive for belief accuracy; wave 2018: respondents encouraged to search the internet for the correct answer. Control mean: mean of the outcome variable for the control group. Regressions weighted by survey weights. Robust standard errors in parentheses. Significance levels: *** p<0.01. ** p<0.05. * p<0.10. Data source: ifo Education Survey 2017 and 2018.

Description of Additional Survey Waves

While our incentive experiments were implemented in the 2017 wave of the ifo Education Survey (described in section 6.2.1), the other two experiments were implemented in the 2016 and 2018 waves of the ifo Education Survey, respectively. The ifo Education Survey 2018 was fielded by Kantar Public between May and June 2018. The sample consists of 4,046 respondents who are representative for the German voting-age population (18 years and older). Respondents completed the survey on their own digital devices. Median completion time was 15 minutes.

The ifo Education Survey 2016 was fielded by Kantar Public between April and June 2016. The sample consists of 3,302 respondents who are representative for the German voting-age population (18 years and older). In contrast to the data of the 2017 and 2018 waves employed in this chapter, respondents were drawn in two strata. Respondents with access to the internet were recruited via an online platform and completed the survey on their own digital devices. Respondents without internet access were polled at their homes by trained interviewers. The interviewers equipped these respondents with tablet computers and asked them to fill out the survey on these devises. Median completion time was 18 minutes.

As with the 2017 data, both the 2016 and the 2018 analyses employ survey weights to achieve representativeness for the German population. The weights match characteristics of the overall German population with respect to age, gender, federal state, school degree, and municipality size. Performing the analyses without using the survey weights does not change the qualitative results.

Bibliography

Abdulkadiroğlu, Atila, Joshua Angrist, and Parag Pathak (2014). “The Elite Illusion: Achievement Effects at Boston and New York Exam Schools”. *Econometrica* 82(1), 137–196.

Acemoglu, D. and David H. Autor (2010). “Skills, Tasks and Technologies: Implications for Employment and Earnings”. *Handbook of Labor Economics*.

Akerlof, George A. and Rachel E. Kranton (2000). “Economics and Identity”. *The Quarterly Journal of Economics* 115(3), 715–753.

– (2010). *Identity Economics: How Our Identities Shape Our Work, Wages, and Well-Being*. Princeton University Press.

Alesina, Alberto (1988). “Credibility and Policy Convergence in a Two-Party System with Rational Voters”. *The American Economic Review* 78(4), 796–805.

Alesina, Alberto, Paola Giuliano, and Nathan Nunn (2013). “On the Origins of Gender Roles: Women and the Plough”. *The Quarterly Journal of Economics* 128(2), 469–530.

Alesina, Alberto and Eliana La Ferrara (2005). “Preferences for Redistribution in the Land of Opportunities”. *Journal of Public Economics* 89(5), 897–931.

Alesina, Alberto, Armando Miano, and Stefanie Stantcheva (2018a). “Immigration and Redistribution”. *National Bureau of Economic Research Working Paper Series* No. 24733.

Alesina, Alberto, Stefanie Stantcheva, and Edoardo Teso (2018b). “Intergenerational Mobility and Preferences for Redistribution”. *American Economic Review* 108(2), 521–554.

Alon, Titan, Matthias Doepke, Jane Olmstead-Rumsey, and Michèle Tertilt (2020). “The Impact of COVID-19 on Gender Equality”. *National Bureau of Economic Research Working Paper Series* 26947.

Aloud, Monira Essa, Sara Al-Rashood, Ina Ganguli, and Basit Zafar (2020). “Information and Social Norms: Experimental Evidence on the Labor Market Aspirations of Saudi Women”. *National Bureau of Economic Research Working Paper Series* No. 26693.

Alt, Christian, Daniela Gesell, Sandra Hubert, Katrin Hüsken, Ralf Kuhnke, and Kerstin Lippert (2017). “DJI-Kinderbetreuungsreport 2017. Inanspruchnahme und Bedarfe aus Elternperspektive im Bundesländervergleich.” *Deutsches Jugendinstitut (DJI), München*.

Altonji, Joseph G. and Rebecca M. Blank (n.d.). “Chapter 48 Race and Gender in the Labor Market”. *Ashenfelter, Card (Hg.) 1999 – Handbook of Labor Economics*. Vol. 3, 3143–3259.

Andre, Peter, Carlo Pizzinelli, Christopher Roth, and Johannes Wohlfart (2019). “Subjective Models of the Macroeconomy: Evidence from Experts and a Representative Sample”. *Cesifo Working Papers* 7850.

Andresen, Martin Eckhoff and Emily Nix (2019). “What Causes the Child Penalty and How Can It Be Reduced? Evidence from Same-Sex Couples and Policy Reforms”. *Statistics Norway Discussion Paper* No. 902.

Andrew, Alison, Sarah Cattan, Monica Costa Dias, Christine Farquharson, Lucy Kraftman, Sonya Krutikova, Angus Phimister, and Almudena Sevilla (2020). “Inequalities in Children’s Experiences of Home Learning during the COVID-19 Lockdown in England”. *IFS Working Paper* W20/26.

Andrietti, Vincenzo and Xuejuan Su (2018). “The Impact of Schooling Intensity on Student Learning: Evidence from a Quasi-Experiment”. *Education Finance and Policy* 14(4), 679–701.

Angelov, Nikolay, Per Johansson, and Erica Lindahl (2015). “Parenthood and the Gender Gap in Pay”. *Journal of Labor Economics* 34(3), 545–579.

Anger, Silke, Hans Dietrich Patzina, Malte Sandner, Adrian Lerche, Sarah Bernhard, and Carina Toussaint (2020). *School Closings during the COVID-19 Pandemic: Findings from German High School Students*. Ed. by Institute for Employment. Nuremberg.

Angrist, Joshua D. and Victor Lavy (1999). “Using Maimonides’ Rule to Estimate the Effect of Class Size on Scholastic Achievement”. *The Quarterly Journal of Economics* 114(2), 533–575.

Angrist, Joshua and Victor Lavy (2009). “The Effects of High Stakes High School Achievement Awards: Evidence from a Randomized Trial”. *American Economic Review* 99(4), 1384–1414.

Ansolabehere, Stephen, Marc Meredith, and Erik Snowberg (2013). “Asking About Numbers: Why and How”. *Political Analysis* 21(1), 48–69.

Arcidiacono, Peter (2004). “Ability Sorting and the Returns to College Major”. *Journal of Econometrics* 121(1), 343–375.

Arcidiacono, Peter, V. Joseph Hotz, and Songman Kang (2012). “Modeling College Major Choices Using Elicited Measures of Expectations and Counterfactuals”. *Journal of Econometrics* 166(1), 3–16.

Arcidiacono, Peter, V. Joseph Hotz, Arnaud Maurel, and Teresa Romano (2014). “Recovering Ex Ante Returns and Preferences for Occupations using Subjective Expectations Data”. *National Bureau of Economic Research* No. 20626.

Arechar, Antonio A., Simon Gächter, and Lucas Molleman (2018). “Conducting Interactive Experiments Online”. *Experimental Economics* 21(1), 99–131.

Armantier, Olivier, Wändi Bruine de Bruin, Simon Potter, Giorgio Topa, Wilbert van der Klaauw, and Basit Zafar (2013). “Measuring Inflation Expectations”. *Annual Review of Economics* 5(1), 273–301.

Armantier, Olivier, Scott Nelson, Giorgio Topa, Wilbert van der Klaauw, and Basit Zafar (2016). “The Price Is Right: Updating Inflation Expectations in a Randomized Price Information Experiment”. *The Review of Economics and Statistics* 98(3), 503–523.

Armona, Luis, Andreas Fuster, and Basit Zafar (2019). “Home Price Expectations and Behaviour: Evidence from a Randomized Information Experiment”. *The Review of Economic Studies* 86(4), 1371–1410.

Athey, Susan and Guido Imbens (2018). *Design-based Analysis in Difference-In-Differences Settings with Staggered Adoption*. Cambridge, MA.

Attanasio, Orazio, Teodora Boneva, and Christopher Rauh (2020). “Parental Beliefs about Returns to Different Types of Investments in School Children”. *Journal of Human Resources*, 0719–10299R1.

Bach, Maximilian and Mira Fischer (2020). “Understanding the Response to High-Stakes Incentives in Primary Education”. *IZA Discussion Paper Series* 13845.

Bacher-Hicks, Andrew, Joshua Goodman, and Christine Mulhern (2021). “Inequality in Household Adaptation to Schooling Shocks: Covid-induced Online Learning Engagement in Real Time”. *Journal of Public Economics* 193, 104345.

Baker, Michael (2013). “Industrial Actions in Schools: Strikes and Student Achievement”. *The Canadian Journal of Economics / Revue canadienne d’Economique* 46(3), 1014–1036.

Barabas, Jason and Jennifer Jerit (2010). “Are Survey Experiments Externally Valid?” *American Political Science Review* 104(2), 226–242.

Barro, Robert J. (2001). “Human Capital and Growth”. *American Economic Review* 91(2), 12–17.

Barrow, Lisa and Cecilia Elena Rouse (2016). “Financial Incentives and Educational Investment: The Impact of Performance-based Scholarships on Student Time Use”. *Education Finance and Policy* 13(4), 419–448.

Bartels, Larry M. (1996). “Uninformed Votes: Information Effects in Presidential Elections”. *American Journal of Political Science* 40(1), 194–230.

– (2002). “Beyond the Running Tally: Partisan Bias in Political Perceptions”. *Political Behavior* 24(2), 117–150.

Baumert, Jürgen, Cordula Artelt, Eckhard Klieme, Michael Neubrand, Manfred Prenzel, Ulrich Schiefele, Wolfgang Schneider, Klaus-Jürgen Tillmann, and Manfred Weiß, eds. (2002). *PISA 2000 - Die Länder der Bundesrepublik Deutschland im Vergleich*. Opladen: Leske + Budrich.

– eds. (2003). *PISA 2000 — Ein differenzierter Blick auf die Länder der Bundesrepublik Deutschland*. Opladen: Leske + Budrich.

– (2009). *Programme for International Student Assessment 2000 (PISA 2000) (Version 1) [Data set]*. Berlin: IQB - Institut zur Qualitätsentwicklung im Bildungswesen. http://doi.org/10.5159/IQB_PISA_2000_v1.

Baumert, Jürgen, Cordula Artelt, Ekhard Klieme, and Petra Stanat (2001a). *PISA -Programme for International Student Assessment: Zielsetzung, theoretische Konzeption und Entwicklung von Messverfahren. In: Leistungsmessungen in Schulen (285–310)*. Weinheim: Beltz.

Baumert, Jürgen, Werner Blum, Rainer Lehmann, Detlev Leutner, and Michael Neubrand (2004). *PISA 2003. Der Bildungsstand der Jugendlichen in Deutschland: Ergebnisse des zweiten internationalen Vergleichs*. Münster: Waxmann.

Baumert, Jürgen, Cornelia Gresch, Nele McElvany, and Kai Maaz (2010). *Der Übergang von der Grundschule in die weiterführende Schule - Leistungsgerechtigkeit und regionale, soziale* *und ethnisch-kulturelle Disparitäten: Zusammenfassung der zentralen Befunde*. Berlin: Bundesministerium für bildung und Forschung.

Baumert, Jürgen, Eckhard Klieme, Michael Neubrand, Manfred Prenzel, Ulrich Schiefele, Wolfgang Schneider, Petra Stanat, Klaus-Jürgen Tillmann, and Manfred Weiß, eds. (2001b). *PISA 2000. Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich*. Opladen: Leske + Budrich.

Beaman, Lori, Esther Duflo, Rohini Pande, and Petia Topalova (2012). "Female Leadership Raises Aspirations and Educational Attainment for Girls: A Policy Experiment in India". *Science* 335(6068), 582.

Beblavý, Miroslav, Sara Beblavý, Zachary Kilhoffer, Mehtap Akgüç, and Jacquot Manon (2019). *Index of Readiness for Digital Lifelong Learning: Changing How Europeans Upgrade their Skills*. Brussels.

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

Behrman, Jere R., Susan W. Parker, Petra E. Todd, and Kenneth I. Wolpin (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.

Belfield, Chris, Teodora Boneva, Christopher Rauh, and Jonathan Shaw (2020). "What Drives Enrolment Gaps in Further Education? The Role of Beliefs in Sequential Schooling Decisions". *Economica* 87(346), 490–529.

Belot, Michèle and Dinand Webbink (2010). "Do Teacher Strikes Harm Educational Attainment of Students?" *LABOUR* 24(4), 391–406.

Benabou, Roland and Jean Tirole (2000). *Self-Confidence and Social Interactions*. Cambridge, MA.

– (2011). "Laws and Norms". *National Bureau of Economic Research Working Paper Series* No. 17579.

Benítez-Silva, Hugo, Moshe Buchinsky, Hiu Man Chan, Sofia Cheidvasser, and John Rust (2004). "How Large Is the Bias in Self-Reported Disability?" *Journal of Applied Econometrics* 19(6), 649–670.

Benjamin, Daniel J., James J. Choi, and A. Joshua Strickland (2010). "Social Identity and Preferences". *American Economic Review* 100(4), 1913–1928.

Bergbauer, Annika B., Eric A. Hanushek, and Ludger Woessmann (2018). "Testing". *National Bureau of Economic Research Working Paper Series* No. 24836.

Bertrand, Marianne (2011). "New Perspectives on Gender". *Handbook of Labor Economics*. Ed. by David Card and Orley Ashenfelter. Vol. 3. Elsevier, 1543–1590.

Bertrand, Marianne, Claudia Goldin, and Lawrence F. Katz (2010). "Dynamics of the Gender Gap for Young Professionals in the Financial and Corporate Sectors". *American Economic Journal: Applied Economics* 2(3), 228–255.

Bertrand, Marianne, Emir Kamenica, and Jessica Pan (2015). “Gender Identity and Relative Income within Households”. *The Quarterly Journal of Economics* 130(2), 571–614.

Bertrand, Marianne and Sendhil Mullainathan (2001). “Do People Mean What They Say? Implications for Subjective Survey Data”. *American Economic Review* 91(2), 67–72.

Besley, Timothy and Stephen Coate (1997). “An Economic Model of Representative Democracy”. *The Quarterly Journal of Economics* 112(1), 85–114.

Betts, Julian (2011). “The Economics of Tracking in Education”. *Handbook of the Economics of Education*. Ed. by Eric A. Hanushek, Stephen Machin, and Ludger Woessmann. Elsevier, 341–381.

Bisgaard, Martin and Rune Slothuus (2018). “Partisan Elites as Culprits? How Party Cues Shape Partisan Perceptual Gaps”. *American Journal of Political Science* 62(2), 456–469.

Björklund, Anders and Kjell G. Salvanes (2011). “Education and Family Background: Mechanisms and Policies”. *Handbook of the Economics of Education*. Ed. by Eric A. Hanushek, Stephen Machin, and Ludger Woessmann. Elsevier, 201–247.

Blanco, Mariana, Dirk Engelmann, Alexander K. Koch, and Hans-Theo Normann (2010). “Belief Elicitation in Experiments: Is there a Hedging Problem?” *Experimental Economics* 13(4), 412–438.

Blau, Francine D. and Lawrence M. Kahn (2017). “The Gender Wage Gap: Extent, Trends, and Explanations”. *Journal of Economic Literature* 55(3), 789–865.

Bleemer, Zachary and Basit Zafar (2018). “Intended College Attendance: Evidence from an Experiment on College Returns and Costs”. *Journal of Public Economics* 157, 184–211.

Blinder, Alan S. and Alan B. Krueger (2004). “What Does the Public Know About Economic Policy, and How Does It Know It?” *Brookings Papers on Economic Activity* 1, 327–397.

Boelmann, Barbara, Anna Raute, and Uta Schoenberg (2020). “Wind of Change? Cultural Determinants of Maternal Labor Supply”. *Working Paper*.

Böhme, Katrin, Dirk Richter, Sebastian Weirich, Nicole Haag, Heike Wendt, Wilfried Bos, Hans Pant, and Petra Stanat (2014). “Messen wir dasselbe? Zur Vergleichbarkeit des IQB-Ländervergleichs 2011 mit den internationalen Studien IGLU und TIMSS 2011”. *Unterrichtswissenschaft* 42, 342–365.

Boneva, Teodora and Christopher Rauh (2017). “Socio-Economic Gaps in University Enrollment: The Role of Perceived Pecuniary and Non-Pecuniary Returns”. *Cesifo Working Papers* 6756.

– (2018). “Parental Beliefs about Returns to Educational Investments—The Later the Better?” *Journal of the European Economic Association* 16(6), 1669–1711.

Bonferroni, Carlo E. (1935). *Il calcolo delle assicurazioni su gruppi di teste*. Tipografia del Senato.

Borghans, By Lex, Ron Diris, Wendy Smits, and Jannes de Vries (2020). “Should We Sort it Out Later? The Effect of Tracking Age on Long-Run Outcomes”. *Economics of Education Review* 75, 101973.

Bos, Wilfied, Irmela Tarelli, Albert Bremerich-Vos, and Knut Schwippert (2012). *Lesekompetenzen von Grundschulkindern in Deutschland im internationalen Vergleich*. Münster: Waxmann.

Bos, Wilfried, Sabine Hornberg, Karl-Heinz Arnold, Gabriele Faust, Lilian Fried, Eva-Maria Lankes, Knut Schwippert, and Renate Valtin (2010). *Internationale Grundschul-Lese-Untersuchung 2006 (IGLU 2006) [Progress in International Reading Literacy Study 2006 (PIRLS 2006)] (Version 1) [Data set]*. Berlin: IQB – Institut zur Qualitätsentwicklung im Bildungswesen. http://doi.org/10.5159/IQB_IGLU_2006_v1.

Bos, Wilfried, Eva-Maria Lankes, Manfred Prenzel, Knut Schwippert, Gerd Walther, and Renate Valtin (2007). *Internationale Grundschul-Lese-Untersuchung 2001 (IGLU 2001) [Progress in International Reading Literacy Study 2001 (PIRLS 2001)] (Version 1) [Data set]*. Berlin: IQB – Institut zur Qualitätsentwicklung im Bildungswesen.
http://doi.org/10.5159/IQB_IGLU_2001_v1.

Boschini, Anne, Astri Muren, and Mats Persson (2012). "Constructing Gender Differences in the Economics Lab". *Journal of Economic Behavior & Organization* 84(3), 741–752.

Bowles, Samuel (1998). "Endogenous Preferences: The Cultural Consequences of Markets and Other Economic Institutions". *Journal of Economic Literature* 36(1), 75–111.

Breyton, Ricarda (2018). "Grundschulempfehlung kommt zurück". *Die Welt* 2018.

Broockman, David E. and Daniel M. Butler (2017). "The Causal Effects of Elite Position-Taking on Voter Attitudes: Field Experiments with Elite Communication". *American Journal of Political Science* 61(1), 208–221.

Bullock, John G., Alan S. Gerber, Seth J. Hill, and Gregory A. Huber (2015). "Partisan Bias in Factual Beliefs about Politics". *Quarterly Journal of Political Science* 10(4), 519–578.

Burns, Tracey (2007). *Evidence in Education: Linking Research and Policy*. Knowledge management. Paris: OECD.

Bursztyn, Leonardo (2016). "Poverty and the Political Economy of Public Education Spending: Evidence from Brazil". *Journal of the European Economic Association* 14(5), 1101–1128.

Bursztyn, Leonardo, Alessandra L. González, and David Yanagizawa-Drott (2020). "Misperceived Social Norms: Women Working Outside the Home in Saudi Arabia". *American Economic Review* 110(10), 2997–3029.

Busemeyer, Marius R., Philipp Lergetporer, and Ludger Woessmann (2018). "Public Opinion and the Political Economy of Educational Reforms: A survey". *European Journal of Political Economy* 53, 161–185.

Campbell, Angus (1980). *The American Voter*. Unabridged ed. Midway reprints. Chicago Ill.: University of Chicago Press.

Canaan, Serena (2020). "The Long-Run Effects of Reducing Early School Tracking". *Journal of Public Economics* 187, 104206.

Cappelen, Alexander W., Ranveigh Falch, and Bertil Tungodden (2019). "The Boy Crisis: Experimental Evidence on the Acceptance of Males Falling Behind". *NHH Dept. of Economics Discussion Paper* No. 06/2019.

Card, David (1999). “The Causal Effect of Education on Earnings”. *Handbook of Labor Economics*. Ed. by Orley C. Ashenfelter and David Card. Vol. 3. Elsevier, 1801–1863.

– (2001). “Estimating the Return to Schooling: Progress on Some Persistent Econometric Problems”. *Econometrica* 69(5), 1127–1160.

Carlsson, Magnus, Gordon Dahl, and Dan-Olof Rooth (2020). “Backlash in Policy Attitudes After the Election of Extreme Political Parties”. *National Bureau of Economic Research Working Paper Series* No. 21062.

Cattaneo, Maria, Philipp Lergetporer, Guido Schwerdt, Katharina Werner, Ludger Woessmann, and Stefan C. Wolter (2020). “Information Provision and Preferences for Education Spending: Evidence from Representative Survey Experiments in Three Countries”. *European Journal of Political Economy* 63, 101876.

Cavallo, Alberto, Guillermo Cruces, and Ricardo Perez-Truglia (2017). “Inflation Expectations, Learning, and Supermarket Prices: Evidence from Survey Experiments”. *American Economic Journal: Macroeconomics* 9(3), 1–35.

Charles, Kerwin Kofi, Jonathan Guryan, and Jessica Pan (2018). “The Effects of Sexism on American Women: The Role of Norms vs. Discrimination”. *National Bureau of Economic Research Working Paper Series* No. 24904.

Checchi, Daniele and Luca Flabbi (2013). “Intergenerational Mobility and Schooling Decisions in Germany and Italy: The Impact of Secondary School Tracks”. *Rivista di Politica Economica* (3), 7–57.

Chetty, Raj, John N. Friedman, Nathaniel Hendren, and Michael Stepner (2020). “The Economic Impacts of COVID-19: Evidence from a New Public Database Built Using Private Sector Data”. *National Bureau of Economic Research* No. 27431.

Chetty, Raj, John N. Friedman, and Jonah E. Rockoff (2014). “Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood”. *American Economic Review* 104(9), 2633–2679.

Chetty, Raj, Adam Looney, and Kory Kroft (2009). “Salience and Taxation: Theory and Evidence”. *American Economic Review* 99(4), 1145–1177.

Cialdini, Robert B. and Noah J. Goldstein (2004). “Social Influence: Compliance and Conformity”. *Annual Review of Psychology* 55, 591–621.

Cialdini, Robert B. and Melanie R. Trost (1998). “Social Influence: Social Norms, Conformity and Compliance”. *The Handbook of Social Psychology*. New York, NY, US: McGraw-Hill, 151–192.

Clifford, Scott and Jennifer Jerit (2014). “Is There a Cost to Convenience? An Experimental Comparison of Data Quality in Laboratory and Online Studies”. *Journal of Experimental Political Science* 1(2), 120–131.

Cohen, Geoffrey L. (2003). “Party Over Policy: The Dominating Impact of Group Influence on Political Beliefs”. *Journal of Personality and Social Psychology* 85(5), 808–822.

Coibion, Olivier, Yuriy Gorodnichenko, and Saten Kumar (2018). “How Do Firms Form Their Expectations? New Survey Evidence”. *American Economic Review* 108(9), 2671–2713.

Cortes, Patricia and Jessica Pan (2020). “Chidren and the Remaining Gender Gaps in the Labor Market”. *National Bureau of Economic Research Working Paper Series* No. 27980.

Cotter, David, Joan M. Hermsen, and Reeve Vanneman (2011). “The End of the Gender Revolution? Gender Role Attitudes from 1977 to 2008”. *American Journal of Sociology* 117(1), 259–289.

Cruces, Guillermo, Ricardo Perez-Truglia, and Martin Tetaz (2013). “Biased Perceptions of Income Distribution and Preferences for Redistribution: Evidence from a Survey Experiment”. *Journal of Public Economics* 98, 100–112.

Cunha, Flávio, Irma Elo, and Jennifer Culhane (2013). “Eliciting Maternal Expectations about the Technology of Cognitive Skill Formation”. *National Bureau of Economic Research Working Paper Series* No. 19144.

Cunha, Flavio and James Heckman (2007). “The Technology of Skill Formation”. *American Economic Review* 97(2), 31–47.

Cunha, Flavio, James J. Heckman, Lance Lochner, and Dimitriy V. Masterov (2006). “Interpreting the Evidence on Life Cycle Skill Formation.” *Handbook of the Economics of Education* (1), 697–812.

Cunha, Flavio, James J. Heckman, and Susanne M. Schennach (2010). “Estimating the Technology of Cognitive and Noncognitive Skill Formation”. *Econometrica* 78(3), 883–931.

Currie, Janet and Enrico Moretti (2003). “Mother’s Education and the Intergenerational Transmission of Human Capital: Evidence from College Openings”. *Quarterly Journal of Economics* 118(4), 1495–1532.

Cygan-Rehm, Kamila (2018). “Is Additional Schooling Worthless? Revising Zero Returns to Compulsory Schooling in Germany”. *Cesifo Working Papers* (No. 7191).

Das, Aniruddha and Edward O. Laumann (2010). “How to Get Valid Answers from Survey Questions: What We Learned from Asking about Sexual Behavior and the Measurement of Sexuality”. *The SAGE Handbook of Measurement*. Ed. by Geoffrey Walford, Eric Tucker, and Madhu Viswanathan. London: SAGE Publications, 9–26.

de Chaisemartin, Clément and Xavier D’Haultfœuille (2020). “Two-Way Fixed Effects Estimators with Heterogeneous Treatment Effects”. *American Economic Review* 110(9), 2964–2996.

de Quidt, Jonathan, Johannes Haushofer, and Christopher Roth (2018). “Measuring and Bounding Experimenter Demand”. *American Economic Review* 108(11), 3266–3302.

Delavande, Adeline (2014). “Probabilistic Expectations in Developing Countries”. *Annual Review of Economics* 6(1), 1–20.

Delavande, Adeline, Xavier Giné, and David McKenzie (2011). “Eliciting Probabilistic Expectations with Visual Aids in Developing Countries: How Sensitive are Answers to Variations in Elicitation Design?” *Journal of Applied Econometrics* 26(3), 479–497.

Delavande, Adeline and Basit Zafar (2018). “University Choice: The Role of Expected Earnings, Nonpecuniary Outcomes, and Financial Constraints”. *Journal of Political Economy* 127(5), 2343–2393.

– (2019). “University Choice: The Role of Expected Earnings, Nonpecuniary Outcomes, and Financial Constraints”. *Journal of Political Economy* 127(5), 2343–2393.

DellaVigna, Stefano (2009). “Psychology and Economics: Evidence from the Field”. *Journal of Economic Literature* 47(2), 315–372.

DellaVigna, Stefano, John A. List, Ulrike Malmendier, and Gautam Rao (2017). “Voting to Tell Others”. *The Review of Economic Studies* 84(1), 143–181.

Die Welt (2014). “Schülerfrust auf dem Gymnasium”. *Die Welt*.

Dollmann, Jörg (2011). “Verbindliche und unverbindliche Grundschulempfehlungen und soziale Ungleichheiten am ersten Bildungsübergang”. *KZfSS Kölner Zeitschrift für Soziologie und Sozialpsychologie* 63(4), 595–621.

– (2016). “Less Choice, Less Inequality? A Natural Experiment on Social and Ethnic Differences in Educational Decision-Making”. *European Sociological Review* 32(2), 203–215.

Downs, Anthony (1957). “An Economic Theory of Political Action in a Democracy”. *Journal of Political Economy* 65(2), 135–150.

Druckman, James N., Erik Peterson, and Rune Slothuus (2013). “How Elite Partisan Polarization Affects Public Opinion Formation”. *American Political Science Review* 107(1), 57–79.

Dustmann, Christian, Patrick A. Puhani, and Uta Schönberg (2017). “The Long–Term Effects of Early Track Choice”. *The Economic Journal* 127(603), 1348–1380.

Eichenberger, Reiner and Angel Serna (1996). “Random Errors, Dirty Information, and Politics”. *Public Choice* 86(1), 137–156.

Elango, Sneha, Jorge Luis García, James Heckman, and Andrés Hojman (2015). “Early Childhood Education”. *National Bureau of Economic Research Working Paper Series* No. 21766.

Engelhardt, Carina and Andreas Wagener (2014). “Biased Perceptions of Income Inequality and Redistribution”. *Cesifo Working Papers* 4838.

Engzell, Per, Arun Frey, and Mark Verhagen (2020). “Learning Inequality during the COVID-19 Pandemic”. *Mimeo, University of Oxford*.

Ertl, Hubert (2006). “Educational Standards and the Changing Discourse on Education: the Reception and Consequences of the PISA Study in Germany”. *Oxford Review of Education* 32(5), 619–634.

European Commission (2019). *2nd Survey of Schools: ICT in Education: Germany Country Report.* Luxembourg: European Commission.

– (2020). “Educational Inequalities in Europe and Physical School Closures During COVID-19”. *Fairness Policy Brief Series* 04.

Falk, A. and F. Zimmermann (2013). “A Taste for Consistency and Survey Response Behavior”. *CESifo Economic Studies* 59(1), 181–193.

Falk, Armin, Anke Becker, Thomas Dohmen, David Huffman, and Uwe Sunde (2016). “The Preference Survey Module: A Validated Instrument for Measuring Risk, Time, and Social Preferences”. *IZA Discussion Paper Series* No. 9674.

Fernández, Raquel (2007). “Women, Work, and Culture”. *Journal of the European Economic Association* 5(2-3), 305–332.

– (2013). “Cultural Change as Learning: The Evolution of Female Labor Force Participation over a Century”. *American Economic Review* 103(1), 472–500.

Fernández, Raquel and Alessandra Fogli (2006). “Fertility: The Role of Culture and Family Experience”. *Journal of the European Economic Association* 4(2/3), 552–561.

– (2009). “Culture: An Empirical Investigation of Beliefs, Work, and Fertility”. *American Economic Journal: Macroeconomics* 1(1), 146–177.

Fernández, Raquel, Alessandra Fogli, and Claudia Olivetti (2004). “Mothers and Sons: Preference Formation and Female Labor Force Dynamics”. *The Quarterly Journal of Economics* 119(4), 1249–1299.

Fetzer, Thiemo, Lukas Hensel, Johannes Hermle, and Christopher Roth (2020). “Coronavirus Perceptions and Economic Anxiety”. *The Review of Economics and Statistics*, 1–36.

Fichtl, Anita, Timo Hener, and Helmut Rainer (2012). “Betreuungsgeld”. *ifo Schnelldienst* 21(65), 38–44.

Fiorini, Mario and Michael P. Keane (2014). “How the Allocation of Children’s Time Affects Cognitive and Noncognitive Development”. *Journal of Labor Economics* 32(4), 787–836.

Fokken, Silke (2020). “Die Abschaffung des Gymnasiums wäre politischer Selbstmord”. *Der Spiegel* 2020.

Forehand, Mark R., Rohit Deshpandé, and Americus Reed (2002). “Identity Salience and the Influence of Differential Activation of the Social Self-Schema on Advertising Response”. *The Journal of Applied Psychology* 87(6), 1086–1099.

Fortin, Nicole M. (2005). “Gender Role Attitudes and the Labour-market Outcomes of Women across OECD Countries”. *Oxford Review of Economic Policy* 21(3), 416–438.

– (2015). “Gender Role Attitudes and Women’s Labor Market Participation: Opting-Out, AIDS, and the Persistent Appeal of Housewifery”. *Annals of Economics and Statistics* (117/118), 379–401.

Fraillon, Julian, John Ainley, Wolfram Ainley, Tim Friedman, and Daniel Duckworth (2018). *Preparing for Life in a Digital World: IEA International Computer and Information Literacy Study 2018 International Report.* Amsterdam.

Friedman-Sokuler, Naomi and Moshe Justman (2020). “Gender, Culture and STEM: Counter-Intuitive Patterns in Arab Society”. *Economics of Education Review* 74, 101947.

Fujiwara, Thomas (2015). “Voting Technology, Political Responsiveness, and Infant Health: Evidence From Brazil”. *Econometrica* 83(2), 423–464.

Fuster, Andreas, Ricardo Perez-Truglia, Mirko Wiederholt, and Basit Zafar (2020). “Expectations with Endogenous Information Acquisition: An Experimental Investigation”. *The Review of Economics and Statistics*, 1–54.

Gabel, Matthew and Kenneth Scheve (2007). "Estimating the Effect of Elite Communications on Public Opinion Using Instrumental Variables". *American Journal of Political Science* 51(4), 1013–1028.

Gächter, Simon and Elke Renner (2010). "The Effects of (Incentivized) Belief Elicitation in Public Goods Experiments". *Experimental Economics* 13(3), 364–377.

Ganzeboom, Harry, Paul Graaf, and Donald Treiman (1992). "A Standard International Socio-economic Index of Occupational Status". *Social Science Research* 21(1), 1–56.

Gass-Bolm, Torsten (2005). *Das Gymnasium 1945 - 1980: Bildungsreform und gesellschaftlicher Wandel in Westdeutschland: Zugl.: Freiburg (Breisgau), Univ., Diss., 2004 u.d.T.: Gass-Bolm, Torsten: Das Ende der Penne*. Vol. 7. Moderne Zeit. Göttingen: Wallstein.

Geis-Thoene, Wido (2018). "Familien müssen für die gleiche Betreuung in der Kita unterschiedlich viel zahlen - Ein Vergleich der Gebührenordnungen der größten Städte in Deutschland". *IW-Report* (50).

Gentzkow, Matthew and Jesse M. Shapiro (2006). "Media Bias and Reputation". *Journal of Political Economy* 114(2), 280–316.

Giavazzi, Francesco, Fabio Schiantarelli, and Michel Serafinelli (2013). "Attitudes, Policies, and Work". *Journal of the European Economic Association* 11(6), 1256–1289.

Giuliano, Paola (2018). "Gender". *The Oxford Handbook of Women and the Economy*. Ed. by Susan Averett, Laura M. Argys, and Saul D. Hoffman. New York, NY: Oxford University Press, 644–672.

Glaeser, Edward, Giacomo Ponzetto, and Jesse Shapiro (2005). "Strategic Extremism: Why Republicans and Democrats Divide on Religious Values". *The Quarterly Journal of Economics* (120.4), 1283–1330.

Goerges, Luise and Daniele Nosenzo (2020). "Social Norms and the Labor Market". *Handbook of Labor, Human Resources and Population Economics*. Ed. by Klaus F. Zimmermann. Cham: Springer International Publishing, 1–26.

Goldin, Claudia (2014). "A Grand Gender Convergence: Its Last Chapter". *American Economic Review* 104(4), 1091–1119.

Goldin, Claudia, Lawrence F. Katz, and Ilyana Kuziemko (2006). "The Homecoming of American College Women: The Reversal of the College Gender Gap". *Journal of Economic Perspectives* 20(4), 133–156.

Gong, Yifan, Todd Stinebrickner, and Ralph Stinebrickner (forthcoming). "Marriage, Children, and Labor Supply: Beliefs and Outcomes". *Journal of Econometrics*.

Goren, Paul, Christopher M. Federico, and Miki Caul Kittilson (2009). "Source Cues, Partisan Identities, and Political Value Expression". *American Journal of Political Science* 53(4), 805–820.

Green, Donald, Bradley Palmquist, and Eric Schickler (2002). *Partisan Hearts and Minds: Political Parties and the Social Identities of Voters*. Yale University Press.

Grewenig, Elisabeth, Philipp Lergetporer, Katharina Werner, Ludger Woessmann, and Lisa Simon (2018). “Can Online Surveys Represent the Entire Population?” *Cesifo Working Papers* 7222.

Grewenig, Elisabeth, Philipp Lergetporer, Katharina Werner, Ludger Woessmann, and Larissa Zierow (2020). “COVID-19 and EducationalInequality: How School Closures Affect Low- and High-Achieving Students”. *Cesifo Working Papers* 8648.

Grigorieff, Alexis, Christopher Roth, and Diego Ubfal (2020). “Does Information Change Attitudes Toward Immigrants?” *Demography* 57(3), 1117–1143.

Guiso, Luigi, Ferdinando Monte, Paola Sapienza, and Luigi Zingales (2008). “Diversity. Culture, Gender, and Math”. *Science* 320(5880), 1164–1165.

Gunderson, Morley and Philip Oreopolous (2020). “Chapter 3 - Returns to Education in Developed Countries”. *The Economics of Education (Second Edition)*. Ed. by Steve Bradley and Colin Green. Academic Press, 39–51.

Guyon, Nina, Eric Maurin, and Sandra McNally (2012). “The Effect of Tracking Students by Ability into Different Schools: A Natural Experiment”. *Journal of Human Resources* 47(3), 684–721.

Haaland, Ingar and Christopher Roth (2020). “Labor Market Concerns and Support for Immigration”. *Journal of Public Economics* 191, 104256.

Haaland, Ingar, Christopher Roth, and Johannes Wohlfart (2020). “Designing Information Provision Experiments”. *Cesifo Working Papers* 8406.

Hampf, Franziska (2019). “The Effect of CompulsorySchooling on Skills: Evidence from a Reform in Germany”. *ifo Working Papers* (No. 313).

Hanushek, Eric A. (1986). “The Economics of Schooling: Production and Efficiency in Public Schools”. *Journal of Economic Literature* 24(3), 1141–1177.

– (2002). “Publicly Provided Education”. *Handbook of Public Economics*. Ed. by Alan Auerbach and Martin J. Feldstein. Vol. 4. Elsevier, 2045–2141.

– (2020). “Education Production Functions”. *The Economics of Education (Second Edition)*. Ed. by Steve Bradley and Colin Green. Academic Press.

Hanushek, Eric A., Susanne Link, and Ludger Woessmann (2013). “Does School Autonomy Make Sense Everywhere? Panel Estimates from PISA”. *Journal of Development Economics* 104, 212–232.

Hanushek, Eric A., Marc Piopiunik, and Simon Wiederhold (2018). “The Value of Smarter Teachers: International Evidence on Teacher Cognitive Skills and Student Performance”. *Journal of Human Resources*, 0317–8619R1.

Hanushek, Eric A. and Ludger Woessmann (2006). “Does Educational Tracking Affect Performance and Inequality? Differences-in-Differences Evidence across Countries”. *The Economic Journal* 116(510), C63–C76.

– (2008). “The Role of Cognitive Skills in Economic Development”. *Journal of Economic Literature* 46(3), 607–668.

– (2011). “The Economics of International Differences in Educational Achievement”. *Handbook of the Economics of Education*. Ed. by Eric A. Hanushek, Stephen Machin, and Ludger Woessmann. Vol. 3. Elsevier, 89–200.

– (2015). *The Knowledge Capital of Nations: Education and the Economics of Growth*. Cambridge: The MIT Press.

Hanushek, Eric, Lavinia Kinne, Philipp Lergetporer, and Ludger Woessmann (2020). “Culture and Student Achievement: The Intertwined Roles of Patience and Risk-Taking”. *National Bureau of Economic Research Working Paper Series* 27484.

Harmon, Colm, Hessel Oosterbeek, and Ian Walker (2003). “The Returns to Education: Microeconomics”. *Journal of Economic Surveys* 17(2), 115–156.

Hartig, Johannes and Andreas Frey (2012). “Validität des Tests zur Überprüfung des Erreichens der Bildungsstandards in Mathematik: Zusammenhänge mit den bei PISA gemessenen Kompetenzen und Varianz zwischen Schulen und Schulformen”. *Diagnostica* (58), 3–14.

Heckman, James J., Lance J. Lochner, and Petra E. Todd (2006). “Earnings Functions, Rates of Return and Treatment Effects: The Mincer Equation and Beyond”. *Handbook of the Economics of Education*. Ed. by E. Hanushek and F. Welch. Vol. 1. Elsevier, 307–458.

Helbig, Marcel and Rita Nikolai (2015). *Die Unvergleichbaren: Der Wandel der Schulsysteme in den deutschen Bundesländern seit 1949*. Bad Heilbrunn: Verlag Julius Klinkhardt.

Herrlitz, Hans-Georg, Wulf Hopf, Hartmut Titze, and Ernst Cloer (2009). *Deutsche Schulgeschichte von 1800 bis zur Gegenwart: Eine Einführung*. 5., aktualisierte Aufl. Weinheim: Juventa.

Hill, John P. and Mary Ellen Lynch (1983). “The Intensification of Gender-Related Role Expectations during Early Adolescence”. *Girls at Puberty: Biological and Psychosocial Perspectives*. Ed. by Jeanne Brooks-Gunn and Anne C. Petersen. Boston, MA: Springer US, 201–228.

Horton, John J., David G. Rand, and Richard J. Zeckhauser (2011). “The Online Laboratory: Conducting Experiments in a Real Labor Market”. *Experimental Economics* 14(3), 399–425.

Hoxby, Caroline M. (2000). “The Effects of Class Size on Student Achievement: New Evidence from Population Variation”. *The Quarterly Journal of Economics* 115(4), 1239–1285.

Hoxby, Caroline and Sarah Turner (2013). “Expanding College Opportunities for High-Achieving, Low-Income Students”. *SIEPR Working Paper* 12-014.

Huber, Stephan Gerhard and Christoph Helm (2020). “COVID-19 and Schooling: Evaluation, Assessment and Accountability in Times of Crises—Reacting Quickly to Explore Key Issues for Policy practice and research with the school barometer”. *Educational Assessment, Evaluation and Accountability* 32(2), 237–270.

Hurley, Terrance M. and Jason F. Shogren (2005). “An Experimental Comparison of Induced and Elicited Beliefs”. *Journal of Risk and Uncertainty* 30(2), 169–188.

Jackson, C. Kirabo (2010). “A Little Now for a Lot Later: A Look at a Texas Advanced Placement Incentive Program”. *Journal of Human Resources* 45(3), 591–639.

Jähnen, Stefanie and Marcel Helbig (2015). “Der Einfluss schulrechtlicher Reformen auf Bildungsungleichheiten zwischen den deutschen Bundesländern”. *KZfSS Kölner Zeitschrift für Soziologie und Sozialpsychologie* 67(3), 539–571.

Jaume, David and Alexander Willén (2019). “The Long-Run Effects of Teacher Strikes: Evidence from Argentina”. *Journal of Labor Economics* 37(4), 1097–1139.

Jayachandran, Seema (2019). “Social Norms as a Barrier to Women’s Employment in Developing Countries”. *Northwestern Working Paper*.

Jessen, Jonas, Sophia Schmitz, and Sevrin Waights (2020). “Understanding Day Care Enrolment Gaps”. *Journal of Public Economics* 190, 104252.

Karadja, Mounir, Johanna Mollerstrom, and David Seim (2016). “Richer (and Holier) Than Thou? The Effect of Relative Income Improvements on Demand for Redistribution”. *Review of Economics and Statistics* 99.

Karagözoğlu, Emin and Martin G. Kocher (2019). “Bargaining under Time Pressure from Deadlines”. *Experimental Economics* 22(2), 419–440.

Kaufmann, Katja Maria (2014). “Understanding the Income Gradient in College Attendance in Mexico: The Role of Heterogeneity in Expected Returns”. *Quantitative Economics* 5(3), 583–630.

Kemp, Simon (2003). *Public Goods and Private Wants: A Psychological Approach to Government Spending*. Cheltenham: Elgar.

Kemptner, Daniel, Hendrik Jürges, and Steffen Reinhold (2011). “Changes in Compulsory Schooling and the Causal Effect of Education on Health: Evidence from Germany”. *Journal of Health Economics* 30(2), 340–354.

Keusch, Florian and Chan Zhang (2015). “A Review of Issues in Gamified Surveys”. *Social Science Computer Review* 35(2), 147–166.

Kleven, Henrik and Camille Landais (2017). “Gender Inequality and Economic Development: Fertility, Education and Norms”. *Economica* 84(334), 180–209.

Kleven, Henrik, Camille Landais, Johanna Posch, Andreas Steinhauer, and Josef Zweimüller (2019a). “Child Penalties across Countries: Evidence and Explanations”. *AEA Papers and Proceedings* 109.

Kleven, Henrik, Camille Landais, and Jakob Egholt Søgaard (2019b). “Children and Gender Inequality: Evidence from Denmark”. *American Economic Journal: Applied Economics* 11(4), 181–209.

– (2020). “Does Biology Drive Child Penalties? Evidence from Biological and Adoptive Families”. *National Bureau of Economic Research Working Paper Series* No. 27130.

Kling, Jeffrey R., Jeffrey B. Liebman, and Lawrence F. Katz (2007). “Experimental Analysis of Neighborhood Effects”. *Econometrica* 75(1), 83–119.

Kocher, Martin G., David Schindler, Stefan T. Trautmann, and Yilong Xu (2019). “Risk, Time Pressure, and Selection Effects”. *Experimental Economics* 22(1), 216–246.

Köller, Olaf, Michel Knigge, and Bernd Tesch (2011). *IQB-Ländervergleich Sprachen 2008/2009 (IQB-LV 2008-9) [IQB National Assessment Study 2008/2009 (IQB-LV 2008-9)] (Version 2) [Data set]*. Berlin: Berlin: IQB – Institut zur Qualitätsentwicklung im Bildungswesen. http://doi.org/10.5159/IQB_LV_2008_v2.

Krueger, Alan B. (1999). "Experimental Estimates of Education Production Functions". *The Quarterly Journal of Economics* 114(2), 497–532.

Kultusministerkonferenz (2006). *Gesamtstrategie der Kultusministerkonferenz zum Bildungsmonitoring*. München: Wolters Kluwer.

– (2015). *Übergang von der Grundschule in Schulen des Sekundarbereichs I und Förderung, Beobachtung und Orientierung in den Jahrgangsstufen 5 und 6 (sog. Orientierungsstufe)*. Berlin: Sekretariat der Ständigen Konferenz der Kultusminister der Länderin der Bundesrepublik Deutschland.

Kunter, Mareike, Gundel Schümer, Cordula Artelt, Jürgen Baumert, Eckhard Klieme, Michael Neubrand, Manfred Prenzel, Ulrich Schiefele, Wolfgang Schneider, Petra Stanat, Klaus-Jürgen Tillmann, and Manfred Weiß (2002). *PISA 2000: Dokumentation der Erhebungsinstrumente (Materialien aus der Bildungsforschung Bd. [vol.] 72)*. Berlin: Max-Planck-Inst. für Bildungsforschung.

Kunz, Johannes S. and Kevin E. Staub (2020). "Early subjective completion beliefs and the demand for post-secondary education". *Journal of Economic Behavior & Organization* 177, 34–55.

Kuziemko, Ilyana, Michael I. Norton, Emmanuel Saez, and Stefanie Stantcheva (2015). "How Elastic Are Preferences for Redistribution? Evidence from Randomized Survey Experiments". *American Economic Review* 105(4), 1478–1508.

Kuziemko, Ilyana, Jessica Pan, Jenny Shen, and Ebonya Washington (2018). "The Mommy Effect: Do Women Anticipate the Employment Effects of Motherhood?" *National Bureau of Economic Research Working Paper Series* No. 24740.

Lau, Richard R. and David P. Redlawsk (2001). "Advantages and Disadvantages of Cognitive Heuristics in Political Decision Making". *American Journal of Political Science* 45(4), 951–971.

Lavy, Victor (2015). "Do Differences in Schools' Instruction Time Explain International Achievement Gaps? Evidence from Developed and Developing Countries". *The Economic Journal* 125(588), F397–F424.

Lee, David S., Enrico Moretti, and Matthew J. Butler (2004). "Do Voters Affect or Elect Policies? Evidence from the U. S. House". *The Quarterly Journal of Economics* 119(3), 807–859.

Lenski, Anna E., Martin Hecht, Christiane Penk, Felix Milles, Manuel Mezger, Patricia Heitmann, Petra Stanat, and Hans A. Pant (2016). *IQB-Ländervergleich 2012. Skalenhandbuch zur Dokumentation der Erhebungsinstrumente*. Berlin: Humboldt-Universität zu Berlin, Institut zur Qualitätsentwicklung im Bildungswesen. 10.20386/HUB-42547.

Lenz, Gabriel S. (2009). "Learning and Opinion Change, Not Priming: Reconsidering the Priming Hypothesis". *American Journal of Political Science* 53(4), 821–837.

Lenz, Gabriel S. (2012). *Follow the Leader? How Voters Respond to Politicians' Policies and Performance*. Chicago Studies in American Politics. Chicago: University of Chicago Press.

Lergetporer, Philipp, Guido Schwerdt, Katharina Werner, Martin R. West, and Ludger Woessmann (2018a). "How Information Affects Support for Education Spending: Evidence from Survey Experiments in Germany and the United States". *Journal of Public Economics* 167, 138–157.

Lergetporer, Philipp, Katharina Werner, and Ludger Woessmann (2018b). "Does Ignorance of Economic Returns and Costs Explain the Educational Aspiration Gap? Evidence from Representative Survey Experiments". *Cesifo Working Papers* 7000.

– (2020). "Educational Inequality and Public Policy Preferences: Evidence from Representative Survey Experiments". *Journal of Public Economics* 188, 104226.

Leuven, Edwin, Hessel Oosterbeek, and Bas van der Klaauw (2010). "The Effect of Financial Rewards on Students' Achivement: evidence from a Randomized Experiment". *Journal of the European Economic Association* 8(6), 1243–1265.

Levendusky, Matthew S. (2010). "Clearer Cues, More Consistent Voters: A Benefit of Elite Polarization". *Political Behavior* 32(1), 111–131.

Levitt, Steven D. (1996). "How Do Senators Vote? Disentangling the Role of Voter Preferences, Party Affiliation, and Senator Ideology". *The American Economic Review* 86(3), 425–441.

Lichtman-Sadot, Shirlee (2016). "Improving Academic Performance through Conditional Benefits: Open/Closed Campus Policies in High School and Student Outcomes". *Economics of Education Review* 54, 95–112.

Lindo, Jason M., Nicholas J. Sanders, and Philip Oreopoulos (2010). "Ability, Gender, and Performance Standards: Evidence from Academic Probation". *American Economic Journal: Applied Economics* 2(2), 95–117.

List, John A., Azeem M. Shaikh, and Yang Xu (2019). "Multiple Hypothesis Testing in Experimental Economics". *Experimental Economics* 22(4), 773–793.

Lupia, Arthur (1994). "Shortcuts Versus Encyclopedias: Information and Voting Behavior in California Insurance Reform Elections". *American Political Science Review* 88(1), 63–76.

Malamud, Ofer and Cristian Pop-Eleches (2011). "School Tracking and Access to Higher Education among Disadvantaged Groups". *Journal of Public Economics* 95(11-12), 1538–1549.

Maldonado, Joana and Kristof De Witte (2020). "The Effect of School cCosures on Standardisedstudent Test Outcomes". *KU Leuven Discussion Paper Series* DPS20.17.

Maniadis, Zacharias, Fabio Tufano, and John A. List (2014). "One Swallow Doesn't Make a Summer: New Evidence on Anchoring Effects". *American Economic Review* 104(1), 277–290.

Mankiw, N. Gregory, Ricardo Reis, and Justin Wolfers (2003). "Disagreement about Inflation Expectations". *NBER Macroeconomics Annual* 18, 209–248.

Manski, Charles F. (2004). "Measuring Expectations". *Econometrica* 72(5), 1329–1376.

– (2018). "Survey Measurement of Probabilistic Macroeconomic Expectations: Progress and Promise". *NBER Macroeconomics Annual* 32, 411–471.

Marcus, Jan and Vaishali Zambre (2019). “The Effect of Increasing Education Efficiency on University Enrollment”. *Journal of Human Resources* 54(2), 468–502.

Matthewes, Sönke Hendrik (2020). “Better Together? Heterogeneous Effects of Tracking on Student Achievement”. *The Economic Journal*.

Meghir, Costas and Mårten Palme (2005). “Educational Reform, Ability, and Family Background”. *The American Economic Review* 95(1), 414–424.

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

Mummolo, Jonathan and Eric Peterson (2019). “Demand Effects in Survey Experiments: An Empirical Assessment”. *American Political Science Review* 113(2), 517–529.

Neugebauer, Martin (2010). “Bildungsungleichheit und Grundschulempfehlung beim Übergang auf das Gymnasium: Eine Dekomposition primärer und sekundärer Herkunftseffekte”. *Zeitschrift für Soziologie* 39(3), 202–214.

Nollenberger, Natalia, Núria Rodríguez-Planas, and Almudena Sevilla (2016). “The Math Gender Gap: The Role of Culture”. *American Economic Review* 106(5), 257–261.

Obergruber, Natalie and Larissa Zierow (2020). “Students’ Behavioural Responses to a Fallback Option - Evidence from Introducing Interim Degrees in German Schools”. *Economics of Education Review* 75, 101956.

OECD (2013). *PISA 2012 Results: What Makes Schools Successful (Volume IV)*. Paris: OECD.

– (2020). *Education GPS — Germany — Student Performance (PISA 2018)*. Paris: OECD.

Olivetti, Claudia and Barbara Petrongolo (2016). “The Evolution of Gender Gaps in Industrialized Countries”. *Annual Review of Economics* 8(1), 405–434.

Oreopoulos, Philip and Kjell G. Salvanes (2011). “Priceless: The Nonpecuniary Benefits of Schooling”. *Journal of Economic Perspectives* 25(1), 159–184.

Ortoleva, Pietro and Erik Snowberg (2015). “Overconfidence in Political Behavior”. *American Economic Review* 105(2), 504–535.

Osikominu, Aderonke, Gregor Pfeifer, and Kristina Strohmaier (2021). “The Effects of Free Secondary School Track Choice: A Disaggregated Synthetic Control Approach”. *IZA Discussion Paper Series* 14033.

Otto, Otto and Arnfrid Schenk (2011). “Am Ende entscheiden die Eltern”. *Die Zeit* 2011(51/2011).

Pallais, Amanda (2009). “Taking a Chance on College: Is the Tennessee Education Lottery Scholarship Program a Winner?” *Journal of Human Resources* 44(1), 199–222.

Pannico, Roberto (2017). “Is the European Union too Complicated? Citizens’ Lack of Information and Party Cue Effectiveness”. *European Union Politics* 18(3), 424–446.

Pant, Hans Anand, Petra Stanat, Martin Hecht, Patricia Heitmann, Malte Jansen, Anna Eva Lenski, Chrisitiane Penk, Claudia Pöhlmann, Alexander Roppelt, Ulrich Schroeders, and Thilo Siegle (2015). *IQB-Ländervergleich Mathematik und Naturwissenschaften 2012 (IQB-LV 2012) [IQB National Assessment Study2012 (IQB-LV 2012] (Version 4) [Data set]*. Berlin: IQB – Institut zur Qualitätsentwicklung im Bildungswesen. http://doi.org/10.5159/IQB_LV_2012_v4.

Pekkala Kerr, Sari, Tuomas Pekkarinen, and Roope Uusitalo (2013). “School Tracking and Development of Cognitive Skills”. *Journal of Labor Economics* 31(3), 577–602.

Persson, Torsten and Guido Tabellini (2013). “Political Economics and Public Finance”. *Handbook of Public Economics*. Ed. by Alan J. Auerbach. Vol. 3. Handbooks in economics. Amsterdam: Elsevier/North Holland, 1549–1659.

Pietsch, Marcus, Kathrin Böhme, Alexander Robitzsch, and Tobias Stubbe (2009). *Das Stufenmodell zur Lesekompetenz der länderübergreifenden Bildungsstandards im Vergleich zu IGLU 2006: In: Bildungsstandards Deutsch und Mathematik*. Weinheim: Beltz-Pädagogik.

Piopiunik, Marc (2014). “The Effects of Early Tracking on Student Performance: Evidence from a School Reform in Bavaria”. *Economics of Education Review* 42, 12–33.

Pischke, Jörn–Steffen (2007). “The Impact of Length of the School Year on Student Performance and Earnings: Evidence from the German Short School Years”. *The Economic Journal* 117(523), 1216–1242.

Pischke, Jörn-Steffen and Till von Wachter (2008). “Zero Returns to Compulsory Schooling in Germany: Evidence and Interpretation”. *The Review of Economics and Statistics* 90(3), 592–598.

Prenzel, Manfred, Cordula Artelt, Jürgen Baumert, Werner Blum, Marcus Hammann, Eckhard Klieme, and Reinhard Pekrun (2010). *Programme for International Student Assessment 2006 (PISA 2006) (Version 1) [Data set]*. Berlin: Berlin: IQB – Institut zur Qualitätsentwicklung im Bildungswesen. http://doi.org/10.5159/IQB_PISA_2006_v1.

Prenzel, Manfred, Jürgen Baumert, Werner Blum, Rainer Lehmann, Detlev Leutner, Michael Neubrand, Reinhard Pekrun, Hans-Günter Rolff, Jürgen Rost, and Ulrich Schiefele (2007). *Programme for International Student Assessment 2003 (PISA 2003) (Version 1) [Data set]*. Berlin: IQB – Institut zur Qualitätsentwicklung im Bildungswesen. http://doi.org/10.5159/IQB_PISA_2003_v1.

Prenzel, Manfred, Christine Sälzer, Eckhard Klieme, Olaf Köller, Julia Mang, Jörg-Henrik Heine, Anja Schiepe-Tiska, and Katharina Müller (2015). *Programme for International Student Assessment 2012 (PISA 2012) (Version 5) [Data set]*. Berlin: IQB – Institut zur Qualitätsentwicklung im Bildungswesen. http://doi.org/10.5159/IQB_PISA_2012_v5.

Priess, Heather A., Sara M. Lindberg, and Janet Shibley Hyde (2009). “Adolescent Gender-Role Identity and Mental Health: Gender Intensification Revisited”. *Child Development* 80(5), 1531–1544.

Prior, Markus, Gaurav Sood, and Kabir Khanna (2015). “You Cannot Be Serious: The Impact of Accuracy Incentives on Partisan Bias in Reports of Economic Perceptions”. *Quarterly Journal of Political Science* 10(4), 489–518.

Psacharopoulos, George and Harry Anthony Patrinos (2004). “Returns to Investment in Education: A Further Update”. *Education Economics* 12(2), 111–134.

Puleston, Jon (2011). “Improving Online Surveys”. *Improving Online Surveys* 53, 557–560.

Rahn, Wendy M. (1993). “The Role of Partisan Stereotypes in Information Processing about Political Candidates”. *American Journal of Political Science* 37(2), 472–496.

Reuben, Ernesto, Matthew Wiswall, and Basit Zafar (2017). “Preferences and Biases in Educational Choices and Labour Market Expectations: Shrinking the Black Box of Gender”. *The Economic Journal* 127(604), 2153–2186.

Richter, Dirk, Katrin Böhme, Jana Bastian-Wurzel, Hans A. Pant, and Petra Stanat (2014). *IQB Ländervergleich 2011. Skalenhandbuch zur Dokumentation der Erhebungsinstrumente*. Berlin: Humboldt-Universität zu Berlin, Institut zur Qualitätsentwicklung im Bildungswesen. 10.18452/3127.

Rivkin, Steven, Eric Hanushek, and John Kain (2005). “Teachers, Schools, and Academic Achievement”. *Econometrica* 73(2), 417–458.

Robert Bosch Stiftung (2020). *Das Deutsche Schulbarometer Spezial: Lehrerbefragung zur Corona-Krise.* Stuttgart: Robert Bosch Stiftung.

Rockoff, Jonah E., Douglas O. Staiger, Thomas J. Kane, and Eric S. Taylor (2012). “Information and Employee Evaluation: Evidence from a Randomized Intervention in Public Schools”. *American Economic Review* 102(7), 3184–3213.

Roodman, David, Morten Ørregaard Nielsen, James G. MacKinnon, and Matthew D. Webb (2019). “Fast and wild: Bootstrap inference in Stata using boottest”. *The Stata Journal* 19(1), 4–60.

Roth, Christopher, Sonja Settele, and Johannes Wohlfart (2020). “Beliefs about Public Debt and the Demand for Government Spending”. *Cesifo Working Papers* 8087.

Roth, Christopher and Johannes Wohlfart (2019). “How Do Expectations about the Macroeconomy Affect Personal Expectations and Behavior?” *The Review of Economics and Statistics* 102(4), 731–748.

Roth, Tobias and Manuel Siegert (2015). “Freiheit versus Gleichheit: Der Einfluss der Verbindlichkeit der Übergangsempfehlung auf die soziale Ungleichheit in der Sekundarstufe”. *Zeitschrift für Soziologie* 44(2), 118–136.

– (2016). “Does the Selectivity of an Educational System Affect Social Inequality in Educational Attainment? Empirical Findings for the Transition from Primary to Secondary Level in Germany”. *European Sociological Review* 32(6), 779–791.

Ryan, Timothy J. (2017). “How Do Indifferent Voters Decide? The Political Importance of Implicit Attitudes”. *American Journal of Political Science* 61(4), 892–907.

Sachse, Karoline A., Julia Kretschmann, Aleksander Kocaj, Olaf Köller, Michel Knigge, and Bernd Tesch (2012). *IQB-Ländervergleich 2008/2009. Skalenhandbuch zur Dokumentation der Erhebungsinstrumente*. Berlin: Humboldt-Universität zu Berlin, Institut zur Qualitätsentwicklung im Bildungswesen. 10.20386/HUB-42659.

Samuels, David and Cesar Zucco (2014). “The Power of Partisanship in Brazil: Evidence from Survey Experiments”. *American Journal of Political Science* 58(1), 212–225.

Schenk, Arnfrid (2010). “Elternwille: Fehlanzeige”. *Die Zeit* 2010(31/2010).

Schipolowski, Stefan, Johanna Busse, Camilla Rjosk, Nicole Mahler, Benjamin Becker, and Petra Stanat (2019). *IQB-Bildungstrend 2016. Skalenhandbuch zur Dokumentation der Erhebungsinstrumente in den Fächern Deutsch und Mathematik*. Berlin: Humboldt-Universität zu Berlin, Institut zur Qualitätsentwicklung im Bildungswesen.

Schipolowski, Stefan, Nicole Haag, Felix Milles, Stefanie Pietz, and Petra Stanat (2018a). *IQB-Bildungstrend 2015. Skalenhandbuch zur Dokumentation der Erhebungsinstrumente im Fach Französisch*. Berlin: Humboldt-Universität zu Berlin, Institut zur Qualitätsentwicklung im Bildungswesen. 10.18452/19998.

– (2018b). *IQB-Bildungstrend 2015. Skalenhandbuch zur Dokumentation der Erhebungsinstrumente in den Fächern Deutsch und Englisch*. Berlin: Humboldt-Universität zu Berlin, Institut zur Qualitätsentwicklung im Bildungswesen. 10.18452/19997.

Schlag, Karl H., James Tremewan, and van der Weele, Joël J. (2015). “A Penny for Your Thoughts: A Survey of Methods for Eliciting Beliefs”. *Experimental Economics* 18(3), 457–490.

Schotter, Andrew and Isabel Trevino (2014). “Belief Elicitation in the Laboratory”. *Annual Review of Economics* 6(1), 103–128.

Schröder, Mathis, Rainer Siegers, and Katharina Spieß (2013). “Familien in Deutschland - FiD”. *Schmollers Jahrbuch* (133), 595–606.

Schueler, Beth E. and Martin R. West (2016). “Sticker Shock: How Information Affects Citizen Support for Public School Funding”. *Public Opinion Quarterly* 80(1), 90–113.

Schuler-Harms, Margarete (2010). *„Verfassungsrechtlich prekär”: Expertise zur Einführung eines Betreuungsgeldes*. Berlin: Friedrich-Ebert Stiftung.

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

Schwarz, Norbert and Leigh Ann Vaughn (2002). “The Availability Heuristic Revisited: Ease of Recall and Content of Recall as Distinct Sources of Information”. *Heuristics and Biases: The Psychology of Intuitive Judgment*. Ed. by Thomas Gilovich, Dale Griffin, and Tversky Kahnemann. New York: Cambridge University Press, 103–119.

Slothuus, Rune (2010). “When Can Political Parties Lead Public Opinion? Evidence from a Natural Experiment”. *Political Communication* 27(2), 158–177.

– (2016). “Assessing the Influence of Political Parties on Public Opinion: The Challenge from Pretreatment Effects”. *Political Communication* 33(2), 302–327.

Stanat, Petra, Katrin Böhme, Stefan Schipolowski, Nicole Haag, Sebastian Weirich, Karoline A. Sachse, Lars Hoffmann, and Felicitas Federlein (2018). *IQB-Bildungstrend Sprachen 2015 (IQB-BT 2015) [IQB Trends in Student Achievement 2015 (IQB-BT 2015)] (Version 5) [Data set]*. Berlin: IQB – Institut zur Qualitätsentwicklung im Bildungswesen. http://doi.org/10.5159/IQB_BT_2015_v5.

Stanat, Petra, Hans Anand Pant, Katrin Böhme, Dirk Richter, Sebastian Weirich, Nicole Haag, Alexander Roppelt, Maria Engelbert, and Heino Reimers (2014). *IQB-Ländervergleich Primarstufe 2011 (IQB-LV 2011) [IQB National Assessment Study 2011 (IQB-LV 2011)] (Version 3)*

[Data set]. Berlin: IQB – Institut zur Qualitätsentwicklung im Bildungswesen. http://doi.org/10.5159/IQB_LV_2011_v3.

Stanat, Petra, Stefan Schipolowski, Camilla Rjosk, Sebastian Weirich, Nicole Mahler, Pauline Kohrt, and Julia Wittig (2019). *IQB-Bildungstrend Primarstufe 2016 (IQB-BT 2016) [IQB Trends in Student Achievement 2016 (IQB-BT 2016)] (Version 1) [Data set]*. Berlin: IQB – Institut zur Qualitätsentwicklung im Bildungswesen. http://doi.org/10.5159/IQB_BT_2016_v1.

Statistisches Bundesamt (1991-2016). *Allgemeinbildende Schulen. Fachserie 11 Reihe 1.* Wiesbaden: Statistisches Bundesamt.

– (2016). *Bildungsausgaben: Ausgaben je Schülerin und Schüler 2013*. Wiesbaden: Statistisches Bundesamt.

– (2018). *Bildungsausgaben: Ausgaben je Schülerin und Schüler 2015*. Wiesbaden: Statistisches Bundesamt.

Steinhauer, Andreas (2018). "Working Moms, Childlessness, and Female Identity". *LIEPP Working Paper* No. 79.

Stinebrickner, Ralph and Todd R. Stinebrickner (2013). "A Major in Science? Initial Beliefs and Final Outcomes for College Major and Dropout". *The Review of Economic Studies* 81(1), 426–472.

Strömberg, David (2004). "Radio's Impact on Public Spending". *The Quarterly Journal of Economics* 119(1), 189–221.

The Economist (July 20, 2019). "Changing the Guard: Saudi Arabia Weighs Loosening more Controls on Women". *The Economist* 2019.

Timmermans, A. C., H. de Boer, H. T. A. Amsing, and M. P. C. van der Werf (2018). "Track Recommendation Bias: Gender, Migration Background and SES Bias over a 20-year Period in the Dutch Context". *British Educational Research Journal* 44(5), 847–874.

Trautmann, Stefan T. and Gijs van de Kuilen (2015). "Belief Elicitation: A Horse Race among Truth Serums". *The Economic Journal* 125(589), 2116–2135.

Tversky, Amos and Daniel Kahneman (1974). "Judgment under Uncertainty: Heuristics and Biases". *Science* 185(4157), 1124.

UNESCO (2020a). *Adverse Consequences of School Closures*.

– (2020b). *Education: From Disruption to Discovery*.

Wang, Stephanie W. (2011). "Incentive Effects: The case of Belief Elicitation from Individuals in Groups". *Economics Letters* 111(1), 30–33.

Wippermann, Carsten (2015). "Transparenz für mehr Entgeltgleichheit: Einflüsse auf den Gender Pay Gap (Berufswahl, Arbeitsmarkt, Partnerschaft, Rollenstereotype) und Perspektiven der Bevölkerung für Lohngerechtigkeit zwischen Frauen und Männern". *Bundesministeriums für Familie, Senioren, Frauen und Jugend*.

Wiswall, Matthew and Basit Zafar (2015). "Determinants of College Major Choice: Identification using an Information Experiment". *The Review of Economic Studies* 82(2), 791–824.

Wiswall, Matthew and Basit Zafar (2018). “Preference for the Workplace, Investment in Human Capital, and Gender”. *The Quarterly Journal of Economics* 133(1), 457–507.

– (2020). “Human Capital Investments and Expectations about Career and Family”. *Journal of Political Economy*.

Woessmann, Ludger (2016). “The Importance of School Systems: Evidence from International Differences in Student Achievement”. *Journal of Economic Perspectives* 30(3), 3–32.

– (2020). “Folgekosten ausbleibenden Lernens: Was wir über die Corona-bedingten Schulschließungen aus der Forschung lernen können”. *ifo Schnelldienst*.

Woessmann, Ludger, Vera Freundl, Elisabeth Grewenig, Philipp Lergetporer, Katharina Werner, and Larissa Zierow (2020). “Bildung in der Corona-Krise: Wie haben die Schulkinder die Zeit der Schulschließungen verbracht, und welche Bildungsmaßnahmen befürworten die Deutschen?” *ifo Schnelldienst* 73(9), 25–39.

Woessmann, Ludger, Philipp Lergetporer, Elisabeth Grewenig, Sarah Kersten, and Katharina Werner (2018). “Denken Jugendliche anders über Bildungspolitik als Erwachsene?” *fo Schnelldienst 71* (17), 31–45.

Zafar, Basit (2011). “How Do College Students Form Expectations?” *Journal of Labor Economics* 29(2), 301–348.

– (2013). “College Major Choice and the Gender Gap”. *Journal of Human Resources* 48(3), 545–595.

Zaller, John (2004). “Floating Voters in U.S. Presidential Elections, 1948–2000”. *Studies in Public Opinion*. Ed. by Willem E. Saris and Paul M. Sniderman. Princeton: Princeton University Press, 166–212.

Zimmermann, Florian (2020). “The Dynamics of Motivated Beliefs”. *American Economic Review* 110(2), 337–361.

Zizzo, Daniel John (2010). “Experimenter Demand Effects in Economic Experiments”. *Experimental Economics* 13(1), 75–98.

Curriculum Vitae

Elisabeth Grewenig

08/2016 – 08/2021	Junior Economist and Doctoral Student ifo Institute – Leibniz Institute for Economic Research at the University of Munich
08/2016 – 06/2021	Doctoral Candidate in Economics (Dr. oec. publ.) Munich Graduate School of Economics Ludwig-Maximilians-Universität München
09/2018 – 04/2019	Visiting Research Fellow Program on Education Policy and Government Harvard University
09/2014 – 06/2016	Master of Science in Economics Stockholm University
09/2012 – 01/2013	Undergraduate Studies in Economics University of Nottingham
09/2010 – 06/2014	Bachelor of Science in Economics University of Mannheim