Essays on the Provision of Effective Public Services

by

Michael Ricks

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Economics)
in The University of Michigan
2022

Doctoral Committee:

Professor James R. Hines, Chair
Assistant Professor Sara B. Heller
Professor Brian A. Jacob
Professor Amanda E. Kowalski

Michael Ricks
ricksmi@umich.edu
ORCID iD 0000-0003-1220-0862

“Already Dedicated”

ACKNOWLEDGEMENTS

It seems almost unbelievable that this dissertation is finally complete. If there is a shortage of humility and gratitude in academia, perhaps recognizing those who have changed my trajectory comprises a final, valuable lesson to me as a student.

This journey would not have been possible without my four committee members: Jim Hines, Brian Jacob, Amanda Kowalski, and Sara Heller. I am grateful for their patience and support as I have blundered through four fields, changed dissertation timelines, juggled too many projects, missed too many typos, and created a patchwork dissertation perhaps more aptly called "Three Essays in Early Childhood Education, Energy Policy, and Political Economy: No They Are Not Related." From each committee member, I've learned lessons that shape my perspective on research and mentoring. I hope to emulate Jim's curiosity (reflected in his threats "if you don't write this paper, I might!") and optimism. Sometimes a student just needs to be reminded that someone they admire thinks that he or she is a "smart cookie" or that everything will work out. I'm so grateful for four years of data analysis training as a research assistant to Brian; learning to think before running regressions is invaluable, but I'm even more inspired by his personal concern as a mentor, reflected in little things like asking after Mindee and the kids. Amanda taught me the expectations of the research community and has pushed me to focus on the things that matter most; as a mentor, she is the one I credit with my transformation from consumer to producer of knowledge. Finally, I hope I can forever internalize Sara's dedication to teaching clearly (whether in class sessions, seminars, or journal articles), her focus on research that improves people's lives, and compassionate mentorship that I want to continuously pay forward.

If it normally takes a village, raising me as an economist seems to have been the effort of a town or a small city. Many have shared their time and attention as generously as if they had been my main advisors. Thank you to John Bound, Ash Craig, Ying Fan, Catie Hausman, and Tanya Rosenblat; I couldn't get nine committee members into the Rackham system, but your support has been transformational. Thank you, Chris Williams, for letting me teach your PhD students and for your help and support as a letter writer on the job market and a mentor to my teaching.

When I worked for Joe Price as a research assistant, he would often say "the best papers are built out of seventeen ideas that each come from different places and people; your job is just to hear those ideas and bring them together into a cohesive whole." Those whose timely feedback provided these gems of insight made this dissertation possible. These people include Jason Baron, Daphna Bassok, Julian Betts, John Bonney, Charlie Brown, Zach Brown, Alecia Cassidy, Matias Cattaneo, Celeste Carruthers, Neil Christy, Jeff Denning, Tatyana Deryugina, Florian Gunsilius, Brigham Frandsen, Gonzalo Vazquez-Bare, Peter Hull, Sarah Johnston, Yeliz Kacamak, Matt Kotchen, Lars Lefgren, Brian McCall, Sarah Mills, Michael Moore, Mike Mueller-Smith, Rich Patterson, Ana Reynoso, Danny Schaffa, Joel Slemrod, Jeff Smith, Kevin Stange, Mel Stephens, Sam Stolper, Rich Sweeney, Brenden Timpe, Chirs Walters, Christina Weiland, Justin Wolfers, Basit Zafar, and seminar participants at University of Michigan, Brigham Young University, the National Tax Association, AERE, the Michigan Tax Invitational, and OSWEET.

While I've been blessed by fantastic teachers, none of this would have been possible without similarly stellar administrative experts. The team at the Education Policy Initiative (EPI) made it possible for me to work with the education data and, together with the Youth Policy Lab, provided computing resources and administrative support for my research and presentations. I want to specifically thank Jasmina Camo-Biogradlija, Kyle Kwaiser, and Nicole Wagner Lam at EPI for their support and Richard Lower and others at the Michigan Department of Education for their interest and feedback. Relatedly, staff at the economics department have saved the day again and again. I also don't think I would have made it to this point without Hiba Baghdadi, Laura Flak, and Lauren Pulay for their tireless work behind the scenes and patience with my many (many, **many**) questions. It is in no small measure because of Julie Heintz that I will actually be employed upon graduation, and because of Julie Esch that I have been able to complete research despite the steep learning curve of university bureaucracy.

I am grateful for the Michigan Department of Education and the Center for Educational Performance and Information which provided access to the administrative education records used in Chapter One. Since great statistical power comes with great responsibility, I want to acknowledge that this research result used data structured and maintained by the MERI-Michigan Education Data Center (MEDC). MEDC data is modified for analysis purposes using rules governed by MEDC and are not identical to those data collected and maintained by the Michigan Department of Education (MDE) and/or Michigan's Center for Educational Performance and Information (CEPI). Results, information and opinions solely represent the analysis, information and opinions of the author and are not endorsed by, or reflect the views or positions of, grantors, MDE and CEPI or any employee thereof.

Chapter Three of my dissertation was made possible by data shared from The State Higher Education Executive Officers Association by Sophia Laderman and was funded in part by grants from the University of Michigan Department of Economics, Department of Political Science, and Rackham Graduate School; Brigham Young University; and the Institute for Education Sciences grant R305B150012.

Finishing a dissertation requires surviving the cognitive and emotional whiplash induced by graduate education. To that end I'm grateful for the many friends we have made in Ann Arbor. I think the best thing that happened to me for my grad school career was coming to EPI to share an office with four generations of IES fellows: Jordy Berne, Elizabeth Burland, Max Gross, Shawn Martin, Shwetha Raghuraman, Anna Shapiro, Andrew Simon, Nate Sotherland, and Brittany Vasquez. Thank you for answering my questions; explaining the nuances of econ classes, job markets, publishing, and academia; and dealing with insuppressible bouts of chattiness. The same can be said for officemates and lunch chats with Stacey Brockman and Jeremy Guardiola at the Youth Policy Lab. Paul Organ and Jamie Fogel helped me survive the job market and listened to more renditions of my job market paper than any human should have been subjected to. Ellen Stuart and Andrew Simon were my key leading indicators for both the market and the dissertation process—thank you! So many others have given me academic and social support when I needed it. Thank you and thanks again to my fellow economists-in-training: Elird Haxhiu, Thomas Helgerman, Caitlin Hegarty, Max Huppertz, Steph Karol, Aaron Kaye, Owen Kay, Nate Mather, Russell Morton, John Olson, Tyler Radler, Tereza Ranošova, and Katherine Richard.

A second set of communities that were supremely sustaining were friends from church and Northwood student housing. Thank you, Chad Lee for teaching me about friendship, faith, determination, and the differences between economic subfields. Thank you Reuben Hurst and Kallan Larsen for being friends at school, church, and everywhere else. I'm still convinced I learned econometrics because of you, Reuben. Dan Magelby provided me with more attention and help than any neighbor or professor on sabbatical need share, and I deeply appreciate it. Thanks to Dian Lu He for being the epitome of a neighbor and friend. Having a community to come home to makes any difficulty at school more surmountable. Thanks too to Adam Bennion, Jessi and Chayce Baldwin, Jake and Yanni Collins, Aaron and Emily Frutos, Kolby Gadd, Sara Lee, Amie and Kaleb Martin, Sara and Stephen Mitton, Annie Otte, Amy and Derrell Sonntag, John Richardson, Brent and Jana Ward, and Jared Williams for your examples, love, and repeated support. Finally, to the youth and parents of the Seminary Program for The Church of Jesus Christ in Ann Arbor: Thank you!! Teaching seminary for four of the last five years grounded me in the things that matter most, and your examples were daily reminders to me that in Christ we can accomplish even the hardest things.

I never would have finished a PhD without first setting out on this path. Thank you to Osvaldo and Adriana Martinez and John Hilton for showing me what difference dedicated teachers can make in a nineteen year old's life. Thanks to Craig Palsson for planting the seeds of pursuing a PhD in economics, and to Jim McDonald, Joe Price, Martha Bailey, and Lars Lefgren for giving me exposure to research, mentoring about the confusing world of academia, and having enough faith in me to guide me on that trajectory even though I started only months before applications were due. It's also been a great joy to hear encouragement and support from long-time friends in the real world. Thank you Will and Danielle Adams, Bryson and Emily Ensign, Phil and Reyna Morgan, and Ted and Caroline Tyler for asking about my research (and not about my research!) and for cheering me on. Also, thanks to Tanner Eastmond—it's great when one of your friends is also your coauthor.

When it comes to getting me to the beginning of a doctoral program, no one deserves more thanks than my family. Thank you Mom and Dad for always teaching me wherever we were. My first forays into research were with you, but that was only the smallest portion of what you did to get me here. Thanks Sam and Deb for loving me and for geeking out about economics with me. Thank you Treena for letting us take grandkids so far away from you, and thanks Tyler, Lexie, Kof, Tara, and Dallin for supporting us as only family can. Nowhere was the irreplaceability of family made more obvious than here in Ann Arbor. Thank you Hannah and Braden for always asking how school is going, putting up with weird renditions of research progress, and supporting me with love and my research with curiosity.

In the end no person on this earth deserves more thanks than Mindee. Thank you putting up with my economist-ness, for telling me "pep talk!," and having confidence in me when I find it hard to find. Sometimes I think we economists care too much about comparative advantage; I continue to be convinced that you have an absolute advantage in intellect, determination, patience, creativity, and love. Thank you for sharing that all with me. And thank you, Gracie, Callister, Brigham, and Reyna for hugs and prayers and love that is not conditional on my converged models or corrected typos. I love you!

Getting to this point is the culmination of a lot of hard work, but it is also the result of unmerited opportunity, unbelievable good fortune, and unearnable grace. While attending BYU I was struck by the motto "Enter to Learn, Go Forth to Serve." And though I have tried to serve, I sometimes feel that instead, I went forth to learn even more. Now as I graduate with a terminal degree, I eagerly anticipate the chance to go forth and serve in even deeper capacities. The dissertation is done, the degree complete; may it mark the beginning of a career of mentoring, teaching, and doing all the good I can with the opportunities entrusted to me.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

Table

LIST OF APPENDICES

Appendix

ABSTRACT

This dissertation contains three essays related to the efficient design of public programs. Governments that provide public goods must determine the appropriate policy, the optimal level of provision, and the criteria for participation. Each essay explores one of these aspects in a specific setting. Chapter I studies participation rules for public education programs; Chapter II, policy choice for alternative energy subsidies; and Chapter III, how taxpayers perceptions affect efficient the optimal level of public goods provision in the context of state-funded higher education.

In Chapter I "Strategic Selection Around Kindergarten Recommendations" I explore participation in public education. Specifically, I measure the trade-offs between having parents or policy makers decide when children begin kindergarten. Using data from the state of Michigan, I find that allowing parents to strategically select around kindergarten recommendations increases average test scores but widens income- and racial-achievement gaps. I quantify this efficiency-equity trade-off and demonstrate that it arises from substantial differences between selection on gains between higher-income families and lower-income families. Mechanisms suggest that increased participation in means-tested prekindergarten programs would increase both efficiency and equity.

In Chapter II "Subsidies with Deadlines: Optimal Tax and Subsidy Policy with Multiple Instruments" Owen Kay and I explore the choice between multiple possible subsidy policies. Although standard theory shows Pigouvian subsidies correct externalities efficiently, many real-world subsidies end after a deadline or subsidize investment rather than output. We show that subsidies with deadlines are the best policy in many settings, but the 10-year deadline of the Production Tax Credit in the wind industry is likely too short. Each month energy markets forego the social benefits of over 500 GWh of wind energy because of the deadline.

Finally, in Chapter III "Public Good Perceptions and Support: Evidence from Higher Education Appropriations" Reuben Hurst, Andrew Simon, and I answer the question of how taxpayers' perceptions affect the level of public services they demand. Using a survey experiment, we find that on average information increases support for additional state spending on higher education by 5%. The effects are concentrated among groups with lower levels of support, which means that information also reduces polarization.

CHAPTER I

Strategic Selection Around Kindergarten Recommendations

1.0 Abstract

What are the costs and benefits of allowing parents to choose when their children start public school? This paper uses two birthday-based discontinuities and marginal treatment effects methods to estimate how waiting a year to start kindergarten affects children whose families strategically select around recommendations about when to begin. Data from a cohort of kindergartners at Michigan public schools reveal that—counter to prevailing conjectures—children who wait to enter kindergarten would have been the lowest achieving in third grade, but they benefit the most from added investments the year before kindergarten. Although strategic selection increases average achievement, it widens racial- and income-achievement gaps, partly because only higher-income parents select on children's gains from waiting. Whereas a naive comparison with no selection on gains wrongly suggests that strategic selection reduces both scores and gaps, I show that it raises scores but widens gaps, presenting an equity-efficiency tradeoff. Analyzing mechanisms suggest that enrollment in means-tested public prekindergarten would simultaneously raise average scores and shrink achievement gaps, diminishing limitations on how well lower-income families can take advantage of "the gift of time."

1.1. Introduction

Across the world, countries, states, and school districts use "birthday cutoffs" to assign children to public-education cohorts. These cutoffs are either recommendations that allow parental choice or requirements that do not. By comparing children with birthdays around these cutoffs, research has shown that up to 90% of families follow recommendations (Bassok and Reardon, 2013) and that complying with a recommendation to *wait* until six to start public education increases scholastic achievement through college (Dhuey et al., 2019; Routon and Walker, 2020). How do families make these strategic decisions? And how does selection around recommendations affect academic achievement and equity? In addition to depending on the well-known effects of following a recommendation to wait, answering these questions depends on how waiting affects families who *do not* comply with birthday-based recommendations.

I explore selection into waiting by comparing families with different reluctance to wait: those who would always wait no matter the recommendation, those that eagerly comply with recommendations to wait, and those who only reluctantly comply with requirements to wait. I describe selection and effect heterogeneity in a marginal treatment effects framework (Heckman and Vytlacil, 2005; Mogstad et al., 2018). Using two sequential birthday cutoffs (one recommending waiting and another requiring it), I estimate the selection and effects with a fuzzy multi-cutoff regression discontinuity (Cattaneo et al., 2016) among a cohort of first-time kindergarteners in Michigan public schools. I measure selection in achievement levels with differences in scores at the recommendation cutoff (Black et al., 2022; Bertanha and Imbens, 2019; Kowalski, 2022b). Then I explore selection on achievement gains by comparing effect sizes between these groups, which requires an ancillary assumption for extrapolation (Cattaneo et al., 2020; Brinch et al., 2017). Together, the selection in levels and the selection on gains characterize the efficiency and equity implications of allowing strategic selection around kindergarten recommendations.

I document three main findings. First, there is negative selection on levels in to waiting. In other words, children who are more reluctant to wait have higher third-grade test scores (i.e., comparing scores had no one waited). Compared to children who always wait and who eagerly comply with recommendations to wait, those who reluctantly comply with requirements to wait score at least 0.42 standard deviations higher on third-grade math tests. (For context, grade repeaters score about 0.39 standard deviations below those who advance, see Jimerson, 2001, for a meta-analysis.) This negative selection in levels into waiting contradicts the prevailing wisdom that children who would always wait are positively selected because they come from wealthy, white, highly-educated families (Schanzenbach and Larson, 2017) or are higher achieving (Fortner and Jenkins, 2017).[1] By measuring selection in potential outcomes, rather than covariates or realized outcomes, my contribution suggests that children who wait are negatively selected in third grade achievement. This negative selection in levels is consistent with strategic efforts to start children in school only when they are "kindergarten

[1]Fortner and Jenkins (2017) suggest positive *and* negative selection because some redshirters are higher achieving and others are more likely to be in special education. The share in special education is small enough that this analysis would suggest positive selection on average.

ready."

Second, selection around recommendations increases average test scores because children who are more reluctant to wait experience smaller gains. Waiting increases the third-grade scores of children who always wait by at least 0.63 standard deviations—much more than it affects children who eagerly comply with recommendations (0.32) or who reluctantly comply with requirements (0.23). (For context, a one standard deviation increase in teacher value added raises early math scores by about 0.20 standard deviations.[2]) This pattern of positive selection on gains explains the economic underpinnings of research focused on families who comply with recommendations (e.g., Bedard and Dhuey, 2006; Elder and Lubotsky, 2009; Black et al., 2011; McCrary and Royer, 2011; Bedard and Dhuey, 2012; Cook and Kang, 2016) and on the causal effects of "redshirting" and "early entry" (Cook and Kang, 2018; Jenkins and Fortner, 2019; Molnar, 2020). My contributions are estimating effect heterogeneity, exploring the nature of selection on gains, and showing that selection around recommendations increases average test scores. Failing to account for that selection on gains would lead one to wrongly conclude that sorting around recommendations has *reduced* achievement in Michigan.

Interestingly, the positive selection on gains is driven entirely by higher-income families. Among higher-income families, children who always wait gain more than those who comply with recommendations or requirements to wait, but similar heterogeneity is absent among lower-income families. Documenting this heterogeneity is possible because I estimate selection on gains separately by income groups, relaxing typical shape restrictions in selection models (see discussions in Brinch et al., 2017; Mogstad et al., 2018; Kline and Walters, 2019). Allowing for similar heterogeneity may affect discussions about other public programs that allow for self-selection like voluntary job training, school and major choice, applications for means-tested services, provider choice in universal healthcare systems, and plan choice in utilities, insurance, or other public markets (e.g., LaLonde, 1986; Brand and Xie, 2010; Einav et al., 2010; Walters, 2018; Finkelstein and Notowidigdo, 2019; Ito et al., 2021).

Finally, I find that selection around recommendations widens income-achievement gaps because children from higher-income families experience larger testing gains from waiting. Although higher-income families are only slightly more likely to wait, their children's scores increase three times more than other children (0.48 standard deviations relative to 0.16). Extrapolating away from the cutoff, I find that parental selection around kindergarten recommendations is responsible for up to 15% of the income-achievement gap in third grade,

[2]A meta-analysis of 852 randomized controlled experiments in K-12 education estimates the 99th, 90th, and 80th percentiles of intervention impacts on third-grade math scores at 0.75, 0.36, and 0.23 standard deviations respectively (Kraft, 2020). See Fryer (2017) for a more detailed review of effects.

validating conjectures that selection around recommendations reinforces learning gaps (Graue and DiPerna, 2000; Deming and Dynarski, 2008) and equity-based discussions in support for requirements (e.g., Illinois, General Assembly, 2019). My contribution is estimating heterogeneity over observed and unobserved dimensions to quantify the equity-efficiency tradeoff that results from strategic selection.

Motivated by these gaps, counterfactual policy simulations suggest that eliminating strategic selection is a relatively inefficient way of closing gaps compared to expanding prekindergarten opportunities for lower-income families. Descriptive analyses show that higher-income families tend to invest in their children while they wait to enter kindergarten (especially through preschool) than do lower-income families; however, children in similar preschools experience similar gains regardless of family background. These findings suggest that lower-income families may benefit less from waiting because they have limited access to high quality investments like preschool. Without independent variation in investments, these findings may not capture causal effects, but they complement rigorous research showing that children from lower-income families benefit more from public prekindergarten programs because their counterfactual care arrangements lead to lower achievement (Kline and Walters, 2016; Felfe and Lalive, 2018; Cornelissen et al., 2018). The unequal benefits from the "gift of time" motivate counterfactual policy simulations comparing the entry recommendations in the *status quo* to entry requirements and to expanded means-tested prekindergarten. I find that both policies would reduce achievement gaps, but whereas prekindergarten for low-income families would raise average test scores, enforcing entry requirements would lower them.

While not its main focus, this paper also describes a new method for extrapolation away from an RD cutoff that may be of interest to other practitioners. Researchers who can identify (average) marginal treatment effects at an RD cutoff can use a parallel trends assumption in the spirit of Dong and Lewbel (2015) and Cattaneo et al. (2020) to extrapolate policy-relevant treatment effects away from the cutoff. This assumption yields a class of testable implications, which I evaluate in my data. This approach connects with a large and growing literature on extrapolation away from RD cutoffs. Similar approaches to extrapolation also assume that the relevant heterogeneity is in observed and unobserved characteristics (Angrist and Rokkanen, 2015; Rokkanen, 2015), but my assumption does not require conditional potential outcomes to be constant, only average treatment effects. This approach to extrapolation is also similar to those that use the LATE to extrapolate (such as Dong and Lewbel, 2015; Bertanha and Imbens, 2019), but it does not require a homogeneity or "external validity" assumption across different (types of) individuals. It is this extrapolation result which enables the policy counterfactuals to measure equity and efficiency in the population and not just at the cutoff.

The remainder of the paper includes the following sections: (1.2) defining the conceptual and econometric framework for this research; (1.3) describing the data and policy context; (1.4) estimating the effects of waiting with the regression discontinuity and testing for selection in levels and selection on gains between compliers; (1.5) exploring selection in levels and selection on gains for children who would always wait even when recommended to start kindergarten; (1.6) presenting suggestive results about mechanisms, the welfare framework, and policy simulations; and (1.7) containing my conclusion, discussion of results, and possible future research.

1.2. Conceptual Framework

This section sets out a model for family decision making, discusses the economics of selection, and presents the marginal treatment responses and effects that must be estimated to measure the costs and benefits of allowing strategic selection around kindergarten entry recommendations.

1.2.1 First Stage: Waiting to Start Kindergarten

I begin with a definition of treatment. Given their child's birthday, r, a family can either start kindergarten when the child turns five or wait until the next year.[3] Let $W \in \{0, 1\}$ be the decision to "wait to enter kindergarten" (similar to definitions in Black et al., 2011; Dhuey et al., 2019). Defining treatment as "waiting" as opposed to starting kindergarten may sound counterintuitive to some, but waiting is the treatment because it implies additional investments before starting public school.[4] Heterogeneity in the returns on this investment will have implications for efficiency and equity (as discussed in Appendix A.2).

There are three concrete advantages to defining treatment as waiting to start kindergarten at six. First, waiting is the one behaviorally-relevant decision for parents and policy makers. Although research often studies entry age, relative age, testing age, and their interactions separately, my definition does not separate them because they are not separately manipulable by the agents. Second, this binary treatment avoids well-known identification issues with continuous age-based definitions of treatment (Angrist and Pischke, 2008; Barua and Lang, 2016). Finally, defining treatment as waiting to start kindergarten at six makes

[3] Allowing a third decision would require either enrolling three-year-old children who are about to turn four or six-year-old children who are about to turn seven (a violation of compulsory schooling laws in Michigan and 35 other states).

[4] Investments that could affect a child's level of preparation and the benefit realized from the (delayed) stream of services.

it clear that redshirting (waiting when recommended to start), early entry (starting when recommended to wait), and on-time entry are not three categorical "treatments" as previously characterized (Cook and Kang, 2018; Molnar, 2020; Jenkins and Fortner, 2019). Rather they are manifestations of selection into and out of treatment. Appendix A.3 compares my framework with prior work answering both descriptive and causal questions in more detail.

Having defined waiting as the decision of interest, I refer to parents as the main decision makers. While professionals often comment on kindergarten readiness and districts may provide entry guidelines, families make their own strategic decisions under a policy of recommendation. As decision makers, parents value the anticipated gains from waiting to enter (e.g., physical, cognitive, or social development) against the costs of doing so (e.g., childcare and foregone wages). Different types of families may have heterogeneous preferences over costs and gains and may make different investments if they choose to wait. This heterogeneity motivates the investigation of how strategic selection around recommendations affects average student performance and disparities in achievement across groups.

Although parents have decision-making power, their choices and incentives may be affected by policy recommendations or requirements. Let $z \in \{0, 1, 2\}$ characterize the policy incentives a family faces in deciding whether to wait. Each family will face one of these incentives: a recommendation to start kindergarten at age five ($z = 0$), a recommendation to wait ($z = 1$), or a requirement to wait ($z = 2$).[5] In practice, these policies usually depend on a child's birthday, r, but for now it may be helpful to think of them as randomly assigned to build intuition.

Different families may respond to these incentives in different ways. Let x be a family's observed characteristics, and U_W characterize the family's unobserved reluctance to wait. Without loss of generality let $U_W \sim [0, 1]|r, x$ reflect the percentile of reluctance to be treated (Mogstad et al., 2018). In other words, families with lower values of U_W will choose to wait in the face of weaker incentives whereas families with higher values of U_W require a stronger impetus to induce them to wait. A family will decide to have their child wait if the unobserved cost is less than some threshold that is a function of the child's birthday, r, and (optionally) other characteristics:

$$W = \mathbb{1}[U_W \leq p_z(r, x)] \text{ where } p_z(r, x) = P(D = 1|z, r, x) \tag{1.2.1}$$

[5] A fourth policy, prohibiting waiting and requiring children start at age five, is possible but I exclude it from this exposition because it doesn't occur in my setting. See Molnar (2020) for an evaluation of one such policy in Hungary.

1.2.1.1 Characterizing Families by Their Reluctance to Wait

Strategic responses to these incentives characterize four types of families by their partial compliance (Imbens and Angrist, 1994; Mogstad et al., 2020). I define "always takers" as those who will wait no matter the recommendation. "Eager compliers" are those induced to wait by either recommendations or requirements, and "reluctant compliers" are families who can only be induced wait by a requirement. Families who no matter what cannot be induced to wait are "never takers."[6]

Figure 1.1 illustrates the connection between these groups and families' innate reluctance to wait, U_W. Imagine a set of families with similar characteristics, x, and birthdays, r. Among them always takers will choose to wait even if they are recommended to start kindergarten ($z = 0$) and are revealed to have the lowest reluctance to wait ($U_W \leq p_0(r, x)$). Facing the same incentives, however, eager compliers and reluctant compliers do not wait because they are more reluctant ($U_W > p_0(r, x)$). Thus $p_0(r, x)$ partitions always takers from eager and reluctant compliers as shown at the bottom of Figure 1.1 and defined in Equation 1.2.1. In the same way, the always takers and eager compliers are partitioned from the reluctant compilers by $p_1(r, x)$ among children who are recommended to wait ($z = 1$), and $p_2(r, x)$ would partition the always takers and compliers who wait from the never takers. In my analyses I assume that there are no never takers, or that $p_2 = 1$.[7] Also note that Equation 1.2.1 implies that there are no "defiers."[8]

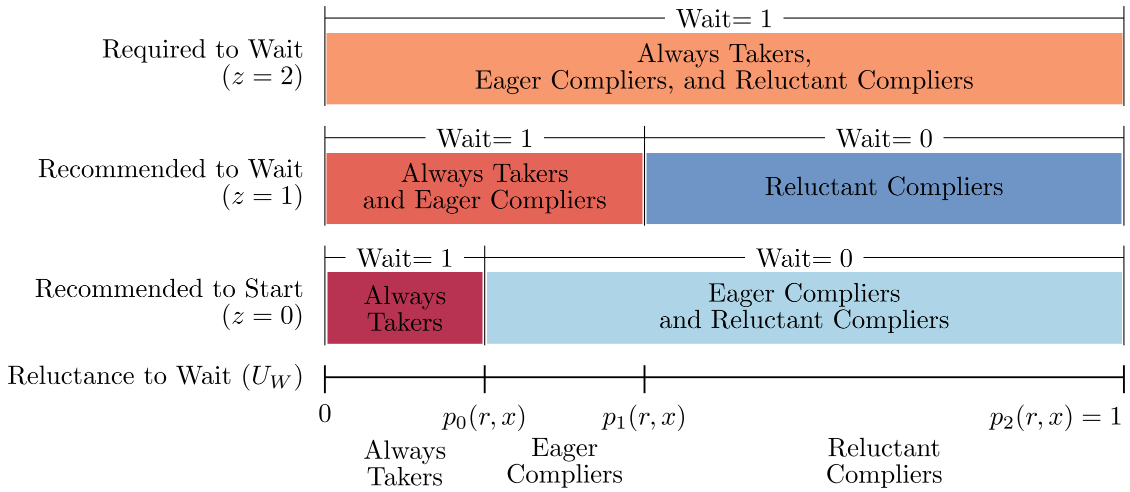
To conclude with the first stage, consider two notes. First, the probabilities $p_z(r, x)$ that characterize the shares of always takers, eager compliers, and reluctant compliers are designed to vary with r (and be allowed to vary with x). As long as $U_W \perp Z|r$, the share of always takers (among children with a given birthday) are those with $U_W \in [0, p_0(r, x)]$, the share of eager compliers $U_W \in (p_0(r, x) - p_1(r, x)]$, and the share of reluctant compliers $U_W \in (p_1(r, x), 1]$. There are two implications of the probabilities varying over r, first, this variation means that identifying and comparing effects at different dates requires extrapolation, and, second, this variation will help identify the nature of selection around kindergarten entry recommendations.

[6]There are only four groups rather than the six of Mogstad et al. (2020) because in this setting z_1 compliers and eager compliers functionally become the same group, as do z_2 compliers and reluctant compliers.

[7]This is reflected in Figure 1.1 where never takers are not depicted. I do this because never takers would seem to be violating a requirement, and in the data there fewer than 100 students identified as never takers in each cohort (less than 0.1%)—a number of which may be due to data or measurement error. Dropping these students eases exposition but does not change any results.

[8]Defiers would respond against any recommendation or requirement, for example families who would enter early if recommended to wait but redshirt if recommended to start. Ruling out such behavior seems plausible and is analogous to LATE "monotonicity" assumptions Imbens and Angrist (1994) used in existing research using birthday cutoffs as instrumental variables. See Vytlacil (2002) for the equivalence between LATE "monotonicity" Imbens and Angrist (1994) and treatment take-up latent index models.

Figure 1.1: Reluctance to Wait for Always Takers, Eager Compliers, and Reluctant Compliers



Note: This figure depicts the way that always takers, eager compliers, and reluctant compliers can be ordered by their reluctance to wait to enter kindergarten. This reluctance to wait is captured by $U_W \in [0,1]$. The values $p_z(r,x)$ represent the probability of waiting when facing different recommendations or requirements, z, for students with birthdays r, and characteristics x. Families with $U_W \leq p_z(r,x)$ will wait to enter kindergarten and others will not.

Second, as Figure 1.1 illustrates, these groups are ordered by U_W which will inform how I compare test scores and the effects of waiting between groups. Specifically, always takers are less reluctant to wait (lower U_W) than eager compliers, who in turn are less reluctant to wait than reluctant compliers. The implications of strategic selection depend on always takers and reluctant compliers—as they *do not* comply with recommendations. The next section leverages the fact that these families are on opposite extremes of U_W to formalize ideas of selection in levels and selection on gains.

1.2.2 Second Stage: The Effects of Waiting

With the treatment decisions and groups defined, I consider the second stage of the data generating process: the effects of waiting. Let each individual have potential test scores had they started kindergarten at five (Y_0) or waited a year (Y_1) given by: $Y_W = g_W(r, x, u, \gamma_W)$. These scores are functions of children's birthdays, r; characteristics, x; reluctance to wait, u; and other unobservables γ_W (that may vary by the decision to wait). Which potential scores are realized depends on the treatment chosen: $Y = Y_1 W + Y_0(1 - W)$.

To characterize the selection around kindergarten entry recommendations, I define three functions borrowed from the marginal treatment effects (MTE) literature. The first and third are the marginal treatment response functions of Mogstad et al. (2018), and the second is the MTE function of Heckman and Vytlacil (2005).

$$\begin{aligned} \text{Selection in levels along } U: &\qquad m_0(u,r,x) = \mathbb{E}[Y_0|R=r, X=x, U_W=u] \\ \text{Selection on gains along } U: &\qquad \tau_{MTE}(u,r,x) = \mathbb{E}[Y_1 - Y_0|R=r, X=x, U_W=u] \\ \text{Selection on levels + gains along } U: &\qquad m_1(u,r,x) = \mathbb{E}[Y_1|R=r, X=x, U_W=u]] \end{aligned}$$

The first marginal treatment response (MTR) function focuses on average test scores had students started kindergarten at age five, capturing how these scores vary between families who are more or less reluctant to wait. This generalization of selection bias (Kowalski, 2022b) is what I use to conceptualize "selection in levels." In the presence of positive selection in levels, families who are more likely to wait (lower U_W) have higher Y_0 than those who are less likely to wait (higher U_W), implying m_0 is decreasing in u. In the presence of negative selection, m_0 would be increasing in u. Recovering the information from m_0 will help answer the question about the nature of selection around entry recommendations.

The second function is the "marginal treatment effect" (MTE) function. It defines how the effects of waiting vary between families who are more or less reluctant to wait. I use this function to conceptualize "selection on gains." In the presence of positive selection on gains, families who are more likely to wait (lower U_W) experience larger test score improvements than those who are less likely to wait (higher U_W), implying τ_{MTE} is decreasing in u. In the presence of negative selection on gains, τ_{MTE} would be increasing in u. In addition to describing the nature of selection around entry recommendations, the information from τ_{MTE} has implications for the equity and efficiency of recommendations and requirements as described in Appendix A.2.

The last function is a MTR function focused on average test scores had students waited until six to begin kindergarten, capturing how these scores vary between families who are more or less reluctant to wait. This second MTR reflects changes in baseline outcomes (what would have happened had they entered at age five) and changes in the average effects of waiting across the quantiles of U_W. Although there is an easy economic intuition for the slopes of m_0 and τ_{MTE}, changes in m_1 are less interpretable on their own. In fact many combinations of monotonic m_0 and τ_{MTE} generate m_1 functions are increasing and decreasing at different u.

1.2.2.1 Identification with Birthday Cutoffs

Although these MTR and MTE functions are economically meaningful, they are not nonparametrically identified with a discrete instrument. For identification this paper estimates more aggregated statistics relevant to the groups of always takers (who wait no matter what), eager compliers (who wait if recommended or required), and reluctant compliers (who wait only if required). Let the following reflect the average potential outcomes of children from each group $g \in \{at, ec, rc\}$ at a given birthday:

$$\mu_{W,g}(r, x) = \mathbb{E}[Y_W | R = r, X = x]$$

These means are weighted averages of the underlying MTR functions m_0 and m_1,[9] but unlike the MTRs, many of these means will be identified. To see this I return to the discussion of birthday cutoffs postponed in Section 1.2.1.

A birthday cutoff is a date r^* where the policy of recommendation or requirement changes from z to z'. Consider two types of policies that will be present in the empirical exercises below. Perhaps the most intuitive policy is to recommend starting for children who turn five before r^* and recommend waiting for children who turn five after r^*. This policy features a change $z = 0$ to $z' = 1$. Using the change in z as an instrument for waiting in a fuzzy RD, this type of cutoff identifies the scores of always takers who wait, $\mu_{1,at}(r^*)$; of eager compliers induced to wait, $\mu_{1,ec}(r^*)$; of eager complies induced to start, $\mu_{0,ec}(r^*)$; and of reluctant compliers who start, $\mu_{0,rc}(r^*)$—suppressing x to save notation. The local average treatment effect (LATE) on eager compliers is also identified: $\tau_{ec}(r^*) = \mu_{1,ec}(r^*) - \mu_{0,ec}(r^*)$ as is the selection between eager and reluctant compliers, $\mu_{0,ec}(r^*) - \mu_{0,rc}(r^*)$.

A second, more standard policy is to recommend children who turn five before r^* start kindergarten and to require that children who turn five after r^* wait another year. This policy features a change $z = 0$ to $z' = 2$. With this policy change I can identify the average effect on eager and reluctant compliers whose behavior is changed by the requirement ($\mathbb{E}[Y_1 - Y_0 | g \in \{ec, rc\}, R = r^*]$) as well as the average scores of always takers who wait, $\mu_{1,at}(r^*)$. This policy change is typical of most of the applied work on kindergarten "entry age" in the United States (e.g., Dhuey et al., 2019).

I will use these conditional expectations by group, $\mu_{W,g}$, to operationalize my study of selection in levels and selection on gains. Since the groups are ordered by their unobserved reluctance to wait, U_W, I measure selection in levels by the comparing expected test scores of children who start kindergarten at five and measure selection on gains by comparing the effects of waiting on test scores. At a given cutoff r^*, $\mu_{0,at} > \mu_{0,ec} > \mu_{0,rc}$ would characterize positive selection in levels, and $\tau_{at} > \tau_{ec} > \tau_{rc}$ would characterize positive selection on gains.

[9] Specifically $\mu_{w,g}(r, x) = \int_0^1 \omega_g(r, x) m_w(u, r, x)\, du$ where $\omega_g(r, x)$ indicate whether u is within the range for group g at (r, x), scaled by the size of the group.

1.2.2.2 Extrapolation from Birthday Cutoffs

While cutoffs are powerful for identification, the big limitation of identification by regression discontinuity is that estimates are local: they only reflect the effects of individuals at the cutoff $r = r^*$. This limitation is doubly true for estimates from a fuzzy RD, as they also are relevant only to a set of compliers who are moved by the change in z: $U_W \in [p_z(r^*, x), p_{z'}(r^*, x)]$. There is a growing interest in external validity and extrapolation in RD settings. Here I will discuss the approach used in this paper and how it compares to other approaches to extrapolation.

My research question focuses on how families select around kindergarten recommendations and how they are affected by waiting. To answer the question I need to know how waiting to enter kindergarten affects different types of families. Because each birthday cutoff has its own set of compliers, estimates of the effects of waiting on different groups must leverage different cutoffs. Any difference in the effects could be driven by true selection on gains (heterogeneity over U_W) or differences in composition (heterogeneity over r[10]). I need to be able to extrapolate effects across birthdays to compare effects estimated at different cutoffs and identify the patterns of strategic selection.

To extrapolate, I adopt a parallel trends assumption. I assume that in expectation, children's potential test scores had they waited to enter kindergarten evolve in parallel to their potential test scores had they started kindergarten at age five (for some extrapolation window $[\underline{r}, \bar{r}]$ around a cutoff r^*):

$$\forall\, \tilde{r} \in [\underline{r}, \bar{r}]: \qquad \mathbb{E}[Y_1 | R = r^*, u, x] - \mathbb{E}[Y_1 | R = \tilde{r}, u, x] = \mathbb{E}[Y_0 | R = r^*, u, x] - \mathbb{E}[Y_0 | R = \tilde{r}, u, x]$$

This assumption is inspired by the assumption in Cattaneo et al. (2020) for extrapolation in a RD setting with multiple cutoffs. My assumption is that treated and untreated outcomes evolve in parallel for units facing a given cutoff, whereas Cattaneo et al. (2020) assume that untreated outcomes evolve in parallel for all cutoffs. An implication of my parallel trends assumption is that $\tau_{MTE}(u, r^*, x) = \tau_{MTE}(u, \tilde{r}, x)\ \forall \tilde{r} \in [\underline{r}, \bar{r}]$. In other words, conditional on u and x, there is no appreciable heterogeneity in treatment effects over r. This simplification allows me to compare treatment effects estimated at different dates.

Conveniently, the parallel trends assumption has a testable implication. This test is made possible by the longitudinal assessment data. Under parallel trends, the slope of test scores over birthdays is the same for students had they started or waited. Note that we expect this causal effect to be negative: students with later birthdays are younger when they take the

[10]This heterogeneity could include either direct differences in effects over r or compositional differences in x or γ_W over r.

test, and younger students tend to do worse on the same test, all else equal. Although we do not observe the potential third grade test scores of students had they started (and tested in third grade at age eight) or waited (and tested in third grade at age nine), we do observe their test scores in third grade (at age eight) and fourth grade (at age nine). By regressing scores in third and fourth grade on birthdays over some range where the probability of waiting doesn't change (e.g., when $z = 2$ or r is very negative), I can determine how close to parallel the trends are. Failure to reject does not imply parallel trends in potential outcomes, but it builds credibility. Furthermore, to the extent to which the relationship becomes attenuated (amplified) over time, extrapolations to before the cutoff will underestimate (overestimate) the true effects and extrapolations after the cutoff will overestimate (underestimate) the effect.[11] For any violation, I can bound the amount of selection on gains attributable to nonparallel trends, reminiscent of the intuition in Rambachan and Roth (2022) for a traditional difference in differences.

Comparisons to Other Methods for Extrapolation. This approach to extrapolation is unique because it is based in the estimation of heterogeneity. Essentially I can extrapolate over r because for all individuals with a given U_W and x, average treatment effects are the same for all r. My approach to extrapolation shares key similarities and differences from the three other approaches in the literature: assuming r is ignorable, using multiple cutoffs, and leveraging the LATE.

My parallel trends assumption implies that x and u are the main drivers of heterogeneity, as do approaches that assume that r is ignorable conditional on certain characteristics—observed or latent (Angrist and Rokkanen, 2015; Rokkanen, 2015). In this approach treatment effect may over r through compositional changes in x or the predicted latent type, but r cannot directly affect outcomes. While plausible in cases with rich x and predictable r (like test score cutoffs), this approach fails whenever r is not (conditionally) excluded (e.g., birthdays, close elections, and income thresholds). In a fuzzy RD, this approach also requires that r have no effect on $p(x, r)$ conditional on characteristics. My approach is similar in that it assumes that individual characteristics are the main driver of heterogeneity, but the parallel trends assumption allows both the probability of treatment and potential outcomes to vary over r (even conditional on x and u)—just requiring that treated and untreated potential outcomes move in parallel.

My parallel trends assumption utilizes a setting with multiple cutoffs as do extrapolation methods leveraging data on different subpopulations (with similar r) facing different cutoffs.

[11] This is assuming a negative relationship. In general if the slopes become more positive (negative) over time extrapolating to the cutoff will underestimate (overestimate) the true effects and extrapolations after the cutoff will overestimate (underestimate) the effect.

In this case a similar parallel trends assumption on untreated outcomes can extrapolate between the cutoffs (Cattaneo et al., 2020).[12] Essentially, the assumption is that the effect of r is the same on units that respond to each cutoff. This approach is agnostic about what drives heterogeneity over r, but only applies in a sharp RD or settings with no always takers. My approach is similar in that it uses the multiple cutoffs for identification, but my intent is to compare effects at different cutoffs rather than estimate effects at r between cutoffs. Furthermore, by making assumptions about treated outcomes, I can extrapolate in fuzzy RD settings with two-sided noncompliance at one or more of the cutoffs, can extrapolate beyond the first and last cutoffs in r, and can still extrapolate when the relevant cutoff to each unit is unobserved.[13]

My approach is also similar in concept to those that use the LATE to extrapolate, but I estimate richer heterogeneity. For example assuming the LATE is externally valid (no selection and no effect heterogeneity) implies that r is the main driver of heterogeneity (Bertanha and Imbens, 2019).[14] If true, this setting makes extrapolation easy: the treatment effect at r is always identified by $\mathbb{E}[Y|W=1,r]-\mathbb{E}[Y|W=0,r]$. But the external validity assumption may often not be true. In this case, changes in the LATE at the cutoff can be extrapolated by the mean value theorem (Dong and Lewbel, 2015).[15] My approach uses the LATE for extrapolation, but does not require externally validity between compliers and other groups because it directly incorporates estimates of the effects on those groups. The limitation of my approach is that the researcher either needs multiple cutoffs or additional ancillary assumptions to identify effects beyond the (eager) complier LATE.

Relative to the existing approaches, the power of my method for extrapolation is that estimating MTE-related parameters can simultaneously overcome *both* dimensions of locality in the fuzzy RD. By estimating effects beyond the traditional LATE, I overcome the locality in U_W, and then those parameters can be the basis of extrapolation away from r^* the parallel trends. The policy-relevant treatment effects change with the share and composition of compliers (giving a new interpretation to the treatment effect derivative of Dong and Lewbel, 2015), but because the marginal treatment effects are the same, the effects can be extrapolated to explore changes in the strength of cutoff incentives or the placement of the cutoff itself.

[12] And an additional assumption for treated outcomes similar to mine would allow for extrapolation beyond the cutoffs.

[13] In fact if appropriately extended to treated outcomes, the assumption in Cattaneo et al. (2020) would also allow for extrapolation beyond the first and last cutoffs in r, if not for two-sided noncompliance and unobserved groups.

[14] Bertanha and Imbens (2019) also propose testable restrictions of this external validity assumption.

[15] In practice this insight is usually employed in the negative, i.e., showing that the LATE is constant to support external validity, see Cerulli et al. (2017).

1.3. Kindergarten, Policy, and Data in Michigan Public Schools

This section explains the policy and data context from Michigan. There are two birthday cutoffs in Michigan that allow me to identify the effects of waiting to enter kindergarten on different subpopulations and to describe the patterns of selection around recommendations (both in levels and on gains).

1.3.1 Michigan Public School Data Contain Needed Information

I use data on the universe of public school students in the state of Michigan. I employ longitudinal datasets of K-12 enrollment and assessments from the Michigan Education Data Center (MEDC). These data cover all students and all state assessments from the 2001-2002 school year until the 2018-2019 school year. I create a main analysis sample of first-time kindergarteners who turned five between March 1, 2013 and February 28, 2014 and a secondary sample covering March 1, 2002 to February 28, 2015. Appendix A.1 details the data cleaning and sample selection process which drops 1% of first-time kindergarteners. Enrollment records also contain demographics and date of birth, which is fundamental for identification using month as an IV recovers hard-to-interpret effects.[16]

In my analysis, I focus on third grade test scores, the nearest-term outcomes to kindergarten entry. Focusing on third-grade scores allows me to investigate selection and effect heterogeneity without worrying as much about differential outcome dynamics between students with different observed and unobserved characteristics. Note that treatment effects on third-grade scores compare students who test at different ages in the same grade (third graders who are nine instead of eight) rather than at about the same age in different grades (nine-year-old students in third grade instead of fourth grade). These within-grade effects are the comparisons of interest for decision makers. For example, students will be tracked into accelerated or remedial paths based on their in-grade scores; report cards and high school transcripts list in-grade performance; states administer high school exit exams in a given grade not at a given age; and students who apply to college will be evaluated against same-grade rather than same-age peers.[17] Fortunately, in Michigan tests are psychometrically calibrated for comparing scores on a given grade's tests across years.

[16] There are two countervailing issues. First, for a given set of compliers, the month-to-month variation biases the estimated effects toward zero because of increasing noncompliance around the cutoff (The reduced forms of the the RD and the month-to-month differences are similar, the but the first stage estimated in an RD is smaller). At the same time, the month-to-month variation estimates effects over a broader support of U_W, making the estimate a mix of the targeted complier-LATE and selection on gains. Attar and Cohen-Zada (2018) explore other reasons why exact date of birth is important in kindergarten entry research.

[17] And waiting to enter kindergarten does indeed affect these and similar outcomes (Ponzo and Scoppa, 2014; Hemelt and Rosen, 2016; Dhuey et al., 2019; Routon and Walker, 2020).

1.3.2 Michigan Policies Affecting Kindergarten Entry

In Michigan nearly 100,000 students enter kindergarten each year—typically at age five.[18] As in most states, kindergarten recommendations are based on birthday cutoffs. Before 2013 the cutoff date was December 1. Children with birthdays before December 1 were recommended to enter kindergarten in the year they turned five, and those with birthdays after December 1 were required to wait. In this era, selection around these recommendations was one-sided: families who were recommended to start were allowed to wait, but other families were required to wait (functionally the state enforces requirements by not giving districts funding for children who have not turned five by December 1).

In the 2010s, Michigan moved the assignment cutoff from December 1 to September 1. Rather than do this all at once, the cutoff date was moved back one month each year from December 1 in the fall of 2012, to November 1 in 2013, to October 1 in 2014, until the new birthday cutoff was September 1 in the fall of 2015. During this time, the state also eased restrictions on early entry to more flexibly accommodate family's anticipated entry decisions, introducing a waiver system whereby children with birthdays between the assignment cutoff and December 1 (e.g., November 15) could still enroll.[19] Appendix Figure A.6 shows how the probability of waiting changed when the cutoff changed in 2013.

1.3.3 Focusing on the Cutoffs in 2013 Provides the Cleanest Comparisons

I focus on children affected by the November 1, 2013 cutoff for three main reasons. First, the waiver policy generated selection out of waiting that identifies selection in levels. In 2013 the share of early entrants at the cutoff rose to 55% (at November 2, 2013). These children who start when recommended to wait are revealed to be reluctant compliers. On the other side of the cutoff children who start are a mix of eager and reluctant compliers. By comparing the test scores of reluctant compliers to those of similarly-aged children on the other side of the cutoff, I can measure selection on achievement levels, and because there is so much selection out of waiting at the cutoff, that selection in levels is identified over a broad support of U_W.

Second, the two cutoffs identify the effects of waiting on eager and reluctant compliers. Students who turned five after November 1 were *recommended* to wait, but students who turned five after December 1 were *required* to wait. Since students who turned five between

[18]99.86% of students attend elementary school in districts that offer kindergarten. Because kindergarten is not mandatory in Michigan, districts decide whether or not to offer it (but if kindergarten is offered in a district, children must enroll in kindergarten rather than first grade).

[19]And districts could still receive state funds for enrolling them; in fact, districts were required to enroll any student who wanted to enter under this waiver system.

November 2 and December 1 were still allowed to start without waiting, the November cutoff identifies the LATE on eager compliers and the December cutoff identifies the LATE on reluctant compliers. Comparing these two LATEs is the first step in characterizing selection on gains.

Finally, having the two cutoffs close together is useful for extrapolation and external validity. To identify selection on gains, I need to extrapolate with the parallel trends assumption, and having the two cutoffs close together reduces the required scope for extrapolation. Furthermore, with the cutoffs so close together, the large changes in selection at the cutoff leave overall classroom composition relatively unaffected.[20] Because, the 2013 cutoffs reveals the nature of unobserved preferences around the cutoff without changing equilibrium behavior, the patterns of selection I find should generalize well to other settings and other states.

1.3.4 Descriptive Evidence of Selection

This subsection presents descriptive evidence for positive selection on gains and negative selection in levels. After exploring the shares of always takers, eager compliers, and reluctant compliers from the first stage, I explore selection around recommendations descriptively. The reduced form relationship suggests selection on gains, and comparing students who make the same waiting decision on either side of the cutoff suggests negative selection in levels.

First, note that the probabilities of waiting at the November 1 cutoff identify the shares of always takers, eager compliers, and reluctant compliers. Recall that $p_z(r)$, the probability that a children with a given birthday r who faces a recommendation or requirement z, describes the shares of each group. Because these probabilities are identified at the November 1 cutoff, the shares are as well. Panel (a) of Figure 1.2 illustrates this graphically. For example, among children recommended to start at November 1, I find that $p_0(r_{Nov}) = 0.18$, indicating that the share of always takers at November 1 is 18%. Similarly, among children recommended to wait at November 1, I find that $p_1(r_{Nov}) = 0.40$, indicating that the shares of eager compliers and reluctant compliers are 22% and 60% at November 1.

Given the information about group shares, the reduced form relationship suggests positive selection on gains because the two discontinuities in third-grade math scores are not proportional to the two discontinuities in waiting. As shown in Panel (b) of Figure 1.2, the probability of waiting jumps by about 0.48 at the December 1 cutoff and only 0.22 at the November 1 cutoff. Absent selection on gains, the effects on eager and reluctant compliers would be equal, and the jump in scores at December 1 would be more than twice the size

[20] The share of children with November birthdays who waited increased from 18% to 45%, but children with November birthdays only make up one twelfth of the population, so the overall effect was only about 2.25 percentage points : $(45 - 18)/12 = 2.25$.

as the jump at November 1. Nevertheless, Figure 1.2 shows this is not the case. The jump at December 1 is only about 30% bigger, suggesting that eager compliers (moved by the recommendation at the November cutoff) gained more from waiting than reluctant compliers (moved by the requirement at the December cutoff). In other words, it suggests positive selection on gains.

Finally, the Bertanha and Imbens (2019) test for external validity indicates the presence of selection in levels or selection on gains. Their external validity condition is that both treated and untreated outcomes are equal in expectation across always takers, compliers, and never takers (or in my setting reluctant compliers). In my context, this condition is equivalent to no selection in levels and selection on gains. A testable implication of the assumption is that there must be no differences in outcomes at the cutoffs, conditional on the decision to start kindergarten or to wait—an implication that the data do not support.

Panel (c) of Figure 1.2 reports average scores by children's birthdays *and* whether they waited to enter kindergarten and shows jumps in test scores at November 1 even conditional on waiting. Among students who do not wait (dark and light blue lines), scores may increase at November 1, suggesting that eager compliers are negatively selected in relative to reluctant compliers. Among students who do wait (red and orange lines), scores seem to decrease at November 1, suggesting that always takers are positively selected in levels or on gains (or both) relative to eager compliers. At December 1 there is not much of a discontinuity, but this comparison cannot discern whether this is because there is no selection in levels and no selection on gains or because the two net each other out.

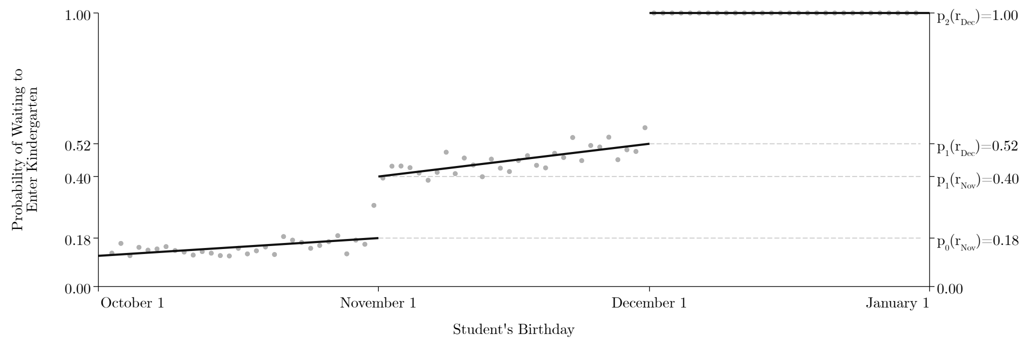
1.4. RD Estimates of Compliers' Selection in Levels and Selection on Gains

This section estimates the causal effects of waiting to enter kindergarten at the two cutoffs. I document selection on gains by testing whether the effects of waiting differ between eager and reluctant compliers. I document selection in levels by comparing the scores of eager and reluctant compliers that enter kindergarten in the same year. I explore selection in levels and on gains for always takers in the following sections.

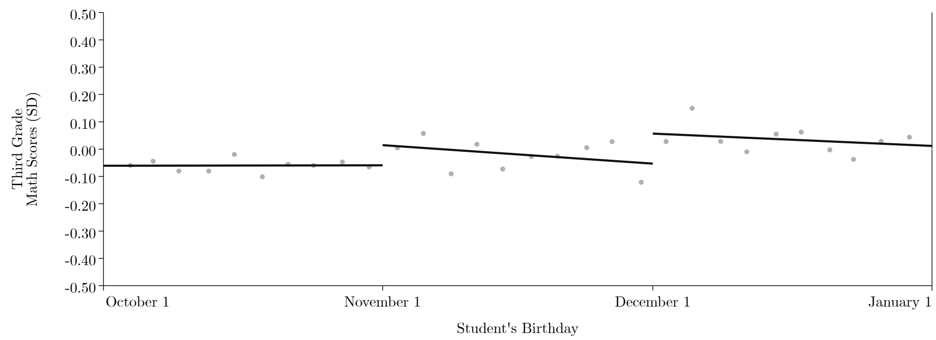
1.4.1 Both Eager and Reluctant Compliers Benefit from Waiting to Enter Kindergarten

I estimate the effect of waiting to enter kindergarteners on eager compliers, $\tau_{ec}(r_{Nov})$ using the November 1 cutoff and on reluctant compliers, $\tau_{rc}(r_{Dec})$, using the December 1 cutoff. I

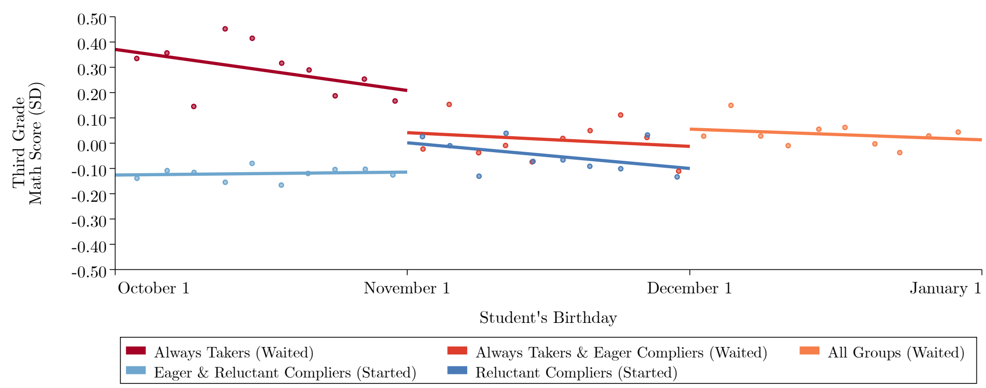
Figure 1.2: Changes in Waiting and Scores at Cutoffs Suggest Selection is Important



(a) Probabilities p_B and p_A Are Identified by the First Stage at November 1



(b) Third-Grade Math Scores Jump at the Two Cutoffs



(c) Even Conditional on Waiting, Outcomes Still Seem to Jump at the Cutoffs

Note: This figure shows patterns in waiting and achievement over birthdays. Panel (a) connects the information from an RD plot of the first stage to the unobserved cost of waiting U_W and the sample proportions illustrated in Figure 1.1. The graph shows both a scatter plot of the probability of waiting to enter kindergarten by birthday and the associated lines of best fit. The limits identify the unconditional probability of waiting on either side of the cutoff, $p_z(r)$. The second two panels display patterns in third-grade math achievement for students on different sides of the cutoffs. In Panel (b) points are three-day averages and in Panel (c) points are four-day averages. Students who turn five before November 1 are recommended to start kindergarten at age five, students who turn five between November 1 and December 1 are recommended to wait another year, and students who turn five after December 1 are required to wait. The third-grade test score measured in standard deviations is plotted over student birthdays. Reported levels for each group are reported local linear regressions with uniform kernels. The sample is comprised of first-time kindergarteners who turned five between October 1 and December 31, 2012 and for whom I observe third grade test scores.

do so with the following local instrumental variables regressions at each cutoff (for $k = 1, 2$):

$$Y_i = \tau_k W_i + \beta_k X_{ik} + v_{ik}$$
$$W_i = \gamma_k z_{ik} + \pi_k X_{ik} + u_{ik}$$

In each regression $z_k = \mathbb{1}(r > r_k)$ is a binary instrument for whether the student turned five after the relevant cutoff, $X_k = (r, r \cdot z_k)$ (with r_k normalized to zero), Y are third-grade math scores measured in standard deviations, and W is an indicator for whether a student waited to enter kindergarten. Weights are rectangular kernel weights for 30 days to either side of r_k.[21]

The main identifying assumption for $\hat{\tau}_1$ and $\hat{\tau}_2$ to be consistent for the target LATEs, τ_{ec} and τ_{rc} (suppressing dependence on r), is that potential outcomes be continuous over both discontinuities. Black et al. (2022) point out that this is an exclusion argument: If z affects outcomes through a mechanism other than W, it would violate the standard LATE exclusion assumption. In my setting this assumption is equivalent to discontinuities in individual potential outcomes at the cutoff. This assumption might be violated if policies or practices led a nonrandom subset of parents to plan births to one side of the cutoffs or encouraged differential childhood investments in children with birthdays to either side of the cutoffs. Appendix A.5 explores these and other potential confounding factors, finding no evidence that this assumption is violated.[22]

I find that the November 1 cutoff induces 22% of students to wait, and that these students can expect gains in their third-grade math scores of about 0.32 σ (τ_{ec}). Similarly, the December cutoff induces 48% to wait. These students can expect gains in their third-grade math scores of about 0.23 σ (τ_{rc}). Table 1.1 displays both sets of results including weighted local linear regressions of the first stage, reduced form, and fuzzy RD (or LATE). Appendix Tables A.2 and A.3 show that these results are not particularly sensitive to bandwidth, weighting, functional form, or covariate specifications. These effects are generally aligned with results from other work, although this is the first characterization to my knowledge of the effect of forcing early entrants to wait (an interpretation of τ_{rc}). Interestingly, the effect on eager compliers is about 45% larger than the effect on reluctant compliers. This finding is suggestive of intentional kindergarten entry decisions and selection on gains—a suggestion that needs to be tested formally.

[21] Note that this will produce equivalent estimates to differencing the limits from above and below at $r = r_k$ using a rectangular kernel $K(\cdot)$ with a bandwidth of 30 (Lee and Lemieux, 2010).

[22] Identification also requires that the instrument actually affect behavior, but this relevance condition is verifiable in the data. Monotonicity is implicit in the theoretical framework which requires that both instruments only increase an individual's probability of treatment.

Table 1.1: Waiting to Enter Kindergarten Increases Eager and Reluctant Compliers' Scores

Panel A: November 1 Cutoff	First Stage (RD: Wait to Enter)	Reduced Form (RD: Third Grade Math)	τ_{ec} (Eager Compliers) (Fuzzy RD: Third Grade Math)
Effect	0.225***	0.073*	0.325*
	(0.015)	(0.032)	(0.144)
Slope Before Cutoff	0.002***	0.000	-0.000
	(0.000)	(0.001)	(0.002)
Change in Slope	0.002*	-0.003	-0.003
	(0.001)	(0.002)	(0.002)

Panel B: December 1 Cutoff	First Stage (RD: Wait to Enter)	Reduced Form (RD: Third Grade Math)	τ_{rc} (Reluctant Compliers) (Fuzzy RD: Third Grade Math)
Effect	0.478***	0.111**	0.232**
	(0.014)	(0.034)	(0.071)
Slope Before Cutoff	0.004***	-0.002	-0.003*
	(0.001)	(0.001)	(0.002)
Change in Slope	-0.004***	0.001	0.002
	(0.001)	(0.002)	(0.002)

Note: This table reports the estimates of the discontinuities in the probability of waiting to enter kindergarten and in third-grade math test scores (measured in standard deviations). Regression discontinuity estimates are weighted liner regressions with 30-day bandwidth around each cutoff and rectangular kernel. Standard errors allow for arbitrary variance-covariance structure within schools, but two-way clustering by birthday changes very little. The sample for the November cutoff comes from 15,066 students who meet the following criteria: entering kindergarten in Michigan public schools in the 2013-14 or 2014-15 school years; turning five within thirty days of November 1, 2013; and taking state math exams in third grade. The sample for the December cutoff comes from 14,873 students who meet the former criteria, but who turn five within thirty days of December 1, 2013. $^{+}p < 0.1$, $*p < 0.05$, $**p < 0.01$, $***p < 0.001$.

1.4.2 Positive Selection on Gains: Among Higher-Income Families Eager Compliers Benefit More than Reluctant Compliers

To test for selection on gains, I compare the LATEs estimated at the November 1 and December 1 cutoffs. Intuitively, comparing effect of waiting on students who are induced to wait by different incentives reveals the nature of selection on gains. The November 1 LATE represents the effect on children those with $U_W \in [0.18, 0.40]$. These children are more willing to wait than reluctant compliers at December 1 who have $U_W \in [0.52, 1.00]$. The larger effect for eager compliers ($0.32\ \sigma$ vs $0.23\ \sigma$) suggests some measure of selection on gains. This subsection tests that intuitive comparison and finds evidence that only higher-income students select on gains.

I want to estimate the difference in treatment effects between eager and reluctant compliers and test whether it is different from zero. To do so I define the heterogeneity statistic:

$$\Delta_{ec,rc} = \tau_{ec} - \tau_{rc}$$

I test the null hypothesis that $\Delta_{ec,rc} = 0$ by simultaneously estimating both effects with GMM.[23] Because τ_{ec} and τ_{rc} are estimated at different values of the running variable, the parallel trends assumption is necessary to attribute the difference to selection on gains. Appendix A.5.2 explain a test of the parallel trends assumption suggesting that parallel trends hold in this setting.

When estimated in the full sample, the difference, $\Delta_{ec,rc} = 0.097$ is economically large, but it is statistically insignificant; however, I find strong evidence for selection on gains from subgroup specific effects. I compare $\hat{\tau}_{ec}(x)$ and $\hat{\tau}_{rc}(x)$ for low-income, high-income, Black, white, female, and male students using the same approach. Table 1.2 reports the results from these comparisons which include all observations within 90 days of either cutoff to remedy the power limitations of subgroup effects.

Some but not all groups select on gains. For example, in the low-income group waiting raises math scores for both types of compliers equally (0.15σ for eager compliers and 0.17σ for reluctant compliers), but in the higher-income group, the two effects are economically and statistically different (0.62σ for eager compliers and 0.15σ for reluctant compliers, p=0.043). These cross-group differences suggest that the failure to reject homogeneity in the full sample resulted from only half of the population selecting on gains. To my knowledge this is the first evidence to document selection on gains into waiting to enter kindergarten. The evidence suggests that strategic selection can benefit families especially when the children who gain the least from waiting are allowed to start kindergarten at age five despite recommendations (i.e., to enter early).

1.4.3 Negative Selection in Levels: Eager Compliers Score Lower than Reluctant Compliers

To test for selection in levels I compare the third-grade math scores of eager and reluctant compliers who both start kindergarten at five. As neither group waited, any differences must stem from the fundamental differences between eager compliers and reluctant compliers, recall that I call these differences selection in levels. Whereas estimating selection on gains required

[23] I use the outcome regression moments restricting the sample to the observations with positive weights under the rectangular kernel. Let $E[\tilde{Z}_{ik}(Y_i - \tau_1 D_i - \beta_1 \tilde{X}_{i1} - v_{i1})] = \mathbb{E}[\tilde{Z}_{ik}(Y_i - \tau_2 D_i - \beta_2 \tilde{X}_{i2} - v_{i2})] = 0$ where $\tilde{X}_1 = (r * (1 - z_2), r * z_1 * (1 - z_2), z_2)$ and $\tilde{X}_1 = ((r - 30) * (1 - z_1), r * z_2 * (1 - z_1), z_1)$ and with $\tilde{Z}_{ik} = (\tilde{X}_k, z_k)$ for instruments for the kth equation. I do this because it generates equivalent tau_k to 2SLS but allows me to test the equality of coefficients.

Table 1.2: Eager Compliers Benefit More than Reluctant Compliers in Some Subgroups

	Eager Complier Effect (τ_{ec})	Reluctant Complier Effect (τ_{rc})	Difference (Δ)
All Students	0.299**	0.218***	0.081
N=49,568	(0.112)	(0.058)	[p=0.431]
Low SES	0.148	0.171*	-0.023
N=28,129	(0.107)	(0.068)	[p=0.823]
Higher SES	0.618*	0.154$^+$	0.465
N=21,439	(0.248)	(0.085)	[p=0.043]
Black	0.273	0.164$^+$	0.108
N=11,197	(0.191)	(0.094)	[p=0.533]
White	0.337**	0.210**	0.166
N=31,946	(0.139)	(0.072)	[p=0.200]
Girls	0.433*	0.229**	0.204
N=23,994	(0.174)	(0.072)	[p=0.213]
Boys	0.196	0.206	-0.010
N=25,530	(0.144)	(0.093)	[p=0.939]

Note: This table compares estimates of the effect of waiting to enter kindergarten on third-grade math test scores (measured in standard deviations) for eager and reluctant compliers in different subgroups. Regression discontinuity estimates are local liner regressions with a rectangular kernel. Standard errors allow for arbitrary variance-covariance structure within schools. The sample includes students who meet the following criteria: entering kindergarten in Michigan public schools in the 2013-14 or 2014-15 school years; turning five within 90 days of either cutoff; and taking state math exams in third grade. Hypothesis tests are two sided tests of the equality of the effects at the two cutoffs estimated simultaneously by GMM. $^+p < 0.1$, $*p < 0.05$, $**p < 0.01$, $***p < 0.001$.

an additional assumption to extrapolate over birthdays, selection in levels is identified at the November 1 cutoff with no additional assumptions. I find strong evidence of negative selection in levels on average and among most subgroups but cannot reject the null of no selection among low-income students.

I want to estimate the difference in scores between eager and reluctant compliers who start kindergarten at five. To do so I define the selection statistic:

$$\mathcal{B}_{ec,rc}(r_{Nov}) = \mu_{0,ec}(r_{Nov}) - \mu_{0,rc}(r_{Nov})$$

I test the null hypothesis of no selection in levels $\mathcal{B}_{ec,rc} \equiv \mu_{0,ec} - \mu_{0,rc} = 0$[24] by applying the results from Imbens and Rubin (1997) and Abadie (2002) to identify the expected outcomes of compliers (See Appendix A.4 for details). As such my test is intuitively similar to testing whether never takers have expected outcomes equal to the control complier mean. Equivalent

[24]Suppressing the dependence on the birthday. As an aside, I call this difference $\mathcal{B}$ because it is analogous to the "bias" term in Cattaneo et al. (2020)

tests for fuzzy RD settings are proposed in both Bertanha and Imbens (2019) and Black et al. (2022).[25]

The estimated selection statistic is large and statistically significant in the full sample and for most subgroups. Table 1.3 reports the results using a nonparametric block bootstrap by school for inference. In the full sample, $\mathcal{B}_{ec,rc} = -0.42$, and I reject the null hypothesis of no selection at a $p = 0.004$. While the differences are negative and economically meaningful for all groups, there are stark difference in estimate size and statistical significance by racial and socioeconomic subgroups. For example, among low-income students the difference between eager and reluctant compliers' third grade test scores is only 0.12σ ($p = 0.289$), but among higher-income students the difference is over 0.66σ ($p = 0.002$). The differences are larger among white students compared to black students, but are similar between boys and girls. Substantively, these findings mean that reluctant compliers outperform eager compliers when both groups start at five. In other words, students are negatively selecting in levels into waiting. In addition to showing the economic nature of selection, this finding demonstrates that children who enter early tend to be higher achieving and also suggests that early entrants are much more positively selected than OLS estimates suggest.[26].

Taken together with the results about effect heterogeneity, we find a compelling story of strategic selection and comparative advantage. Reluctant compliers opt out of waiting because they will perform well even if they start at five and will gain less from waiting. This narrative is not without caveats, however. Lower-income families may be less negatively selected in levels and on average are not selecting on gains. Furthermore, these results do not inform us about the nature of selection into for always takers.

1.5. MTE-Framework Estimates of Always Takers Selection in Levels and Selection on Gains

Knowing that there is negative selection in levels and positive selection on gains between complier groups only answers half of the question. Do always takers are select in the same way? This section begins by reinterpreting the RD results in a marginal treatment effects (MTE) framework to show why the nature of selection into redshirting is critical for identifying

[25] Bertanha and Imbens (2019) note that a discontinuity of $\mathbb{E}[Y_i|D_i = 0, r]$ at the RD cutoff violates their strong "external validity" assumption, meaning that the complier estimated effects may not generalize to always takers or never takers. I employ the test to measure selection, a possibility noted in Black et al. (2022) which notes that rejecting the null "constitutes evidence of either selection or violation of the exclusion restriction." This test can also be framed as a special case of a more general test proposed in (Mogstad et al., 2018).

[26] Such as Bassok and Reardon (2013); Fortner and Jenkins (2017). This is because OLS estimates compare early entrants to a mix of students that include would-be early entrants assigned their ideal entry preference, attenuating the difference

Table 1.3: Eager Compliers Who Start Kindergarten at Five Have Lower Scores than Reluctant Compliers

	Eager Complier Untreated Outcomes ($\mu_{0,ec}$)	Reluctant Complier Untreated Outcomes ($\mu_{0,rc}$)	Difference ($\mathcal{B}_{ec,rc}$)
All Students	-0.431	0.002	-0.432***
N=49,568	(0.088)	(0.034)	(0.115)
Low SES	-0.489	-0.370	-0.120
N=28,129	(0.085)	(0.037)	(0.113)
Higher SES	-0.228	0.438	-0.666**
N=21,439	(0.181)	(0.040)	(0.241)
Black	-0.890	-0.621	-0.270
N=11,197	(0.159)	(0.056)	(0.206)
White	-0.310	0.253	-0.563***
N=31,946	(0.106)	(0.037)	(0.137)
Girls	-0.514	-0.050	-0.464**
N=23,994	(0.144)	(0.041)	(0.178)
Boys	-0.368	0.062	-0.429**
N=25,530	(0.108)	(0.049)	(0.146)

Note: This table compares the expected outcomes of students from different unobserved groups. The top panel shows the results at the November 1 cutoff, and the bottom panel shows those at the December 1 cutoff. Expected complier outcomes are calculated using the procedures from Imbens and Rubin (1997) and Abadie (2002). The final column reports the differences between groups. The sample for the November cutoff come from 15,066 students who meet the following criteria: entering kindergarten in Michigan public schools in the 2013-14 or 2014-15 school years; turning five between August 1 and December 1, 2013; and taking state math exams in third grade. The sample for the December cutoff come from 14,873 students who meet the former criteria, but who turn five within thirty days of December 1, 2013. Nonparametric block bootstrapped standard errors for estimated means and differences are given in parentheses, blocking by school with 1000 replications. $^{+}p < 0.1$, $*p < 0.05$, $**p < 0.01$, $***p < 0.001$.

τ_{at}. Then it and demonstrates that always takers are negatively selected relative to eager compliers, and uses that fact to estimate the effect of waiting to enter on always takers.

1.5.1 Mapping the RD Results into an MTE Framework

Recasting the results from Section 4 in a MTE framework shows how the scores at the cutoff identify averages of the marginal treatment response and marginal treatment effect curves, $m_0(u, r_{Nov})$, $m_1(u, r_{Nov})$, and $\tau_{MTE}(u, r_{Nov})$. Figure 1.3 shows these connections graphically in three panels. The top two panels of Figure 1.3 map the average scores identified by the RD to $\mu_0(u, r_{Nov})$,$\mu_1(u, r_{Nov})$. The top left panel displays average third-grade math scores around the November 1 cutoff separately by whether students waited to enter. The regression lines identify four limits at November 1. The top right panel plots the implied average test scores identified by these limits as the average values of m_0 and m_1 over each group's reluctance to wait, U_W. For example the RD identifies $\mu_{1,at}(r_{Nov}) = 0.21\sigma$. Since 18% of students redshirt at the cutoff, $\mu_{1,at} = \mathbb{E}\left[m_1(u, r_{Nov})|u \in [0.00, 0.18]\right]$, so I plot a line at 0.21 over $U_W \in [0.00.018]$. The other three line segments report the averages for the other groups: always takers and eager compliers who wait, eager and reluctant compliers who start, and reluctant compliers who start.[27] Each is an average of the relevant marginal outcome function $\mu_1(u, r_{Nov})$ or $\mu_0(u, r_{Nov})$.

This interpretation of the results also identifies portions of the MTE curve, $\tau_{MTE}(u)$ and MTR curve measuring selection in levels, $m_0(u, r_{Nov1})$. This is reflected in the bottom panel of Figure 1.3 which plots the average outcomes over U_W after recovering the complier means. This presentation of the results makes τ_{ec}, τ_{rc}, and $\mathcal{B}_{ec,rc}$ visible. τ_{ec} reflects the average values of τ_{MTE} over $U_W \in [0.18, 0.40]$ and $\mathcal{B}_{ec,rc}$ the difference in test scores after starting kindergarten at five between $U_W \in [0.18, 0.40]$ and $U_W \in [0.40, 1.00]$.

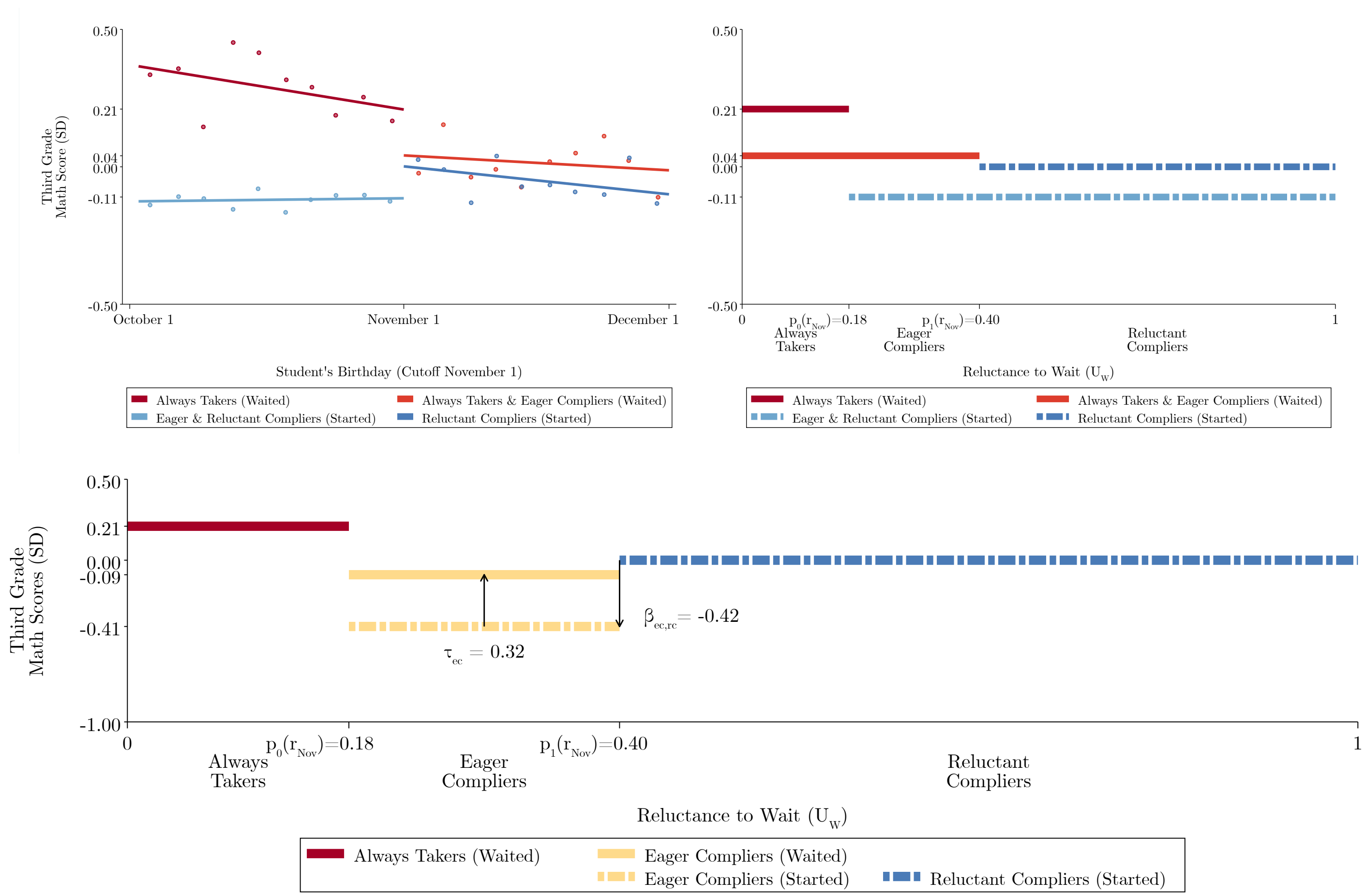
Mapping the RD results into an MTE Framework also makes it visually apparent that identifying the causal effect on groups other than eager compliers requires additional information or assumptions. I need more information because the November cutoff has no information about the scores of reluctant compliers if they wait to start kindergarten or about always takers if they start at five. Fortunately, the December cutoff gives me information about the scores of reluctant compliers who wait, but I will need to make an assumption about the test score of always takers had they entered without waiting. Although it is common to make *a priori* assumptions about the untreated outcomes of always takers, the following section empirically tests the nature of selection in levels instead. Identifying how always takers select in levels will inform what assumptions are reasonable to identify the treatment effects on always takers and whether or not they also positively select on gains.

1.5.2 Negative Selection in Levels: If Not for Waiting, Always Takers Would Score Lower than Eager Compliers

This subsection explores two sources of information that tell us more about the selection in levels: early-elementary-school outcomes and how average achievement changes as the share of redshirters increases. Each approach demonstrates that always takers are negatively selected on average and within most subgroups.

[27] Note the overlap occurs because eager compliers ($U \in [0.18, 0.40]$) show up in two groups both as treated and untreated—this is what identifies τ_{ec}.

Figure 1.3: Reduced From Shows Effects, Selection, and Heterogeneity



Note: This figure displays average standardized third-grade math scores over two different dimensions to illustrate the mapping between the two. In both figures the sample is made up of 15,066 students who entered kindergarten in a Michigan public school, who have birthdays within thirty days of November 1 in 2013, and for whom I observe third-grade math scores. The panel on the top left is an RD plot of outcomes separated by treatment status, similar to Panel (c) in Figure 1.2. Instead of showing average outcomes for all students at a given index value, this graph plots those average outcomes separately for treated students (who waited to enter kindergarten) and untreated students (who did not wait). Average test scores are reported by three-day bins for students who did and did not wait, and regression lines are displayed for each subgroup. Lines of best fit are also displayed to visualize how the limits of $\mathbb{E}[Y_i|d, z_1]$ are estimated. The panel on the top right maps the expected performance of each group as identified by those limits at November 1 onto the support of their unobserved cost U_W. In this graph the average outcomes of treated groups are displayed in solid lines, and the average outcomes untreated groups are displayed with dashed lines. The bottom panel illustrates the local average treatment effect (LATE) and selection between eager and reluctant compliers implied by these means at the November 1 cutoff.

1.5.2.1 Always Takers Receive More Special Education Services and Testing Accommodations

This section estimates the differences in outcomes between always takers and eager compliers to assess whether the treatment effects implied by different assumptions about selection in levels are plausible. Under the null hypothesis of no selection between always takers and eager compliers, differences in outcomes between always takers and eager compliers captures the difference in effect heterogeneity between these groups.[28] Results that have counterintuitive signs or magnitudes suggest that the null of no selection in levels is implausible. I estimate these differences using the 10 cohorts of children who turned five between 2002 and 2012 to increase power. Similar patterns are visible but imprecise using the differences between always takers and eager compliers in 2013 (see Appendix Table A.7).

The results show large differences between groups and produce perplexing results under no selection or positive selection in levels. Table 1.4 reports the average outcomes of always takers, compliers who wait, and compliers who enter without waiting. Despite having the highest third-grade math scores (see Figure 1.3), always takers have higher rates of testing accommodation, non-testing,[29] and special education service receipt.[30] These results are also consistent across demographic subgroups (see Appendix Table A.8) and suggest negative selection in levels because generally it is low achieving students who receive these accommodations.

In addition to the fact that always takes have worse level outcomes than eager compliers, only large and unintuitively signed treatment effects on always takers could rationalize positive selection in levels in these outcomes. Positive selection in levels would suggest that always takers who start kindergarten at age five should experience lower rates of accommodation, non-testing, and special education than compliers who start at age five. But given the observed outcomes for always takers who wait would then imply that waiting to enter kindergarten increases the prevalence of these outcomes in always takers by about 100%; whereas for compliers, waiting to enter kindergarten *reduces* the likelihood of missing a test and of receiving special education by around 20%.[31] These differences in sign are statistically

[28]Since these outcomes occur after waiting decisions, the comparison combines information about the baseline differences across groups (selection in levels) and the effect heterogeneity, so there exists a hypothetical treatment effect on always takers that can rationalize any assumption about selection in the counterfactual outcome.

[29]Note the affected portion of students is too small to change the cross-group rankings at the cutoff: Even if all these students and students who did not test scored a half standard deviation below average $\mu_{at,1}$ would still be around 0.13 σ.

[30]This does not seem to be driven by less sever diagnoses: always takers are the most likely to be diagnosed with cognitive impairment, emotional impairment, language impairment, and early childhood developmental delay than any other group (see Appendix Table A.6

[31]Finding that waiting reduces special education assignment is not unique to my sample. See also Elder

significant and constitute strong evidence that always takers are *not* positively selected in levels relative to compliers. Alternatively, under negative selection in levels, these are exactly the relationships we would expect to see.

There are two possible concerns with this suggestive evidence about selection in levels. First, it may not be true that the causal effects of waiting to has the same sign for eager compliers and always takers. For example, if early interventions are valuable and special needs are hard to detect, waiting a year to enter kindergarten could possibly increase special education diagnoses and accommodated testing for some groups. Second, the relationship between baseline achievement levels and the early elementary school outcomes may not be the same for always takers as for compliers. For example, for compliers special education or accommodated testing might be correlated with lower baseline achievement, but for always takers they might be correlated with having very pushy parents (and possibly higher baseline scores). Although these concerns likely are not large enough to suggest positive selection in levels on average, the following subsection provides additional evidence of negative selection in levels from third-grade achievement.

1.5.2.2 "Marginal" Always Takers Score Lower than Same-Aged Compliers

My second approach directly tests for selection in levels using variation in average achievement as birthdays approach the cutoff. This subsection sets out the intuition for this test and reports the results. As birthdates approach the cutoff more families choose to wait, changing the composition of students in each group. In this section I show that always takers who are induced to wait by being closer to the cutoff have lower test scores than compliers when they start kindergarten at five.

The comparison of interest exploits variation in average achievement as birthdays approach the cutoff. In a sharp RD design where all units to the left of the cutoff are untreated, any change in average untreated outcomes as the running variable r increases captures the direct effect r (or if r and x are not independent any changes conditional on x). This intuition changes in a fuzzy RD because as r increases, the share of always takers increases—the upward sloping lines in the first stage (Panel (a) Figure 1.2). In a fuzzy RD, the changing share of always takers implies that changes in average untreated outcomes to the left of the cutoff reflect both the direct effect of r as well as the changing composition of the still untreated group.

Because changing scores reflect the direct effect of r as well as the changing composition, if a researcher knows something about the direct effect, the change in scores can reflect

(2010); Evans et al. (2010); Layton et al. (2018); Dee and Sievertsen (2018); Sharpiro (2020) for other examples.

Table 1.4: Early Elementary School Outcomes Suggest that Always Takers Are Negatively Selected

	Mean	Always Takers	Compliers Wait	Compliers Enter	Difference	LATE
Sample Shares		20.3%	77.3%			
Special Education (Kindergarten)	0.099	0.170	0.069	0.086	0.101***	-0.018***
	(0.001)	(0.006)	(0.003)	(0.002)	(0.008)	(0.004)
Special Education (Third Grade)	0.147	0.210	0.114	0.138	0.096***	-0.025***
	(0.001)	(0.007)	(0.004)	(0.003)	(0.009)	(0.005)
No Third Grade Math Score	0.110	0.154	0.083	0.116	0.071***	-0.033***
	(0.001)	(0.006)	(0.003)	(0.003)	(0.008)	(0.004)
Accommodated Test in Third Grade	0.008	0.013	0.006	0.006	0.008***	0.000
	(0.000)	(0.002)	(0.001)	(0.001)	(0.002)	(0.001)

Note: This sample shows the average elementary school outcomes for always takers, eager compliers, and reluctant compliers. The sample is comprised of students who started kindergarten in Michigan public schools in the fall of 2002-2012 and turned five within thirty days of December 1. Note that in this table I do not restrict to students who took third-grade math tests. Block bootstrapped standard errors for the estimated means and differences are given in parentheses, blocking by school with 1000 replications. Note that the shares do not sum to 100% because in some of the earlier years there were loopholes to the requirements allowing some never takers to still start kindergarten at age five. Those students are not dropped in the analysis.

the selection in levels revealed by the compositional change. For example, in the case of elementary school achievement, it is well-known that students with later birthdays score worse on tests (conditional on waiting). This is called an "age at test" effect and is visible in the reduced form (Panel (b) of Figure 1.2). If we observed the scores of children who started kindergarten at five increasing instead of decreasing in some range of r it must be due to compositional changes and would reveal negative selection in levels for the range of $p_0(r)$ where test scores were increasing. Interestingly, this is exactly what Figure 1.3 revealed.

Figure 1.3 showed that average test scores of students who start are increasing, implying that always takers are negatively selected. Between October 1 and November 1, there is a large increase in the probability of waiting. Inducing lower-achieving students to wait is increasing the average scores in the remaining group who starts at five. A linear regression of test scores on this window (with controls for lower-income, black, and female) reports a positive slope and rejects a two sided test of the slope being zero at $p = 0.046$ level. Because this result is somewhat sensitive to bandwidth and covariate inclusion, I explore an alternative test in Appendix A.4 based on changes in slope of test scores over birthdates that yields the same results and demonstrates that always takers are negatively selected in levels on average and among higher-income, white, and female subgroups.

Together with the other results we have repeated evidence that always takers are negatively selected. The main limitation of this test is that although we know the direction of selection, we cannot measure the magnitude without knowing the true causal effect of r. This test also cannot determine whether all always takers are negatively selected, or just the marginal ones. However, combined with the evidence that eager compliers have fewer accommodations in elementary school, these results suggest that $\mu_{0,at} < \mu_{0,ec}$, both on average and within the majority of the subgroups. This finding is new compared to a large descriptive literature that has suggested that students who select into waiting are positively selected since they tend to come from affluent, educated, white families. Instead it is consistent with a comparative advantage story in which children who are "not ready" for kindergarten are most likely to wait.

1.5.3 Selection on Gains: Always Takers Benefit More Than Compliers from Waiting (Especially Higher-Income Always Takers)

This subsection leverages the new information about selection in levels to identify the treatment effect on always takers, documenting large gains. Despite the large average gains to always takers, not all always takers are positively selected on gains into waiting. For example, among black and lower-income children, the treatment effects are small and vary little across the unobserved groups.

Because identifying the effect of waiting on always takes requires an assumption about the nature of selection in levels, I assume that achievement without waiting evolves linearly over unobservables U_W. Assumptions like this are ubiquitous whenever researchers are trying to test for selection on gains. For example, control function methods usually rely on a functional form assumption (see the examples in Kline and Walters, 2019)–often linearity (as in Kline and Walters, 2016; Walters, 2018, for two recent examples). Functional form assumptions are also common when using an MTE framework (Heckman and Vytlacil, 2005; Carneiro et al., 2011; Brinch et al., 2017; Kowalski, 2022b), but shape restrictions are becoming the new frontier (Mogstad et al., 2018; Kowalski, 2022a). While my results about selection are robust to weaker shape restrictions as well (see Appendix A.5), I prefer making a functional form assumption to obtain point estimates rather than bounds—and point estimates are necessary to quantifying the magnitudes of effects that the selection patterns have on efficiency and equity. Rather than make a functional form assumption on both m_0 and m_1 I only make an assumption about the MTR reflecting selection in levels, m_0.[32] Specifically, I assume

$$m_0(u, r, x) \equiv \alpha_0(x) + \beta_0(x)u_i(x) + v_i \text{ with } E[v_i|u, x] = 0$$

This assumption is consistent with the facts that always takers are negatively selected relative to eager compliers and that eager compliers are negatively selected relative to reluctant compliers. I measure the effects of waiting on always takers by comparing their average scores after waiting with the scores m_0 imply had they started at five. After estimating the effects, I test the null of zero, $\tau_{at} = \mu_{1_at} - \mu_{0,at} = 0$, and the null of homogeneity between always takers and eager compliers, $\tau_{at} = \tau_{ec}$.

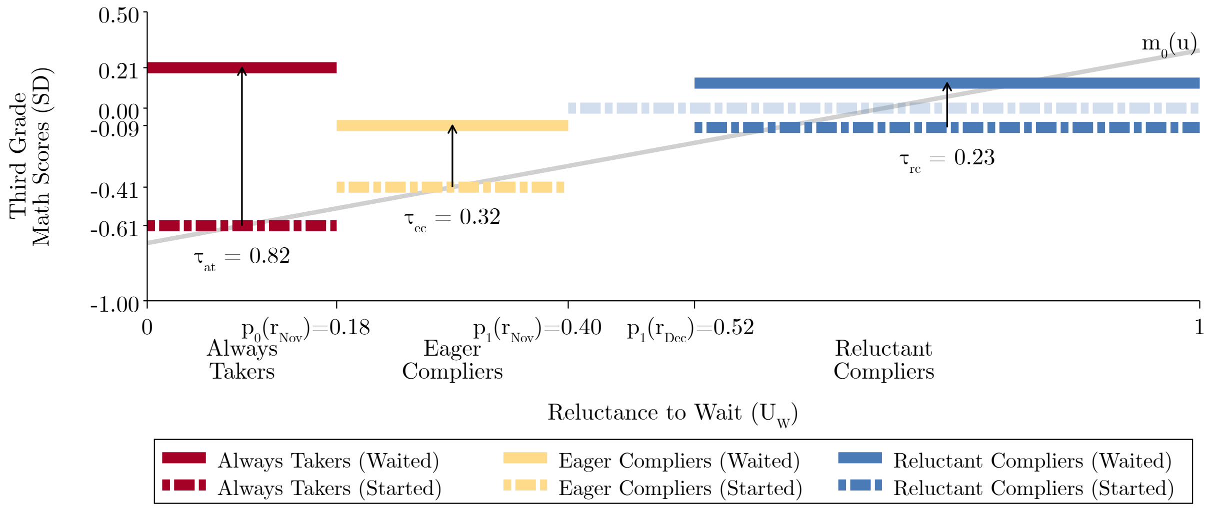
I find that always takers at the November 1 cutoff benefit enormously from waiting to enter kindergarten, suggesting more positive selection on gains. Figure 1.4 combines earlier results with the effects on always takers to illustrate the selection on gains. This figure adds the scores of reluctant compliers at December 1 over their values of U_W. Then by plotting the showing the linear $m_0(u)$ over the support of U_W and the implied average scores of always takers had they not waited ($\mu_{0,at} = -0.63$), Figure 1.4 visualizes the gains to always takers, $\hat{\tau}_{at} = 0.84\sigma$. This large effect is significantly different from zero and from τ_{ec} at a $p = 0.001$ level (Table 1.5 has all standard errors and p-values as well as results from a bounding assumption presented in detail in Appendix A.5. Taken together it is visually apparent that the three shrink grow monotonically as the reluctance to wait increases. In other words it clearly displays positive selection on gains.

In addition to showing that always takers benefit from waiting on average, I explore heterogeneity across different demographic groups by splitting the sample on observable characteristics x. A key advantage to nonparametrically estimating heterogeneity by observables is that it allows for different patterns of selection on gains across groups. As Cornelissen et al. (2016) point out, splitting the sample by x and estimating effects separately is the ideal way to estimate marginal treatment effects when there is strong enough support. Splitting the sample also removes the need to make common assumptions of additive separability between x and u, as these assumptions imply that all groups must select uniformly on gains.

After splitting the sample, I find evidence that only certain types of always takers are positively selected on gains. Table 1.5 shows that not all groups positively select on gains. For example, the treatment effect on always takers is larger than the effect on eager compliers

[32] Although not as weak as a shape restriction, assuming that $m_0(u)$ is linear in u is actually weaker than any of the other functional form assumptions mentioned above. Because I have identification from two different cutoffs, I only need make one assumption to identify the effects for each group rather than assuming linearity in both treated and untreated outcomes (as in Brinch et al., 2017; Kowalski, 2022b) or that treatment effects are linear in the unobservable (as in Kline and Walters, 2016; Walters, 2018). This has an added advantage of leaving shape of the treatment effect unrestricted.

Figure 1.4: Late Entrants Benefit Enormously from Waiting to Entering Kindergarten



Note: This figure graphically illustrates the average treatment effect for always takers at the cutoff τ_{at}, which is recoverable from the limits of student achievement at the cutoff and the ancillary assumption of linearity in untreated outcomes. The sample is made up of 15,066 students who entered kindergarten in a Michigan public school, who have birthdays within thirty days of November 1 in 2013, and for whom I observe third-grade math scores. Average outcomes for treated and untreated compliers are backed out of observed data and choice probabilities. For block-bootstrapped standard errors see Table 1.5.

(which in turn is larger than the effect on reluctant compliers) among the white children and children from higher-income families, but there are not significant differences for other children. There is strong evidence of selection on gains among girls and boys. Table 1.5 also reports the full results p-values of tests for heterogeneity under weaker assumptions of monotonicity rather than linearity in untreated outcomes with somewhat larger p-values but qualitatively similar results.

These results would be consistent with the explanation that narrative that regardless of demographics, parents tend to be aware of which children are at risk of underperforming in kindergarten. That said, the varying slopes of the selection in levels may suggest that certain parents respond more strongly to the possibility of poor performance: high-income parents seem to be especially responsive. Simultaneously, the lack of selection on gains suggests that identifying children who are at risk for underperforming is not sufficient to enable them to succeed, they need to actually benefit from those decisions to wait. I explore possible mechanisms and explanations for this in the subsection that follows.

Table 1.5: Only Some Demographic Groups Are Positively Selecting on Gains

	Reluctant Complier Effect (τ_{rc})	Eager Complier Effect (τ_{cc})	Always Taker Effect (τ_{at})	Difference $\tau_{at} - \tau_{ec}$	Always Taker Bound ($\tilde{\tau}_a t$)	Test Statistic
All Students	0.232	0.325	0.838	0.505	0.631	1
N=116,506	(0.071)	(0.144)	(0.197)	[$p = 0.000$]	(0.133)	[$p = 0.011$]
Lower-Income	0.171	0.132	0.170	0.0376	0.122	0
N=53,568	(0.070)	(0.102)	(0.145)	[$p = 0.732$]	(0.101)	-
Higher-Income	0.153	0.587	1.14	0.552	0.800	1
N=42,938	(0.856)	(0.215)	(0.285)	[$p = 0.006$]	(0.176)	[$p = 0.094$]
White	0.212	0.385	0.990	0.605	0.694	1
N=63,307	(0.073)	(0.129)	(0.162)	[$p = 0.000$]	(0.098)	[$p = 0.002$]
Black	0.170	0.258	0.439	0.181	0.338	1
N=20,995	(0.096)	(0.184)	(0.255)	[$p = 0.325$]	(0.197)	[$p = 0.343$]
Girls	0.229	0.4434	1.04	0.573	0.827	1
N=47,435	(0.073)	(0.169)	(0.177)	[$p = 0.000$]	(0.154)	[$p = 0.006$]
Boys	0.208	0.183	0.902	0.719	0.671	1
N=48,993	(0.090)	(0.137)	(0.189)	[$p = 0.000$]	(0.114)	[$p = 0.002$]

Note: This table reports estimated treatment effects and tests for heterogeneity for different subgroups. Effects are estimated as discussed in the text. For the case of linearity, the tests for heterogeneity is a two-sided test on the null: $\tau_{at} = \tau_{ec}$. For the case of the monotonicity in untreated outcomes, it is a one-sided test of the null that $\tilde{\tau}_a g \leq \tau_{ec}$, i.e., it rejects the null if the estimated bound for always takers excludes the effect on eager complier effect. Block bootstrapped standard errors for estimated means and differences are given in parentheses blocking by school with 1000 replications. The numbers in brackets below the test statistics are the fraction of bootstrapped replications in which the lower bound on the treatment effect for always takers is greater than the effect on eager compliers.

1.6. Discussion, Extensions, and Policy Implications

The previous sections answered the positive question of how children are selecting into waiting and showed negative selection in levels and positive selection on gains. This section turns to the normative questions about achievement and learning gaps. First, it measures the average effects of waiting to enter kindergarten and explores investments in the year of waiting as a possible mechanism through which the effect heterogeneity may operate. Then, it returns to the theoretical framework and uses it to formally describe equity and efficiency. Finally, it combines all of the results to answer the questions of equity and efficiency. Does allowing strategic selection increase or decrease achievement? What does selection imply for learning gaps across different types of students? And what might these results mean for improving policy?

1.6.1 Exploring Inequities: The Case for High Quality Pre-School

Before turning to equity and efficiency, this subsection explores the mechanisms through which the observed patterns of selection may operate. The absence of selection on gains I find is not the typical reverse-Roy sorting where low-income parents fail to invest in the children who would benefit the most;[33] rather, the results in Table 1.5 suggest that waiting does not make as much of a difference for the low-income students. In other words there are not large gains to sort on. This section formally explores these differences and possible drivers. I find that low-income students benefit much less on average from waiting a year to enter kindergarten and show suggestive evidence that this is the result of heterogeneity in the preschool investments made in the intervening year (as also suggested by the tapering of gains over time, see Elder and Lubotsky, 2009).

I want to estimate the average treatment effect $\tau_{ATE}(x)$ for different subgroups from the estimated subgroup effects $\tau_g(x)$. Considering the effects estimated at the November 1 cutoff:

$$\tau_{ATE}(x) \equiv \mathbb{E}[\tau_i|X = x] = \tau_{at}(x)p_0(r_{Nov}, x)+\tau_{ec}(x)(p_1(, r_{Nov}, x)-p_0(r_{Nov}, x))+\tau_{rc}(x)(1-p_1(r_{Nov}))$$

I then test the null of no difference between children with different characteristics $\tau_{ATE}(x = 1) = \tau_{ATE}(x = 0)$. Because the sample is made up of always takers, eager compliers, and reluctant compliers, then the population average treatment effect should be a weighted average of the groups' respective effects.

Note that estimating τ_{ATE} this way requires extrapolating from r_{Dec} to r_{Nov}. This extrapolation is necessary because τ_{rc} is only identified over $U_W \in [0.52, 1.00]$; children with $U_W \in [0.40, 0.52]$ are reluctant compliers at November 1, but eager compliers at December 1. For this exercise I assume that $\mathbb{E}[\tau_{MTE}|u \in [p_1(r_{Nov}), p_1(r_{Dec})] = \tau_{rc}$ to extrapolate (in addition to the "parallel trends" assumption). This assumption is fairly weak since the share of children who would enter if their birthday was right after November 1 but would wait right before December 1 is small (about 12%), but Appendix Table A.9 shows robustness to other assumptions about the effect on these students.

I find that the ATE is smaller for low-income students than for higher-income students. The first column of Table 1.6 reports the results. The average effect on higher-income children in 0.48σ, whereas it is only 0.16σ for the low-income children. This difference is significant at the $p = 0.05$ level and means that on average higher-income children benefit three times more from waiting to enter kindergarten than do low-income students. The difference between the average treatment effect on male and female students is inconsequential, and the difference

[33]For example in Kline and Walters (2016); Cornelissen et al. (2018) poorer children with a large u have the biggest gains.

between black and white students is large but falls just shy of statistical significance at a $p = 0.10$ level.

Because caring for children is costly, one possible explanation for the differences in average effect is that different families make different human-capital investments during the year of waiting. This would be consistent with evidence that the year of waiting seems to drive achievement increases (Elder and Lubotsky, 2009) and children from low-income families face more barriers to high quality preschool (see Shapiro et al., 2019, for a thorough review of results). On the other hand, it may be that families make similar investments in the year of waiting, but the different gains come from dynamic complementarities with earlier family decisions. For example, children whose parents read more to them before age four (or who had better nutrition or watched less television etc.) may learn more in the same preschool setting than children with fewer early-life investments. The evidence in early childhood dynamic complementarities is more scarce but research suggests they have large effects (Johnson and Jackson, 2019; Adhvaryu et al., 2020). These two mechanisms would require very different policy interventions to address the inequity in gains.

I compare these mechanisms, by examining families' investment decisions. Although the Michigan state administrative data does not have much information on early childhood programs, I do observe whether children participate in the Great Start Readiness Program (GSRP). The GSRP is a one-year a program for four-year-old children from single-parent households, lower-income families, or with special needs or other risk criteria. Children participate in GSRP the year before they enter kindergarten. To explore whether the mechanisms of differences in ATEs across groups, I compare the ATE for students who do and do not participate in the GSRP program. If the effects are not different it would be suggestive evidence that there are important differences in investments made during the year between deciding to wait and when the children begin kindergarten. Note that these exercises are descriptive because there is only exogenous variation in the decision to wait to enter kindergarten, not in participation in the GSRP.

Analyzing the GSRP yields suggestive evidence that the differences in ATEs across groups are mainly driven by different quality investments in the year before kindergarten. The main evidence for this claim is the fact that while the ATE for low- and higher-income students are very different in the full sample, the effects are almost identical on those who participate in GSRP. The second and third columns of Table 1.6 detail the results. For comparisons by economic status and by race, the ATE is statistically indistinguishable among students who did GSRP the year before kindergarten; in fact, the difference between the higher-income ATE and the low-income ATE is less than 0.02σ. On the other hand, among children who do not participate in GSRP, the differences in the effects of waiting are extreme: the large for higher-income children and white children dwarf the statistically insignificant effects on low-income and black children (both differences are statistically significant).

Table 1.6: Average Treatment Effects Are Larger for Higher-Income Families, Except Among Children in Public Pre-K

	Average Effect (All)	Average Effect (GSRP)	Average Effect (No GSRP)
All Students	0.361***	0.396***	0.356***
	(0.082)	(0.116)	(0.112)
Higher-SES Students	0.481***	0.268	0.552***
	(0.128)	(0.172)	(0.162)
Low-SES Students	0.164**	0.284***	0.085
	(0.071)	(0.108)	(0.097)
Female	0.345***	0.433***	0.363***
	(0.094)	(0.130)	(0.112)
Male	0.369***	0.247*	0.409***
	(0.086)	(0.130)	(0.137)
White	0.412***	0.343***	0.473***
	(0.083)	(0.117)	(0.112)
Black	0.207*	0.435***	0.076
	(0.109)	(0.166)	(0.131)

Note: This Figure reports the average treatment effects of students by subgroup and preschool decisions. Average treatment effects are recovered by taking a weighted average of the effect on always takers, eager compliers, and reluctant compliers, using the sample proportions as weights. Standard errors are obtained by a nonparametric block bootstrap blocking on school with 1000 replications.

These results are suggestive of sharp differences in the quality of the human capital investment entailed in waiting a year to enter kindergarten. Because preschool enrollment decisions are not random, these results must be interpreted as differences in the average effects for students who choose to (or not to) participate in GSRP, and as low-income children are more likely to qualify, selection is likely not uniform across groups. The patterns suggest that differences in the effects of waiting are not driven dynamic complementarities. In the presence of such complementarities we would expect the higher-income children to benefit more from a given preschool program because they have had more intensive investments earlier on.

Together with what is already known about the importance of early-childhood investments, these results suggest that different returns to waiting to enter kindergarten do not stem from differential investment decisions. The cost to participate in high quality preschools seems important in these decisions. In addition to financial costs that may be barriers to many low-income families, other studies have found that less educated families are less likely to participate in publicly provided programs (see, for example, Felfe and Lalive, 2018; Kline and Walters, 2016). This distinction in mechanisms is important because the policy recommendation in the presence of dynamic complementarities (early interventions) is very different from the recommendation in the face of underinvestment (increase access to investment opportunities right before public school). Furthermore, this information about mechanisms suggests that the small gains among lower-income families do not stem from valuing gains less or from uninformed decision making, but from binding constraints that impede access to high-impact investments as a part of the "gift of time."

1.6.2 Measuring the Efficiency and Equity Implications of Strategic Selection

On its own identifying selection in levels and selection on gains is insufficient to determine the efficiency and equity implications of selection around recommendations. Not even the average treatment effects and heterogeneity by investment can do that. To explore the normative questions I define one allocation as being more efficient than another if two conditions are met: the allocation implements choices that are revealed preferred to families, *and* the allocation results in higher average test scores. Equity for students of type $X = x$ is measured in average differences in realized test scores $\mathbb{E}[Y|X = x] - \mathbb{E}[Y|X \neq x]$, so allocations can be compared by measuring the resulting change in achievement gaps. Appendix A.2 details these definitions, how the partially order allocations, and what types of strategic selection and specifications of social welfare they are robust to.

With these conceptualizations of equity and efficiency, I compare the welfare implications of different kindergarten policies. The first is the 2013 policy with its implemented rules about redshirting, early entry, and its empirical availability of prekindergarten. The second policy is to make the November 1 cutoff impose a requirement (on both sides) to eliminate strategic selection. For this policy, I estimate a naive counterfactual using the eager complier LATE and a more thorough counterfactual using the heterogeneity I estimate. The final policy is a counterfactual is one that that allows strategic selection but increases enrollment in prekindergarten programs among low-income populations.

The first two columns of Table 1.7 show how banning strategic selection would lower scores but shrink gaps. To estimate these effects I impose two assumptions: I assuming that the effects on reluctant compliers estimated at December 1 reflect the effects on all reluctant compliers (as when calculating the ATE) and that the effects on always takers estimated at November 1 reflect the effects on all always takers. Note that if marginal treatment effects are monotonic over U_W, these assumptions minimize the efficiency gains of strategic selection (relative to a policy with requirements). Currently the income-achievement gap is

Table 1.7: Allowing Strategic Selection Raises Scores But Widens Gaps

	Recommendation Baseline	Requirement Policy (MTE)	Requirement Policy (LATE)	Increase Low-SES Pre-K (MTE)
Average achievement:				
Higher-Income	0.404	0.310	0.497	0.404
Lower-Income	-0.384	-0.344	-0.349	-0.293
Gap:	0.788	0.654	0.836	0.693
Efficiency:				
Raises Scores	Baseline	No	Yes	Yes
Revealed Preferred	Baseline	No	No	Yes
Equity				
Shrinks Gap	Baseline	Yes	No	Yes

Note: This table shows the results of counterfactual policy simulations that explore the efficiency and equity implications of recommendations and requirements and the role for increased early childhood investments in promoting both objectives.

about 0.79 standard deviations. Banning redshirting and early entry reduces this gap by about 18%. The gap is closed in both directions. Scores among children from lower-income families increase because the gains to reluctant compliers more than compensate the losses to always takers (since there are many more reluctant compliers than always takers). But scores among children from higher-incomer families are lowered tremendously because the gains to reluctant compliers are small, and the losses to always takers are large. Also note that in addition to lowering average scores, the allowing strategic selection is revealed preferred, so the requirement policy imposes large costs on parents. These results highlight a real equity-efficiency tradeoff between requirements and recommendations.

A policy simulation using the LATE from eager compliers would get this efficiency equity implications completely backwards. The third column of results from Table 1.7 shows this result by assuming that the eager-complier LATE for higher- and lower-income families is the effect on all students. Because there is no selection on gains (positive or negative) among lower-income families, extrapolating using the LATE to always taker and reluctant compliers does not change much among that group. On the other hand, among higher-income families the eager-complier LATE largely overstates the benefits of a requirement to reluctant compliers and understates the losses to always takers. In net this means that using the LATE for policy analysis gives the wrong answer. It says that a requirement will raise scores for both groups, but because the gains are larger to higher-income families, it will widen gaps. This answer is wrong on both accounts and shows the importance of allowing for selection on gains especially for questions about allowing strategic noncompliance with treatment recommendations.

Finally, I show that increasing low-income children's participation in prekindergarten could raise average scores and shrink gaps. This simulation assumes that the average effect of waiting on children who participate in GSRP would generalize to those who do not participate in GSRP and assumes that all children benefit 0.07 standard deviations in third grade from participating in GSRP. These assumptions minimize the possible impact of this counterfactual because the direct benefits to preschool and the benefits of being in preschool while waiting are likely larger for children who are less likely to participate (Kline and Walters, 2016; Cornelissen et al., 2018; Felfe and Lalive, 2018). I find that if all lower-income children enrolled in public prekindergarten, it would shrink gaps by at least 12% and would increase average achievement. These results reinforce the importance of considering the investments made in the intervening year when considering equity and efficiency.

1.7. Conclusion

This paper proposed two main questions: (1) How are families strategically selecting around kindergarten recommendations? and (2) What are the implications of that selection for efficiency and equity? The purpose of answering these questions was to characterize the economics of this important human capital decision in order to inform the policy questions surrounding recommendations and requirements.

Comparing narratives about parents strategically manipulating educational systems with those about private information about child readiness, my results suggest that on average the second story comes closer to the truth. I find negative selection in levels (children who are more likely to wait would perform worse if they did not wait) and positive selection on gains (children who are more likely to wait experience larger score increases). This evidence is consistent with parents weighing the costs (of money, time, and opportunity) of waiting against the potential early-elementary-school benefits for their children and trying to have the children wait who would benefit the most from it. Indeed, this seems to be the case for both redshirters (who gain the most from waiting) and early entrants (who gain the least from waiting). In fact, allowing parents to use their private information to make these decisions is raising average test scores in the status quo.

The fact that children who are negatively selected are more likely to wait to enter kindergarten does not dismiss the equity concerns about strategic selection. In fact, my results show that these concerns are well founded: Despite being negatively selected on average, after waiting, always takers (including academic redshirters) are among the highest performing students. This pattern of selection on gains is concentrated among higher-income families, implying that in the status-quo allowing selection perpetuates racial- and income-based gaps in achievement.

The evidence also suggests that there are major structural barriers preventing disadvantaged groups from benefiting from waiting. Despite evidence that lower-income families are trying to redshirt the children who need it the most, they do not tend to benefit as much from waiting as their peers from higher-income families. My results suggest that this stems from unequal access to high-quality preschool programming. If the policy priority is to close gaps by improve the achievement of lower-income children, reducing barriers to high quality programming is a much better policy than completely banning selection around recommendations. Note, however, that expanding access to preschool is insufficient. Added availability must be accompanied by effective outreach. Otherwise the students who would benefit the most from preschool are the least likely to participate (Kline and Walters, 2016; Cornelissen et al., 2018; Sharpiro, 2020).

I conclude by exploring some promising avenues for future research. One mechanism that was difficult to explore was the role of participating in kindergarten for two years in mediating the effects I find. Because of data limitations I cannot distinguish between kindergarten repetition and formal developmental kindergarten programs in my sample period, so I cannot explore which children entered kindergarten intending to (or with the option value of being able to) repeat it the next year. Disentangling these pieces could be important research, especially because children from lower-income families are more likely to enroll in kindergarten twice (Dhuey et al., 2019).

Another important policy-relevant question that this paper does not answer is the peer effects from always takers as compared to compliers. The fact that waiting is good for individual children does not suggest that it is necessarily good for their peers. Some research indicates that compliers are more likely to be tracked into gifted and talented programs if they are assigned to wait. If spots in these programs are scarce, then noncompliance could have a negative externality on other students (i.e., redshirters might take slots from other children). On the other hand, research suggests that for a given class, having peers who wait to enter has positive spillovers onto other children (Bedard and Dhuey, 2012; Cascio and Schanzenbach, 2016; Peña, 2017). But does having a peer who is an always taker have a similar effect to having a peer who is a complier? If so, the always takers are providing a positive classroom externality. Estimating heterogeneity in peer effects would be a fascinating area of further study.

While this paper focuses on the economic nature of selection into waiting, it is entirely focused on short term outcomes. Applying my framework to research on longer-term outcomes would be important and policy relevant. I show that reluctant compliers are positively selected in levels compared to eager compliers, and that eager compliers are positively selected relative to always takers—but this finding is limited to third-grade test scores. In the same way that treatment effects on compliers have been shown to fade out over more extensive time horizons, the magnitude (or even direction) of selection need not be constant across outcomes measured at different lengths of time since treatment. This too seems like an important avenue to consider the effects of redshirting in the long run.

Finally, the stark differences in the nature of selection between higher- and low-income students highlights the importance of heterogeneity in selection between different individuals (i.e., relaxing common assumptions of additive separability). Whether in labor, health, education, or public economics, any applied topic that involves heterogeneous costs and benefits could benefit from relaxing the assumptions about the nature of selection in levels and selection on gains. Relaxing these assumptions allows us to explore whether structural inequities prevent disadvantaged groups from benefiting from program participation such as school choice, college admissions, health investments, or other human capital decisions.

CHAPTER II

Subsidies with Deadlines: Optimal Tax and Subsidy Policy with Multiple Instruments (with Owen Kay)

2.0 Abstract

Although Pigouvian subsidies correct externalities efficiently in standard economic theory, many real-world subsidies are not Pigouvian: either ending after a deadline or subsidizing investment rather than output. This paper presents an optimal tax framework for "subsidies with deadlines" that nests traditional investment and Pigouvian output subsidies, revealing that a subsidies with deadlines are often efficient. The optimal policy trades off the social costs of a longer deadline against the externality benefits of increased production—a sufficient statistic for deadline length. We estimate this change in production in the US wind industry using the 10-year eligibility deadline for the production tax credit. A regression discontinuity finds an 8% reduction in output after the deadline, implying that US energy markets are foregoing over 500 GWh of wind energy per month due to the deadline. Inverse optimum exercises require implausibly large social costs to justify the ten-year deadline, and calibrations show that even second-best output subsidies with too-short deadlines can forego upwards of 20% of social welfare the efficient market would produce.

2.1. Introduction

Subsidizing positive-externality goods is a hallmark of economic theory and policy. In standard theory, Pigouvian output subsidies make efficient corrections (Diamond and Mirrlees, 1971; Kopczuk, 2003), but in practice many subsidies are not based on output at all, and many output subsidies end after specified deadlines. For example, in the United States the research and experimentation tax credit subsidizes research inputs not output, and investment subsidies support solar energy rather than output subsidies. Similarly, federal production subsidies for electric vehicles phase-out after a quota, and wind facilities only receive output subsidies for their first ten years of operation.This paper establishes a unifying framework for these subsidies and characterizes efficient departures from the Pigouvian standard.

Our framework models Pigouvian subsidies and investment subsidies as special cases of "subsidies with deadlines." Under a Pigouvian policy, an output subsidy covers the entire capital life before reaching the effective deadline. On the other hand, the investment subsidy is paid at construction, and outputs are not subsidized for any of the capital life, so the effective deadline is before all production. In general, a deadline could be set anywhere between these two extremes. This insight yields three policy for the optimal policy problem: the size of the investment subsidy for fixed inputs, the size of a per-unit output subsidy, and the deadline for how much of the capital life the output subsidy covers. In this framework, subsidies with deadlines nest four classes of subsidies: Pigouvian subsidies with no deadlines and no investment subsidy, investment-only subsidies, subsidies with deadlines that mix output and investment, and output-only subsidies with deadlines.

Solving the model reveals that subsidies with deadlines are an efficient policy response to social costs of subsidization. This result holds for both first- and second-best optimal tax problems. There are two key insights. First, the efficient deadline trades off the marginal external value of increased production from a longer deadline against its marginal social cost (such as administrative costs or the marginal cost of public funds, see Ng, 1980; Dharmapala et al., 2011; Keen and Slemrod, 2017). As such the production response at the deadline—or the change in output when the subsidy ends—is a sufficient statistic for the optimal length. Second, efficiency requires that the value of the investment subsidy increase as the deadline becomes shorter. For example, if the optimal deadline is the entire capital life, the optimal policy is the standard Pigouvian ideal. This would occur if the change in production is large compared to the social costs. On the other extreme, if the optimal deadline is none of the capital life, the optimal policy is a traditional investment subsidy. This would occur if the change in production is small compared to the social costs. Between these extremes an output-only subsidy with a deadline can never be efficient because without an investment subsidy it cannot ensure an efficient choice of capital.

Because the production response at the deadline is the key statistic in determining efficient deadline length, we estimate the production response at the ten-year deadline of the production tax credit (PTC) in the US wind industry. In the wind industry the production response at the deadline may be small because turbines are essential, and wind is free. This makes it a limiting case of the model since a (shorter) deadline is only justified when the change in production is small.[1] When we estimate the change in production using a regression-discontinuity at the 120th month of operation, we find that the deadline results in a 8% reduction in output. This suggests that PTC ineligibility is resulting in over 500 GWh of forgone production each month. Inverse optimum exercises would require frictions of over $200,000 per firm per year to justify the ten-year deadline. Since the time horizon is too short even though the change in production is small, we conclude that many other subsidies with deadlines have horizons that are also too short.

Our paper makes its contribution to three main bodies of research. First, we provide a general model that nests output and investment subsidies and provides the first theoretical treatment of subsidies with deadlines. Furthermore, we show that investment subsidies and subsidies with deadlines may be efficient (even when revenue can be raised without distortions) whereas traditional theory suggests that output subsidies dominate (Diamond and Mirrlees, 1971), unless there are budget concerns (Parish and McLaren, 1982) or many compounding market imperfections (Yi et al., 2018). If the marginal cost of public funds is one, the optimal subsidy mix is always an output subsidy equal to the size of the externality plus an investment subsidy calibrated to the optimal deadline length. This manifestation of the additivity principle parallels other results on externality targeting in optimal taxation. For example, the targeting principle holds for commodity taxes (Sandmo, 1975), international tax policy (Dixit, 1985), public good provision (Bovenberg and van der Ploeg, 1994), joint income and commodity taxation (Cremer et al., 1998), and other domains generalized by Kopczuk (2003).

Our second contribution is providing an empirical comparison between investment subsidies, output subsidies, and subsidies with deadlines. There is a growing empirical literature exploring subsidies in specific industries, usually through the estimation of structural models of entry or technology adoption. These papers study industries on a case-by-case basis and show that one subsidy instrument can generate large savings and market improvements relative to another (for example Dunne et al., 2013; Burr, 2016; Yi et al., 2018; De Groote and Verboven, 2019). In these papers the preferred subsidy varies from case to case, but our model can explain these differences. We are among the first to explore subsidies with deadlines in general, although Lohawala (2022) quantifies implications of the dynamic incentives that deadlines and quotas create in electric vehicle markets. Furthermore, whereas previous work has focused exclusively on cost effectiveness or maximizing gains from budget neutral changes, we extend our comparisons to welfare and efficiency. The optimal subsidies we characterize in the face of revenue-raising costs reflect the intuition from previous work on cost effectiveness

[1] Here the positive externality is carbon offset. It is estimated that each megawatt hour (MWh) of energy produced by wind offsets 1,382 lbs. of CO_2 emissions (Estimate from EPA AVERT 2020).

while actually ensuring efficient allocations.

Our third contribution is that the empirical results extend the conversation about subsidies for wind energy. Despite the fact that the PTC has been a staple of US policy for alternative energy development for 30 years, little is known about how the deadline structure has affected the industry. We find that firms decrease output by about 8% after the PTC deadline. Economic research on the wind industry has explored topics including siting (Jarvis, 2021), investment and technology (Cook and Lawell, 2020; Lee and Howard, 2021), externalities (Cullen, 2013; Novan, 2015; Fell et al., 2021), intermittency (Ambec and Crampes, 2019; Kaffine et al., 2020) and welfare in energy markets (Callaway et al., 2018; Liski and Vehviläinen, 2020; Karaduman, 2021). Of the papers studying or comparing subsidies (including Schmalensee, 2012; Johnston, 2019; Abrell et al., 2019; Aldy et al., 2021; Helm and Mier, 2021; Petersen et al., 2021), to our knowledge only Hamilton et al. (2020) considers the nature of the PTC deadline, and they focus on degradation leading up to the deadline and changes in degredation after the deadline. Aldy et al. (2021) compare firms who choose to receive the PTC versus firms who choose to receive an investment subsidy (Section 1603 Grants), finding slightly larger differences in production than we find. Our contribution is quantifying how the PTC deadline affects firms' incentives to produce, showing that US energy markets are forgoing over 500 GWh each month because of the deadline (before accounting for any investment responses).

The rest of the paper is organized as follows: Section 2.2 explores the theoretical justifications and policy context for different subsidies; Section 2.3 sets out the theoretical model and results; Section 2.4 presents the empirical design, results, and robustness; Section 2.5 discusses calibrations and their theoretical and market implications; and Section 2.6 concludes and discusses the broader policy implications of the results.

2.2. Policy Background

This section explores the context of subsidy policy. We describe existing theoretical justifications for output subsidies, investment subsidies, and subsidies with deadlines with accompanying examples.

2.2.1 Subsidization in Theory and Practice

Traditionally, economists have advocated for Pigouvian taxes and subsidies in response to externalities.[2] In Principles classes around the world, new economists are taught that

[2]Corrective and Coasian solutions to internalizing social costs have strengths and shortcomings (Shavell, 2011), but policymakers seem to be more apt to use corrective measures.

corrective taxation addresses externalities when calibrated to the marginal externality of each unit produced (e.g., Mankiw, 2020; Stevenson and Wolfers, 2020). This focus on output subsidies is theoretically justified by results that production efficiency requires no taxes on inputs (Diamond and Mirrlees, 1971), and that an externality can be directly subsidized (or taxed) by the targeting principle (Kopczuk, 2003). Under this paradigm investment subsidies are undesirable because while they may correct an externality, in so doing they distort the efficient input mix. It is common for there to be production taxes on negative externality goods like gasoline, air travel, cigarettes, alcohol, and sweetened drinks, and these policies almost never feature deadlines; however, with the exception of ethanol subsidies, nearly all output subsidies for positive externality goods we are aware of in the United States feature deadlines.[3]

The absence of true "Pigouvian" subsidy programs for positive externality goods may have to do with concerns about cost effectiveness or market imperfections. When there are social costs to raising additional tax revenue, standard subsidization results can be completely reversed (Ng, 1980), and a social planner may reasonably choose a subsidy program focused on the social gains from marginal (rather than inframarginal) units of output. Input subsidies can be justified by decreasing returns to scale (Parish and McLaren, 1982)[4] or by price uncertainty in imperfectly competitive markets (Yi et al., 2018). Investment and input subsidies are much more common for positive externality goods such as research and experimentation subsidies, low-income housing development, and adoption of renewable energy generation technology.[5]

While there are many investment subsidies for positive externality goods, output subsidies with deadlines are much more common than true Pigouvian subsidies. For example, in the United States wind energy, geothermal energy, closed- and open-loop biomass, municipal solid waste, and qualified hydroelectric and hydrokinetic energy all receive output subsidies with deadlines. These generally take the form of tax credits with a ten-year deadline for each facility. There are also output subsidies with deadlines for the purchase of new electric vehicles, although in these industries the deadlines are triggered by quotas rather than being fixed.[6]. Interestingly, the original legislation for the alternative energy production tax credits did not include a subsidy deadline (United States Congress, 1991), but it was added as a part

[3]The extent to which ethanol actually is a positive externality good is disputable, but it is the best example of a truly Pigouvian-like production subsidy we could find in the US.

[4]There is a slight tension here because the results in Parish and McLaren (1982) hinge on increasing or decreasing returns to scale (and an output subsidy to be equally cost effective to an input subsidy for all inputs under constant returns), but Diamond and Mirrlees (1971) assume constant returns to scale in production to guarantee their zero-profit condition. Both agree that in general subsidizing some inputs will disrupt production efficiency.

[5]We are likewise unaware of investment taxes used in the production of negative externality goods.

[6]See Lohawala (2022) for an exploration of how the dynamic incentives created by this policy feature affects producer and consumers.

of the political process, suggesting the presence of additional frictions. Although subsidies with deadlines are common—especially for alternative energy—we are aware of no theoretical or empirical results about how they compare to other subsidy designs.

2.3. Optimal Subsidies with Deadlines

This section models optimal subsidies with deadlines when firms make investment and output decisions. The model represents a unifying framework for output subsidies, investment subsidies, and subsidies with deadlines. We define a sufficient statistic for the optimal deadline and characterize the optimal taxes in both first- and second-best settings, exploring conditions under which subsidies with deadlines are efficient departures from Pigouvian subsidization.

2.3.1 Firm and Government Problems:

The firm's objective is to choose a level of capital and variable inputs to maximize profits. The firm produces output according to a production function $q(x, v)$ where q is the quantity produced from the combining a fixed input, x, and a variable input, v. Output is sold in a competitive market with price normalized to one, and variable inputs and capital are purchased in competitive factor markets with prices m and c. The capital life is divided into two portions of length T and $1 - T$. There is an investment subsidy on the capital good τ_i and an output subsidy of size τ_o with a deadline at T.

Firms choose capital investments and variable inputs to maximize profits. Firms make all decisions upon entry with perfect foresight about prices and policy. Specifically, the firm invests in a level of capital, x, which remains fixed, and chooses two levels of variable inputs v_1 and v_2, corresponding to production before and after the deadline, T. In Appendix B.3 we show that this simplified setup is isomorphic to the continuous-time version with discounting where firms choose a function $v(t)$ for each moment of the capital life ($t \in [0, 1]$). We define the firm's problem as follows:

$$\max_{x,v_1,v_2} \pi(x, v_1, v_2; \tau_i, \tau_o, T) = T[q(x, v_1) + \tau_o q(x, v_1) - mv_1] + (1-T)[q(x, v_2) - mv_2] - cx(1-\tau_i) \tag{2.3.1}$$

Note how this subsidy framework nests other output and investment subsidies. A Pigouvian policy features an output subsidy that runs for a firm's entire capital life, $T = 1$, and no accompanying investment subsidy $\tau_i = 0$. On the other hand, only subsidizing investment is functionally equivalent to having an immediate deadline at the beginning of the capital life, $T = 0$. For subsidies with binding deadlines, $T \in (0, 1)$, policies could potentially subsidize only output or could subsidize both output and investment.

Solving this problem requires some regularity conditions:

Assumption 1. Assume (1) that $q(x, v)$ is increasing in both arguments with decreasing returns such that there exists an interior solution (x^f, v_1^f, v_2^f); and (2) that the firm choices (x^f, v_1^f, v_2^f) are implicit functions of the policy parameters (τ_i, τ_o, T) such that all first order conditions are continuously differentiable with respect to all arguments and produce a matrix $F = (f_x, f_{v_1}, f_{v_2}) = 0$ with a non-singular Jacobian with respect to x and v_t.

The solution to the firm's problem under Assumption 1 is in Appendix B.2 where we implicitly define x^f and v_2^f as functions of the policy parameters using a second-order Taylor expansion of around $q(x^f, v_2^f)$.

The government takes the solution to the firms problem as given and tries to design a tax/subsidy system to maximize welfare. Welfare in our model is the sum of four components: firm profits, the external benefit of production, the social cost of raising tax revenue, and an administrative (or compliance) cost of the policy. The externality is proportional to production, with each unit of output producing γ of external benefit. Tax expenditures are multiplied by the marginal cost of public funds (λ) that captures the social value of \$1 of government revenue relative to the value of giving \$1 to the firm.

In addition to the costs of production, we allow there to be an administrative or compliance cost associated with running subsidies over time, denoted by $\phi(T)$. This term captures all real costs associated with running the subsidy program for a longer time, T. It could capture the managerial burden to tax administrators, firms, and accountants, as well as reflect pricing uncertainty, policy uncertainty, and issues of political economy that make *ad infinitum* subsidies complicated to implement in practice.

The government's problem, therefore, is to select the investment and output tax rates (τ_i and τ_o) and the length of the output subsidy T to solve the following maximization problem:

$$\begin{aligned}
\mathcal{W}(\tau_i^*, \tau_o^*, T^*) &= \max_{\tau_i, \tau_o, T} \Pi(x^f, v_1^f, v_2^f) + \gamma\big[Tq(x^f, v_1^f) + (1-T)q(x^f, v_2^f)\big] - \lambda\big[cx^f\tau_i + T\tau_o q(x^f, v_1^f)\big] - \phi(T) \\
&= \max_{\tau_i, \tau_o, T} \Bigg\{ \max_{x, v_1, v_2} \Big\{ T[q(x, v_1)(1+\tau_o) - mv_1] + (1-T)[q(x, v_2) - mv_2] - cx(1-\tau_i) \Big\} \\
&\quad + \gamma\big[Tq(x^f, v_1^f) + (1-T)q(x^f, v_2^f)\big] - \lambda\big[cx^f\tau_i + T\tau_o q(x^f, v_1^f)\big] - \phi(T) \Bigg\} \qquad (2.3.2)
\end{aligned}$$

Note the absence of a demand side in the model. Implicitly we are assuming (1) that there are other technologies for producing the good—but without the positive externality—and (2) that firms with the externality-generating technology are "small" relative to the other firms.

To solve the social planner's problem, we will consider an assumption of convex administrative costs:

Assumption 2. Assume $\phi(T)$ is twice continuously differentiable with $\phi'(T) \geq 0$ and $\phi''(T) \geq 0$.

Although costs of compliance may not be convex, the real costs of uncertainty and political economy are almost certainly convex in the deadline length. We present solutions to this problem under Assumption 1 and under Assumptions 1 and 2 in the following subsections.

2.3.2 Welfare Maximizing Subsidies with Deadlines

This section proposes and discusses three theoretical results about first- and second-best subsidies with deadlines at the efficient solution. In both cases deadlines an efficient policy instrument in the face of social costs. Two points are of particular note. First, if there is a deadline, efficiency requires that the value of the investment subsidy increase as deadlines become shorter. Second, the efficient deadline must trade off the marginal external value of increased production against the marginal social costs of the deadline (whether administrative or revenue-raising costs).

2.3.2.1 Optimal Subsidy Values with Deadlines

Because investment subsidies and Pigouvian subsidies are on a spectrum based on the deadline length T, our first theoretical result, Proposition 1, describes the optimal taxes (τ_o, τ_i) for each deadline T. It is convenient to allow the taxes to be functions of T for two reasons. First, characterizing the optimal policy over T describes the set of second-best allocations given T in addition to the first-best allocation. Policymakers may be interested in knowing the optimal subsidy policy given some T that is politically feasible or already chosen. Second, as will be shown in Proposition 2, the optimal T depends crucially on the nature of the $\phi(T)$ function. As this function is difficult to measure, there may be cases where the optimal T is not known but policy makers nevertheless want approximate guidance improve subsidies given a proposed deadline.

For this first result, we focus on the simplified first-best case where there is no dead-weight loss from raising the required tax revenue (or welfare gain from spending the generated revenue). This benchmark case does not require a lump-sum instrument to raise revenue, but is justified more generally in any optimized tax system (Jacobs, 2018); nevertheless, we relax this assumption later in this section.

Proposition 1. Given Assumptions 1, if the marginal cost of public funds is $\lambda = 1$, then a second-order Taylor approximation around $q(x^f, v_2^f)$ implies the following implicit definitions of the optimal taxes τ_o^*, τ_i^* given any value of T:

$$\begin{aligned}
\tau_o^* =& \gamma \\
\tau_i^* =& \frac{\gamma - T\tau_o}{c} \frac{\frac{\mathrm{d}q}{\mathrm{d}\tau_i}}{\frac{\partial x^f}{\partial \tau_i}} = \frac{\gamma(1-T)}{c} \frac{\frac{\mathrm{d}q}{\mathrm{d}\tau_i}}{\frac{\partial x^f}{\partial \tau_i}}
\end{aligned}$$

Proof in Appendix B.2.

Proposition 1 has three main implications worth discussing: the Pigouvian-like nature of the ideal output subsidy, the form of investment subsidy, and a connection between these results and the targeting principle.

First, we note that the first-best policy is always to set $\tau_o = \gamma$ equal the production externality. If $T = 1$, this result follows given the results from Diamond and Mirrlees (1971), Sandmo (1975), and Kopczuk (2003), but interestingly it also holds true even when the subsidy operates for only a fraction of the capital life ($T < 1$). Because the same *ad valorem* subsidy amount will increase production less the sooner its deadline comes, it was not clear *ex ante* that the optimal output subsidy would be equal to the marginal externality. In fact, without the other policy instruments, this result would not hold in general (see derivation in Appendix B.2). The key to this Pigou-like formulation is also having the investment subsidy.

The second important implication from Proposition 1 is the critical role of the investment subsidy in obtaining first-best allocations. Although the ideal investment subsidy is $\tau_i = 0$ when $T = 1$, it is not the case that there should be no investment subsidy in general. In fact, the ideal investment subsidy is always weakly positive but is decreasing in T whenever $c > 0$ and $\frac{\mathrm{d}q}{\mathrm{d}\tau_i} \neq 0$. It is also worth noting that whenever $T < 1$, maximizing welfare hinges on the availability of investment subsidies since without it the output subsidy with a deadline cannot appropriately target both the production externality and achieve the efficient level of investment.

The third implication is a connection with the targeting principle for externalities. For any deadline T, the optimal subsidy is an output subsidy equal to the size of the externality plus an investment subsidy that decreases in the deadline length. The presence of the additive principle here parallels other results on externality targeting in optimal taxation. For example, the targeting principle holds for commodity taxes (Sandmo, 1975), international tax policy (Dixit, 1985), public good provision (Bovenberg and van der Ploeg, 1994), joint income and commodity taxation (Cremer et al., 1998), and other domains generalized by Kopczuk (2003). Intuitively, this is because a shorter deadline reduces output through two channels.

First, firms produce less when the subsidy is not active ("moral hazard"), and, second, firms invest less when the subsidy is going to be run for less time ("under-investment"). Because increasing the investment subsidy offsets the under-investment effect, the output subsidy need only correct the externality, leading to the Pigouvian-like calibration. Neither subsidy can address the "moral hazard" effect; this is something that only extending the deadline can improve.

2.3.2.2 Optimally Choosing Deadlines

Given the dependence of optimal subsidies on the deadline, T, our second theoretical result, Proposition 2, describes the optimal deadline T^*. We continue to focus on the baseline case of efficiency when there are no social costs to raising revenue.

Proposition 2. Under Assumptions 1 and 2, if $\lambda = 1$, a second-order Taylor approximation around $q(x^f, v_2^f)$ yields the following implicit characterization of the optimal deadline T^* for interior solutions:

$$\phi'(T^*) = \gamma\big[q(x^f, v_1^f) - q(x^f, v_2^f)\big] \equiv \gamma\Delta q$$

Proof in Appendix B.2. Note that Assumption 2, convex administrative costs, is necessary for an interior solution (although the proposition is trivially true in other cases). We relax this assumption to allow administrative costs to be concave (but still increasing) hereafter.

We discuss two implications of Proposition 2: a connection to marginal costs and benefits of extending deadlines, and the resulting sufficient statistic for the optimal deadline length.

Proposition 2 shows that the optimal deadline trades off the marginal (administrative) cost of a longer output subsidy against the implied marginal (externality) benefits. In general the social benefits of a longer deadline could accrue through reducing "moral hazard" (firms go longer before reducing production) or reducing "under-investment." As the deadline gets longer, the under-investment channel is controlled by the investment subsidy, and the remaining marginal benefit of extending the deadline comes only from increased production.[7] The size of this benefit is equal to the external value of the units produced because of the extended deadline, $\gamma\Delta Q$. At the optimal T^* the marginal administrative cost of extending the deadline is equal to the marginal social benefit of the extension (or a corner solution, $T \in \{0, 1\}$).

Because the change in quantity characterizes he optimal deadline, T^*, it is a sufficient statistic for the optimal policy. In this view, the argument about marginal costs and benefits

[7]This also suggests that if an investment subsidy is not feasible, the optimal time deadline should be later all else equal, since the marginal benefit to extending the subsidy will be much larger.

from the previous paragraph can also be interpreted in terms of elasticities. Industries with more elastic production will face larger changes at the deadline and should be subsidized longer, all else equal. This is because there is a greater social benefit to extending the deadline when production is more elastic (conditional on investment). In the context of policy this means that if the social planner has accurate beliefs about the values of γ and $\phi(T)$, then ΔQ contains all the relevant information about the optimal length of the production subsidy. We estimate this change in production at the deadline in Section 2.4.

Taken together, Propositions 1 and 2 show that subsidies with deadlines and even pure investment subsidies are efficient policies. Traditional theory suggests that output subsidies dominate (Diamond and Mirrlees, 1971), but Propositions 1 and 2 reveal that in the face of large administrative cost, investment subsidies and subsidies with deadlines can strictly dominate output subsidies. We explore the degree to which a (second-best) policy constrained to $T < 1$ and $\tau_i = 0$ reduces overall welfare in calibrations in Section 2.5. Another common justification for non-Pigouvian taxation comes from budget concerns (Ng, 1980) cost effectiveness (Parish and McLaren, 1982; Yi et al., 2018), but assessing whether these concerns can justify deadlines requires relaxing the assumption that $\lambda = 1$.

2.3.2.3 Second-Best Subsidies with Deadlines

Finally, we explore the case where raising tax revenue to fund the subsidy programs imposes a social cost. There are two reasons why this might be the case. First, although the marginal cost of public funds should be 1.0 at the optimal tax system (Jacobs, 2018), there is no guarantee that every other aspect of the system is perfectly optimal. Second, even if revenue was raised with lump sum taxes, the marginal utility of firm owners (who receive the subsidy) may be lower than the marginal utility of average consumers (who pay the lump sum tax). In either case, we can still solve for the optimal policy parameters taking $\lambda > 1$ as given.

Corollary 1. Under Assumptions 1 and 2 and allowing the marginal cost of public funds to be $\lambda > 1$, the optimal τ_o^*, τ_i^*, and T^* have the following implicit definitions (in the case of T

for internal solutions):

$$\tau_o : \qquad \frac{m^2(\gamma - \lambda\tau_o^*)}{(1+\tau_o^*)^3} = (1-\lambda)\left[\left(q_x - m\frac{q_{xv}}{q_{vv}}\right)x^f - q(x^f, v_1^f)\right]$$

$$\tau_i : \qquad \tau_i^* = \frac{\gamma - T^*\tau_o^*}{c}\frac{\frac{\mathrm{d}q}{\mathrm{d}\tau_i}}{\frac{\partial x^f}{\partial \tau_i}} + \frac{(1-\lambda)x^f}{\lambda\frac{\partial x^f}{\partial \tau_i}}$$

$$T : \qquad \Delta Q\gamma = \phi'(T^*) - (1-\lambda)\tau_o^*\left[\frac{c}{T^*} - q(x^f, v_1^f) + \gamma\left(q_x\frac{\partial x}{\partial T} + q_v\frac{\partial v}{\partial T}\right)\right]$$

Proof in Appendix B.2

We note three implications of Corollary 1: An ambiguity about whether it is optimal to tax or subsidize production, changes in optimal subsidies with respect to λ, and connections to the large literature on subsidy cost effectiveness.

Regarding to the subsidy size, Corollary 1 shows that when there is a social cost of raising tax revenue, it may be optimal to tax—rather than subsidize—output, investment, or both. In these cases, large values of λ imply that funding subsidies generates such a large cost and it becomes optimal to forego producing marginal units of the externality good. When the cost is particularly high, it may even be optimal to tax the positive externality good because the additional revenue has a greater social value than the production. The ambiguity between subsidy and tax values reveals that it is efficient to tax socially beneficial activities. Similarly, when the marginal cost of public funds is high, it has been show than it can be efficient to subsidize harmful activities (Ng, 1980). Furthermore, whereas taxing output does not distort the efficient input mix, any non-zero investment tax or subsidy will. Thus, Corollary 1 reveals that a high costs of public funds can overturn traditional results about production efficiency (Diamond and Mirrlees, 1971).

Interestingly, as λ increases, the change in both τ_o and τ_i is technically ambiguous. For simple production functions (see Cobb Douglas results for intuition in Section 2.5) τ_o monotonically changes from a Pigouvian subsidy to a tax, but this need not hold in general. On the other hand, τ_i is not even monotonic in the simplest parameterizations. Near $\lambda = 1$ it is often optimal to subsidize investment *more* as λ increases. This increase is necessary to avoid the under-investment that can occur as a result of τ_o is becoming more like a tax.[8]

In addition to changing optimal subsidy values, marginal costs of raising revenue may also change optimal deadlines. In fact, subsidy deadlines can improve welfare when raising revenue is costly. In the lump-sum tax setting, deadlines could only be rationalized by an increasing

[8]When τ_o is a tax, a positive investment subsidy can increase in tax revenue, but the total effect on the budget will depend on market primitives.

cost of administration or compliance; however even when there are no administrative or compliance costs ($\phi(T) = 0$), the presence of costly taxation is sufficient to make deadlines optimal ($T^* < 1$). What's more, subsidies with deadlines can strictly dominate both traditional Pigouvian subsidies ($T = 1$ and $\tau_o = \gamma$) and standard investment subsidies ($T = 0$ and $\tau_i = \tau_i^*$).

The fact that deadlines may be efficient second-best policies is new to the conversation of subsidy choice that has often focused on cost minimization (Parish and McLaren, 1982; Yi et al., 2018; Aldy et al., 2021) rather than efficiency. The main difference between these approaches and ours is that our optimal policy parameters minimize the cost of the efficient allocation, whereas the cost minimization exercises are subject to subsidizing a given level of output—one that need not be efficient. For example, we show that in cases with a high marginal cost of public funds, the optimal "subsidy" may actually be a tax, since the tax revenue from a unit may be worth more to society than the value of the externality.

Corollary 1 also highlights the role of returns to scale, complementary, and marginal firm responses in determining the optimal policy. This is seen in the higher-order moments of the production function that appear in each equation. The importance of production technology reflects insights from research on cost effectiveness, but it clarifies previous results in the light of overall efficiency. For example Parish and McLaren (1982) show that in a simple model input subsidies are more cost effective when there are decreasing returns to scale; note a similar pattern in the q_{vv} term appearing in the equation for τ_o^*. Yi et al. (2018) argue that in the presence of social costs (or, in their model, concern for cost effectiveness) efficient investment subsidies may arise through the interaction of dynamics, market imperfections, and price uncertainty, but Corollary 1 shows that only dynamics are necessary.

Given these social costs, whether deadlines improve welfare or not depends on the production response at the deadline. In order to quantify the importance of those effects, we turn to our empirical investigation: how the Production Tax Credit's deadline affects production at wind energy facilities and what that tells us about optimal policy.

2.4. Production Responses at the Production Tax Credit Deadline

Given the critical role that changes in production play in the design of an optimal subsidy, we turn to an exploration of how production of energy from wind responds at the the production tax credit (PTC) deadline. After presenting background on the wind industry, its subsidy policies, and why it is an ideal setting for exploring our optimal policy questions, this section presents evidence that facilities reduce production in response to the loss of output subsidies, evaluates the results' robustness, and demonstrates that the effects do not seem to be driven by facility retirement or entirely explained by reduced deployment in response to low average prices.

2.4.1 The US Wind Industry

We now turn to a more specific description of the wind industry, profit maximizing decision, and policy domain. Wind facilities use kinetic energy from wind to generate electricity (measured in megawatt-hours, MWh). The amount of energy generated is determined in part by the invested capacity and technology, and by the velocity of the wind. In 2020 the average ratio of production to capacity, called the capacity factor, was 36% (Wiser and Bolinger, 2021). Firms obtain revenue by selling electricity on spot markets or through long-term power purchasing agreements. Subsidies make up another major revenue source for wind facilities.

When developing a wind facility at a given site, firms maximize profits by choosing how much capacity to develop (investment) and making operational decisions as they run (production). Investment costs are paid at the outset of the project. These costs include turbine purchase and installation, interconnection costs, and other investments costs, called balance of plant. Estimates for the average investment costs are on the order of $0.8-1.5 million per megawatt (MW) of capacity (Wiser and Bolinger, 2021).[9] Although wind is a free input, facilities continue to incur production costs after investment. These costs can include maintenance and repairs, rent for the land, optimization software or consulting, and wages to workers. In 2020 these fixed and variable operation and management costs averaged about $25 per kW-year, or just under $10 per MWh (Wiser and Bolinger, 2021).[10]

2.4.1.1 Output and Investment Subsidies for Wind Energy

The policy environment for the US wind industry has been characterized by five main sets of subsidy policies. These policies include investment subsides, output subsidies, and output subsidies with deadlines, implemented by federal, state, and local governments.

The most characteristic policy in the US wind industry is an output subsidy with a deadline called the production tax credit (PTC). The PTC has been in place since 1992 and awards $25 (nominal in 2020 and 2021) of non-refundable tax credits to firms for every MWh of wind-generated electricity they produce for 10 years of operation. This is a large subsidy: for context, wholesale prices average between $30-45 per MWh, so the PTC is in the range of 55-80% of average wholesale prices. Research on the effects of these subsidies on firms has

[9] Capacity is measured in MW and reflects the production potential of a generator. For example, a 1 MW wind turbine running at full capacity for one hour would produce 1 MWh of energy.

[10] Wiser and Bolinger (2021) report that average costs are $25 per kW-year. Assuming a capacity factor between 0.3 and 0.4 implies average costs of $7-$9.5 per MWh.

shown that production degradation accelerates in the years after passing the subsidy deadline (Hamilton et al., 2020, though we know less about the immediate effects) and that because the subsidy is nonrefundabile and politically uncertain, it's real value is less than $25 per MWh (Grobman and Carey, 2002; Johnston, 2019).

A second federal policy was a temporary investment subsidy that paid for 30% of a firms' investment costs in cash, known as the Section 1603 investment grant. Firms entering between 2009 and 2012 chose whether to receive the PTC or this grant. Aldy et al. (2021) describe the history and implications of this policy and show that although the subsidy may have made some marginal entrants profitable, firms who were induced to take up the investment subsidy produced less and invested less efficiently (as predicted by Diamond and Mirrlees, 1971; Parish and McLaren, 1982). They also show that for a wide range of quantity targets the 30% investment subsidy would be relatively less cost effective than the existing PTC.

A third federal policy is an active investment subsidy through accelerated depreciation. Traditionally, tax deductions for an investment are claimed over its capital life. Accelerated depreciation allows firms to deduct these expenses over one to five years rather than the twenty-year baseline; thus, an investment in wind that generates one dollar of profits costs less in real terms than an investment in another industry with slower (or no) bonus depreciation. We are unaware of any research on accelerated depreciation as an investment subsidy; however, current policy renders the benefits of accelerated depreciation for alternative energy investments moot. The CARES Act allows taxpayers to immediately expense the full cost of qualified assets (including but not limited to wind generation infrastructure) bought and placed in service between September 2017 and December 2022 (see Guenther, 2018, for more), removing any marginal incentive to invest in wind generation.

In addition to these federal projects, many states have Renewable Portfolio Standards (RPS) that incentivize alternative energy production. In states with mandated standards (e.g., 35% renewable), producing qualified energy generates Renewable Energy Credits (RECs) which can be sold to buyers who need to meet the standard. Because firms receive a market price for REC, RPS function as a state-level output subsidy with a varying price. Research on RPS finds that they lead to increases in renewable production and reductions in carbon emissions locally and in neighboring states (Greenstone and Nath, 2020; Hollingsworth and Rudik, 2019, although they may not be cost effective).

Finally, some states and localities also give preferential property and sales tax treatment to wind facilities, land, and purchases. Reduced property tax liabilities for sites and reduced sales tax liabilities for turbines function as a *de facto* investment subsidy. These liabilities are often relatively small, however, and our conversations with developers suggest that when making siting and investment decisions, these issues are overall less important than other concerns such as wind resource, interconnection, local zoning, and larger state and federal policies. We are not aware of any empirical research examining local tax treatment as an investment subsidy.

2.4.1.2 Advantages of the Empirical Setting

Given this industrial and policy context, there are three main reasons why we choose the wind industry. The first is theoretical. In Section 2.3 we showed that subsidies with shorter deadlines are justified when the change in production at the deadline is small (i.e., when production conditional on investment is inelastic with respect to the subsidy). Since the fixed inputs are extremely important in the wind industry and most variable inputs are exogenous, one might expect the change in production at the deadline to be small. In this case the wind industry serves as a limiting case of the model, bounding how much deadlines might affect other more elastic industries. If the deadline is too short in the wind industry, other subsidies with deadlines in more elastic energy may reduce the incentive to produce (and thus the external benefits) even more.

The second advantage is more pragmatic: the role of the production tax credit will be well measurable in this industry. Subsidy revenues make up large share of total revenues, and subsidy eligibility is based entirely on time in operation rather than depending endogenously on production (as studied in Lohawala, 2022). Furthermore, whereas other alternative energy producers may be disqualified from claiming the output subsidies because of other tax-credit support (e.g., biomass, landfill gas or municipal waste facilities), for wind facilities claiming the PTC is ubiquitous.[11] Finally, the rapid growth in wind energy over the last 20 powers precise statistical inference—something lacking in small industries like geothermal energy and closed-loop biomass.

The final advantage reflects the policy relevance of evaluating the PTC. The PTC has been a staple of US policy for alternative energy development for 30 years, but little is known about how the deadline structure has affected the industry. As expanding renewable energy subsidies and extending the PTC continue to be important pieces of the national conversation on energy, it is important to inform policy by exploring how subsidies with deadlines affect firms' incentives to produce.

[11] The exception being firms that enter between 2009 and 2012 that had the option of electing the 1603 grant instead. These firms are not old enough to be included in our main regression discontinuity sample but are used as a placebo test.

2.4.2 Data

We utilize detailed data about firms and their decisions, including investment, production, and subsidy receipts. Data on entry and investment are available from the Energy Information Administration (EIA). The EIA surveys all utility-scale wind facilities in the United States. The annual EIA-860 survey contains information like first date of operation, location, regulation, and information about investment such as the number of turbines and total nameplate capacity. Realized production then comes from the monthly EIA-923 survey which reports monthly generation at the plant level. Together this yields a data set with facility-by-month information on generation. We drop the first 36 months of production for each firm since not all capacity comes online at the same time.[12]

To identify the production response to deadlines, we use the 10-year eligibility deadline of the production tax credit (PTC). Empirically, we define this cutoff for firms between the 120th and 121st month of energy production recorded in the EIA-923. Because it takes firms 10 years to reach this threshold and because the EIA data cover production in 2001-2019, we focus on firms who began producing under the PTC from 2002 to 2010. We exclude facilities that started operation from 2009 to 2012 and received the section 1603 investment grants rather than the PTC from our main analysis (but use them in a placebo test later).

2.4.3 Wind Facilities Reduce Production at the PTC Deadline

In theory reducing the number of dollars received per unit of output should incentive less production, but in practice it is an empirical question whether wind facilities respond to this incentive. On one hand, investment decisions are made only once, turbines and wind may be close to Leontief in production, and firms have no control over the amount of wind (their primary variable input) that is available. This would suggest that firms have no margin for response to deadline. On the other hand, firms may still be able to respond by optimizing or maintaining their capital less effectively, engaging in curtailment in the face of low or negative prices, and possibly choosing to exit. In this subsection we demonstrate that facilities do decrease production after the PTC deadline; furthermore, we can compare these mechanisms and find suggestive evidence that the effect is not driven by exit, investment, or curtailment.

To estimate the effect of production subsidies on net generation we estimate a regression discontinuity at the 120-month threshold of PTC eligibility. To examine whether facilities

[12]Because facilities qualify for the PTC at the turbine level, the fact that not all capacity comes online at the same time may mean that some facilities are still receiving some subsidy after the 121th month of production. In this case the results can be interpreted like a reduced form effect in a fuzzy RD. In this analogy there is a strong first stage as 97.4% of capacity is online by month 12. Furthermore, the results are robust to restricting the sample to firms with all capacity online by month 1 or to dropping months 121-126 or 121-132, with effect sizes ranging between 6.5-13%.

are changing investment or exiting we run the same analysis using nameplate capacity and an indicator for whether monthly production was zero. To explore curtailment, we look at heterogeneity by average wholesale prices. Since curtailment happens when prices are negative, we can evaluate whether the entire effect is driven in low-price markets. Finally, we also explore heterogeneity by vintage since wind technologies are changing and previous research has shown that production at newer turbines depreciates much more slowly than at older facilities (Hamilton et al., 2020).

There are two econometric challenges with estimating economically interpretable effects from the regression discontinuity (RD). In fact, a naive local linear regression on a small sample of facility-month observations around the cutoff will not produce consistent results. The first concern is about sample composition because the data do not form a balanced panel. Earlier facilities also tend to have less capacity, less generation, and steeper declines in production over time (Hamilton et al., 2020). Facilities in earlier cohorts have many more observations after the cutoff than later ones, so the compositional change will introduce a negative bias in estimates of the slope after the cutoff. As a result, the limit of production after the cutoff will also not be consistently estimated. To deal resolve this concern, we focus on a balanced panel from 102 facilities who are observed at least 60 months after the deadline.[13] These firms comprise our RD sample. Appendix Figure B.1 shows robustness to the 60-month threshold for inclusion.

The second challenge is the seasonality of wind resource. There is tremendous variation (including geographic, seasonal, and year-to-year) in wind speeds. If these trends were independent of the cutoff, ignoring seasonality would yield consistent but imprecise estimates. In our data, however, more than 50% of facilities began production in December or January. These entry patterns mean that potential outcomes are not unconditionally continuous at the discontinuity. Specifically, the months from September to March are generally increasing in wind production, introducing positive bias to estimates of the production response at the deadline. Including state-by-month-year fixed effects captures the geographic, seasonal, and year-to-year variation, but a sample with 102 facilities is not powered for estimating thousands of fixed effects. To account for these trends, we expand the sample to include newer firms who produce contemporaneously with the facilities in the RD sample. Intuitively, this allows us to estimate the state-by-month-year fixed effects as nuisance parameters. By dropping all observations after these facilities' 120th month, coding their running variable as zero for all observations, and including firm fixed effects, we ensure that these firms help estimate the state-by-month-year fixed effects, but provide no identifying variation to the RD parameters.

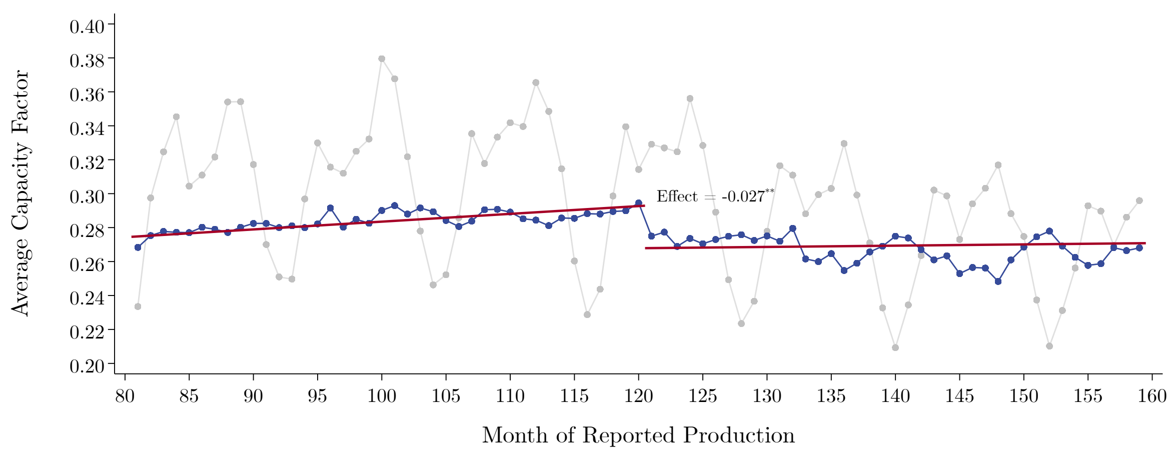
[13] To ensure balance we impute zero production in months with missing from the EIA data.

Resolving these concerns leads to the regression specification:

$$\text{Capacity Factor}_{jst} = f(r_{jt} \cdot \mathbb{1}(j \in \mathcal{J}_{RD}) + \beta \mathbb{1}(r_{jt} > 120)\mathbb{1}(j \in \mathcal{J}_{RD}) + \theta_j + \psi_{s,t} + \varepsilon_{jst}$$

Here r_{jt} is the number of months facility j has been producing at time period t, and $f(\cdot)$ is a function that measures how generation in the RD firms changes as firms age. We approximate this with an adapted local linear specification using a rectangular kernel with a 60-month bandwidth. The term $\mathbb{1}(j \in \mathcal{J}_{RD})$ is an indicator for firms in the RD sample (as opposed to the fixed effects sample). This includes months 60-180 for firms who produce for at least 180 months. The facility fixed effects and state-by-month-year fixed effects are θ_j and $\psi_{s,t}$. Finally, β is the production response at the PTC deadlines among the facilities in the RD sample. Our preferred outcome is capacity factor (the energy produced at a facility as a fraction of the production capacity). Capacity factor implicitly accounts for changes in capacity over time (which are rare) and for the fact that firms with larger capacity have higher variance in total generation given the same weather conditions. We also show specifications for net generation below.

Figure 2.1: Production Response to the Production Tax Credit



Note: This graph shows production for 102 wind facilities that began production by 2006. The semi-transparent series shows average capacity factor over each month of production, the darker series show the seasonality-corrected average capacity factor (monthly average plus average residuals from the state-by-month-year specification), and the lines of best fit show the local linear approximations of the seasonality corrected relationship between capacity factor and on each side of the deadline.

For β to identify the causal effect, the production response at the PTC deadline, we assume that the potential production with and without the subsidy are each continuous at 120 months. Intuitively, these assumptions mean that the limit of production under the PTC must be a good counterfactual for production if the deadline were extended beyond 120 months. This assumption would be violated if firms are changing their production capacity or inputs at the 120th month in ways that are not related to the end of PTC eligibility (for example repowering or seasonality). Although there is no evidence of the first concern,[14] there are monthly fluctuations in the wind that could confound estimation if not accounted for.

Figure 2.1 shows facilities' capacity factor before and after the PTC deadline. The somewhat transparent series shows the average capacity factor unconditionally across all facilities at each month of production r_{jt}. The darker series reports the average values of $y_{jst} - \hat{\theta}_j - \hat{\psi}_{s,t}$, the remaining relationship of capacity factor over time after controlling for the facility and seasonality. Whereas differences in the first series were suggestive but noisy, the difference between month 120 and 121 is clearly visible as the largest month-to-month difference in the darker series. The bold lines plot the fitted values from the local linear regression, showing the estimated difference in output at the cutoff: We estimate a 2.7 percentage point decrease in the capacity factor at the end of PTC eligibility or an 8% reduction (see Table 2.1 for standard errors).

Table 2.1 also presents the results of the regression discontinuity specification for four different dependent variables: capacity factor, net generation, whether a firm stops producing all together, and whether they change their capacity. Columns 1 and 2 report that energy production dropped when firms became ineligible for the PTC, as displayed in Figure 2.1. Column 3 shows the decrease in production is not caused by firms shutting down or exiting at the deadline, and the negative change in slope suggests that, if anything, the exit hazard becomes less steep to the right of the deadline. Similarly, Column 4 shows that there are no confounding changes in total capacity—the standard errors are small enough to rule out changes greater than 1%. Taken together the results in Panel A reveal that facilities do produce less after the PTC deadline, but this is not because of exit or unobserved factors; rather firms are continuing to use their same capital but produce less while doing so. It seems that firms are responding to the incentives implicit in the PTC deadline.

Given the important role of fixed inputs like turbines, some readers may find it striking that there is any response at all. Because the PTC deadline essentially creates a 30% reduction in prices, the implied elasticity is still low (about 0.25). These RD-based estimates are smaller than the differences in Aldy et al. (2021) who show that facilities receiving the PTC produce 10-12% more than those who received the 1603 investment grant instead. By carefully disentangling the mechanical effect of curtailment under negative prices from endogenous

[14] Repowering means investing in new generators after beginning production. And unless the repowering will requalify a facility for the PTC under the 80/20 rule, facilities receiving the PTC have an incentive to repower as long before the cutoff as possible in order to capture as much of the subsidy as they can with their increased capacity.

decisions about maintenance and optimization, they also demonstrate that curtailment only accounts for about one third of this difference.[15] Similarly, estimates of effect heterogeneity from Section 2.4.4 show that our effects also cannot be explained by curtailment under negative prices.

2.4.3.1 Regression Discontinuity Robustness

This subsection explores the robustness of the previous RD results. In addition to presenting a standard battery of RD test, we conduct two placebo tests to show no effects for firms that do not receive the PTC.

First, all of the main results are robust to a number of alternative specifications. Because the panel is balanced, there can be no bunching in the running variable. Appendix Figure B.1 shows the estimated treatment effect for different bandwidths and for including more cohorts by reducing number of required months after the cutoff. The effect remains negative and stable across bandwidths and samples, though the standard errors increase as the number of firms in the RD sample decreases. Similarly Appendix Table B.1 presents alternative estimates of the regression discontinuity using a variety of kernels and polynomial orders. For all specifications, the estimated treatment remains roughly similar with estimates ranging between 1 and 3 percentage points.

Second, we conduct two placebo tests to demonstrate that our results are driven by the PTC deadline and not by other factors in wind production or alternative energy systems. In the first placebo test, we look for effects of the PTC deadline on a set of wind facilities that received investment grants under section 1603 and who were therefore ineligible for the PTC. In the second, we look for effects at the 10-year mark for a set of solar power facilities. Solar facilities are eligible for investment tax credits but not for the PTC and should not be affected by the 10-year PTC deadline.

Panel B of Table 2.1 shows that we find null effects on capacity factor and statistically imprecise effects on net generation for both placebo groups. Columns 1 and 3 show the capacity factor results which are 5-20 times smaller than our headline results in Panel A. Note however, that since the 1603 grant firms and the entry of large scale solar are more recent phenomena, we have fewer years of data after the cutoff. To still estimate the effects, we use a smaller bandwidth to try and include as many observations as we can. Despite this, the estimates on net generation still seem to be somewhat underpowered. Despite the power concerns we find it comforting that most of the results are small and often positively signed if anything.

[15] One reader pointed out to us that because wear and tear is closely related to hours of operation (rather than MWh) another endogenous mechanism could be strategically choosing cut-in speeds.

Table 2.1: All Intensive Margin Measures of Production Respond to Ending PTC Eligibility

Panel A: Main RD Effects	Capacity Factor	Net Generation (MWh)	𝟙(Net Generation = 0)	Nameplate Capacity
Effect of Deadline	-0.027**	-794.3*	0.009	-0.020
	(0.006)	(321.0)	(.012)	(0.035)
Pre-Deadline Slope	0.000*	20.6	0.000	0.000
	(0.000)	(15.2)	(0.000)	(0.001)
Change in Slope	-0.000	1.0	-0.000	0.000
	(0.000)	(10.9)	(0.000)	(0.001)
Control Mean	0.310	18,926	0.006	82.3
Bandwidth	60 Months	60 Months	60 Months	60 Months
Facilities in RD Sample	102	102	102	102

Panel B: Placebo Tests	1603 Wind (No PTC)		Solar (No PTC)	
	Capacity Factor	Net Generation	Capacity Factor	Net Generation
Effect of Placebo Deadline	0.005	-853.1	-0.001	224.6
	(0.006)	(860.8)	(0.015)	(165.3)
Pre-Placebo-Deadline Slope	-0.001*	15.0	0.000	-0.8
	(0.001)	(68.6)	(0.001)	(6.5)
Change in Slope	0.001	47.8	0.001	-8.7
	(0.001)	(78.8)	(0.002)	(7.5)
Control Mean	0.319	21,934	0.193	1,786
Bandwidth	12 Months	12 Months	24 Months	24 Months
Facilities in RD Sample	71	71	26	26

Panel C: Effect Heterogeneity	Wholesale Prices		Turbine Vintage	
	Low Price	High Price	2002-2003	2004-2006
Effect of Deadline	-0.028**	-0.023*	-0.031*	-0.020*
	(0.007)	(0.007)	(0.010)	(0.007)
Pre-Deadline Slope	0.001**	0.000	0.000	0.000*
	(0.000)	(0.000)	(0.000)	(0.000)
Change in Slope	-0.001$^+$	0.000	-0.000	-0.000
	(0.000)	(0.000)	(0.000)	(0.000)
Control Mean	0.330	0.299	0.282	0.342
Bandwidth	60 Months		60 Months	
Facilities in RD Sample	50	52	55	47

Note: This table reports estimates from regression discontinuity analyses of the timeout of the PTC. "Control" facilities are used to identify fixed effects, but have a value of zero for treatment and the running variable. Panel A reports the main results across four different measures, the capacity factor, net generation, an indicator for whether there was zero production in a given month, and total capacity. Panel B reports two regressions of an interacted RD. The first two columns of report results for older facilities (began production in 2002-2004) vs newer facilities (in 2005-2006) and the last two columns report results for facilities who received below vs above average wholesale electricity prices in the year of the deadline. Panel C reports results for two placebo groups of firms that did not receive the PTC, wind firms that elected to receive the 1603 investment grant and solar firms that are not eligible for a PTC. For all regressions standard errors are two-way cluster corrected for arbitrary variance-covariance structure at the facility level and month-of-year level. All regressions control for facility and state-by-time-period fixed effects.
$^+p < 0.10$, $*p < 0.05$, $**p < 0.01$.

These results increase our confidence that the main effects presented in Figure 2.1 and Table 2.1 are what they seem to be. Panel A showed that firms are reducing production, but not responding on other margins, and Panel B shows that firms who cannot receive the PTC do not change their production at the 10-year mark.

2.4.4 Exploring Mechanisms

To explore the firms' responses to the PTC deadline, we estimate heterogeneity in the response at the deadline by two facility characteristics. First, we consider the average prices firms receive. Prices matter because of curtailment: In markets with low prices and in time periods with a strong winds, wind facilities may be deployed even though the wholesale prices are negative (because facilities receiving the PTC still make positive revenues). If a facility that has passed the PTC deadline is in this situation, it will have to curtail their generation because the market prices are negative. We estimate prices by dividing the total revenue facilities reported on the EIA form 923 by the total net generation in their tenth year of production. This section compares the production response at the deadline for firms who tended to receive higher wholesale prices in their tenth year with those who received lower prices. Although we don't have data on curtailment, we posit that if curtailment is driving our results, then the effect should be concentrated among firms who were receiving lower prices and who were more likely to get curtailed.

The first two columns of Panel C in Table 2.1 report the results showing significant effects of the deadline on both types of firms. To estimate these effects, we interact the treatment indicator, running variable, and interaction term with indicators for whether a firm received an average wholesale price of more or less than \$38/MWh.[16] We find both types of firms reduce production at the PTC deadline. The point estimate of the effect is larger for firms who face lower prices, but the presence of an effect in both subgroups suggests that there are more margins of response than just curtailment such as maintenance and production optimization. While it is outside of the scope of our paper, which is focused on subsidy deadlines, to quantify all of these mechanisms, both Aldy et al. (2021) and Hamilton et al. (2020) explore the roles of endogenous maintenance and utilization in wind facility performance in greater detail.

The second dimension of heterogeneity we explore is between firms operating with older or newer technology, often referred to as "vintage." Hamilton et al. (2020) find evidence that the PTC leads wind facilities to experience slower degradation in performance during the 10-year eligibility window than facilities in other countries experience. They also find that facilities

[16]Because the model includes firm fixed effects we do not need to include an indicators for whether firms received an average wholesale price of more or less than \$38/MWh.

from older vintages (defined as before 2008) experience steeper declines in production over time, but they cannot determine whether this is because of technological development or differential response to the PTC. This evidence motivates us to explore heterogeneity in how facilities from relatively older and newer vintages change production at the PTC deadline. In addition to helping understand the effects of subsidization on the US wind industry, we care about this dimension of heterogeneity because differences in responsiveness across vintages would have implications for whether the optimal subsidy policy could be changing over time.

The third and fourth columns of Panel B in Table 2.1 reveal suggestive evidence of heterogeneity between older and newer facilities. As with heterogeneity by wholesale prices, we estimate these effects by fully interacting the RD variables (treatment indicator, running variable, and interaction term) with indicators for whether a facility began production before or after January 2004 (comparing vintages 2002-2003 to 2004-2006). We find that whereas newer firms experience a 2 percentage point (5.8%) decrease in capacity factor after the PTC deadline, older firms experience a 3.1 percentage point (11.0%) decrease. Although we cannot reject the null hypotheses that these two effects are the same because splitting the sample results in larger standard errors, the magnitude of the difference is striking. We know that older vintages experience larger decreases in output leading up to the deadline (Hamilton et al., 2020). We show that it may also be the case that older vintages experience larger reductions in output at the PTC deadline. This heterogeneity has serious implications for how subsidies with deadlines would need to change across vintages to ensure an efficient allocation.

2.5. Discussion: Implications of Deadlines for Welfare, Theory, and Energy Markets

With estimates of how the PTC deadline affects production and insight into why, we return to a broader discussion of the results including their theoretical questions of optimal taxation. We begin by presenting a simple calibration that builds intuition for how optimal policy parameters change with model priors—like the production response at the deadline. The calibration also allows us to quantify the welfare loss from second-best subsidies. This section also conducts inverse optimum exercises to consider what types of social costs (either administrative or from raising revenue) rationalize current policy, and quantifies the empirical impacts of the PTC deadline on power generation.

2.5.1 Calibrated Evaluation of Welfare under the Second-Best

Although the theoretical results from Section 2.3 are powerful in their generality and the changes in production from Section 2.4 are striking, the empirical effects are difficult to interpret in terms of magnitudes and relative importance in the general model. In this subsection we explore a simple calibration and what it tells us about the quantitative sizes of various policies and their associated welfare implications.

The calibrated results all come from a Cobb-Douglas production function that assumes the following technology with decreasing returns to scale.

$$q_t = x^a v_t^b$$
$$1 > a + b$$

Under this assumption we can obtain exact expressions of the optimal policy parameters (i.e., without the Taylor Approximation, see Appendix B.4). In these expressions τ_i is ambiguous in b[17], increasing in T and γ, and unaffected by other parameters; $\tau_o = \gamma$ as in general; and ΔQ is mainly determined by b (given costs and τ_o).

2.5.1.1 Calibrations of Optimal Second-Best Subsidies

First, we use the calibration to explore the nature of optimal subsidies in response to administrative costs or social costs of taxation. Figure 2.2 shows how the optimal deadline changes under different circumstances. To generate differences in the change in output (given the optimal production subsidy), we change the variable input elasticity b holding $a + b = 0.9$ constant. Because these calibrations report results from different production functions, we report the change in production in percentage terms ($\frac{\Delta Q}{q(x,v_1)}$ to make them comparable. This percent change in output is the x axis in both panels. Note however, that this is a rescaled version of ΔQ which features in the formula for the ideal T^*.

Panel (a) of Figure 2.2 shows how sensitive the optimal time horizon T is to the change in production ΔQ. It plots a simulated correspondence between the two for different possible cost functions $\phi(T)$ (holding $\lambda = 1$).[18] As shown theoretically, the ideal deadline is at the end of the capital life $T^* = 1$ when there are no costs. For any administrative cost function, the ideal time horizon is less than one for a small enough change, ΔQ, but approaches one as ΔQ increases.[19] The specification of the cost function also generates substantial heterogeneity in

[17]The measure of $\{b : \frac{\partial \tau_i}{\partial b} < 0 \, \forall T\}$ is decreasing in γ.

[18]The simulation takes the calibrated values as given and chooses the T that maximizes welfare given b and $\phi(T)$, then plots that over the ΔQ implied by that b and the other calibrated parameter values.

[19]In the case that $\phi''(T) \leq 0$ the change will be a discrete switch from $T^* = 0$ to$T^* = 1$.

the optimal policy. For example, whereas linear, root, and log costs imply a quick switch from an investment-only subsidy to the output-only subsidy, polynomial and exponential costs reveal larger ranges of ΔQ for which a subsidy deadline is the first best.

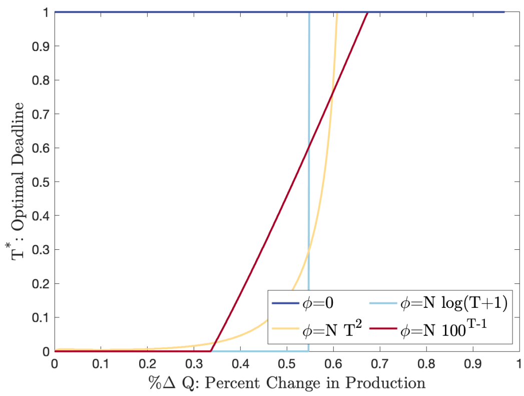
These results have two main implications for policy. First, knowing the responsiveness of firms is of first-order importance. In industries where we have reason to think that production conditional on investment is relatively unresponsive to production subsidies, investment-only subsidies may often be ideal in the face of even relatively small administrative costs. On the other hand, mixed subsidies and Pigouvian output subsidies become optimal when behavioral responses are larger. Furthermore, precision is important. In cases where $\phi''(T) \leq 0$ (e.g., linear, root, log) the first-best policy changes from being an investment-only subsidy to a Pigouvian production subsidy within the range of only a few percentage points' change in the firm response ΔQ; therefore, knowing the elasticity of output is key to knowing which policy to implement.

Second, these simulations suggest that in many cases when the only social costs are administrative subsidies with deadlines (even well-calibrated mixed subsidies) are dominated by either investment-only or production-only subsidies. If administrative costs are increasing in T but at a decreasing rate (i.e., $\phi'() > 0$ and $\phi''() < 0$), the first best subsidy will not have a deadline because there is no interior solution. On one hand, if the firm response to the deadline is small, the early-period administrative costs will outweigh the benefits of having the production subsidy. On the other hand, if the response is large, then the added benefits will tend to be much bigger than the (decreasing) costs of increasing the time horizon to the full capital life. For policy, this means that if administrative costs are concave and the subsidies can be funded with lump sum taxes, production subsidies with deadlines may be likely suboptimal.

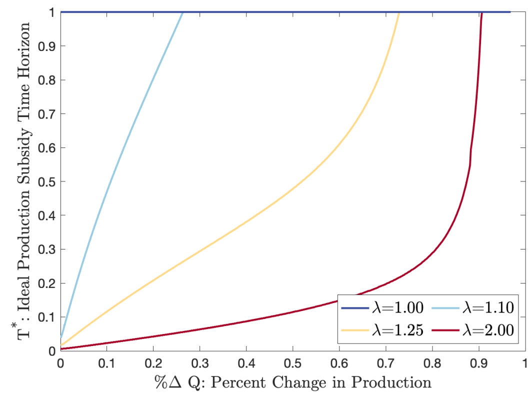
Next, Panel (b) of Figure 2.2 shows how sensitive the optimal time horizon T is to the change in production ΔQ when there are social costs of taxation. This panel plots a simulated correspondence between the $\%\Delta Q$ and T^* for different possible values of λ (holding $\phi(T) = 0$). As shown theoretically, when there are no social costs of raising revenue (as when lump sum taxes are feasible or a perfectly inelastic good is taxed), the ideal time horizon is the entire capital life $T^* = 1$. This result changes for large values of λ. In fact, as the costs of raising revenue grow, the ideal time horizon quickly falls for large ranges of ΔQ. That said, the deadline length is still increasing in the change and eventually approaches one.[20] Note that the x axis in Panel (b) reports the change in quantity that would result from a value of b

[20]This result follows from the assumption that $a + b$ is constant. We feel this assumption is reasonable when considering possible values of b for a given industry (where c and m are also held constant). In the case of comparing multiple industries with different production functions, $a + b$ should be allowed to vary. Holding a constant the optimal T^* may be U-shaped in b especially for large values of λ.

Figure 2.2: Frictions and Revenue Constraints May Make Deadlines Optimal



(a) Optimal Deadline with Administrative Costs



(b) Optimal Deadline with Social Costs of Taxation

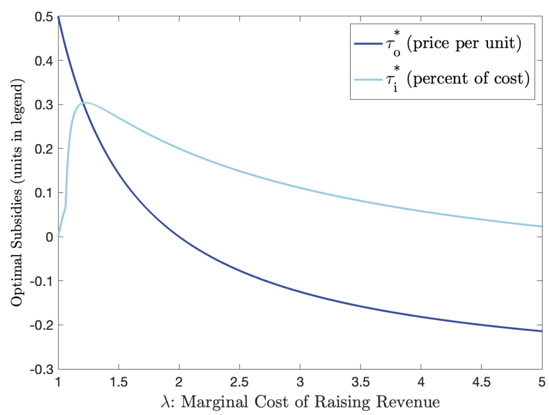
Note: This figure reports the optimal deadline T^* for different values of b and magnitudes of administrative or revenue-raising costs. The results are from an (unapproximated) Cobb-Douglas calibration $q_t = x^\alpha v^\beta$. Calibrated values: $c = 0.62$, $m = 0.3$, and $\gamma = 0.5$. We report results from calibrations that satisfy $a + b = 0.9$ and report the $T*$ in terms of the percent change in quantity $\Delta Q/q(x, v_0)$ to compare magnitudes. Note that at $\beta = 0.2$ $\%\Delta Q = -0.096$, a percent change in output close to what we observe in the empirical section of the paper.

given $\tau_o = -\gamma$. This is because when $\lambda > 1$, τ_o may be different, which in turn changes the ΔQ. The optimal T^* accounts for this but is plotted over the change most related to the primitive b because otherwise the plotted correspondences are not functions. Plotting T^* over b looks identical, but is less interpretable.

The main policy implication here is that knowing the cost of raising revenue is also tremendously important. For a given value of b (and thus ΔQ), there can be large variation in the optimal deadline T^* over a small range of λ. For example, with a 10% change in output, a 6% cost implies no deadline, a 7% cost a deadline at 0.91, an 8% cost 0.74, and a 10% cost 0.52. Interestingly, this means that a cost in the range of $\lambda \in (1.10, 1.12)$ could rationalize the ten-year deadline of the production tax credit given the empirical effect of ΔQ we estimated in Section 2.4.

It is also important to note how the optimal subsidies τ_o and τ_i change with λ. Figure 2.3 depicts this for a stylized calibration with b set to match the empirical change observed in the following section. The y axis depicts the subsidy size (or tax if negative) and the x axis the social cost of raising one dollar of revenue for the subsidies. Here we see that as λ increases τ_o decreases, changing from being a subsidy to a tax. At the same time the ideal time horizon also decreases (as seen through comparisons across curves in Figure 2.2). To compensate for the shrinking subsidy and shortening deadline, there is a rapid increase in

Figure 2.3: Optimal Subsidies with Social Costs of Taxation



Note: This figure reports the optimal subsidies τ_o^* and τ_i^* for different values of the revenue-raising costs λ. The results are from an (unapproximated) Cobb-Douglas calibration $q_t = x^a v^b$. Calibrated values: $c = 0.62$, $m = 0.3$, and $\gamma = 0.5$, $a = 0.7$ and $b = 0.2$. Positive values indicate subsidies and negative values indicate taxes.

the investment subsidy, which tapers out as changes in T^* and τ_o slow and as the costs of raising revenue for the project outweigh the social benefits of correcting under-investment.

Taken together these empirical calibrations show that the theoretical results from the previous subsection have large and economically meaningful implications for optimal policy. Subsidies with deadlines can be justified by administrative or revenue-raising costs, and the length of the deadline can vary substantially based on these policy parameters as well as the production function of the externality good and its elasticity to the subsidy deadline. These results also suggest that there may be large welfare costs to inappropriately calibrated subsidies—especially those with incorrect deadlines.

2.5.1.2 Measuring Welfare Under Imperfect Policy

This subsection explores a second question for the calibrated model: If existing policies are suboptimal, how large is the welfare loss resulting from the imperfect policies? In addition to exploring welfare losses under inefficient deadlines, we also examine a larger set of improperly calibrated subsidies under three classes of policies:

$$\mathcal{W}(\tau_i, \tau_o, T) = \begin{cases} \mathcal{W}(\tilde{\tau}_i(T), -\gamma, T) & \text{(Second Best)} \\ \mathcal{W}(0, \tilde{\tau}_o(T), T) & \text{(No Investment)} \\ \mathcal{W}(0, -\gamma, T) & \text{(Naive Pigou)} \end{cases}$$

Each function corresponds with the total welfare under a certain type of policy. The first type of policy gives the second-best policy for a given T (and achieves the first best when $T = T^*$). It allows τ_i to vary based on the time horizon that the output subsidy runs. The second policy is constrained to only use an output subsidy with a deadline. The welfare under this no-investment restriction does allow τ_o to vary with T to compensate for under-investment. Finally, the third policy is a naive Pigouvian output subsidy, $\tau_o = \gamma$, with a deadline no investment subsidy.

Figure 2.4 reports the results, plotting realized welfare over possible deadlines for each class of policies. The figure reports the achieved welfare of each function given a time horizon T. For all results welfare is scaled so that $W() = 1$ represents the first best allocation net of any compliance costs and the social cost of taxation and that $W() = 0$ represents no investment and no production. Panel (a) shows the case where there are no administrative or compliance costs and Panel (b) includes costs.

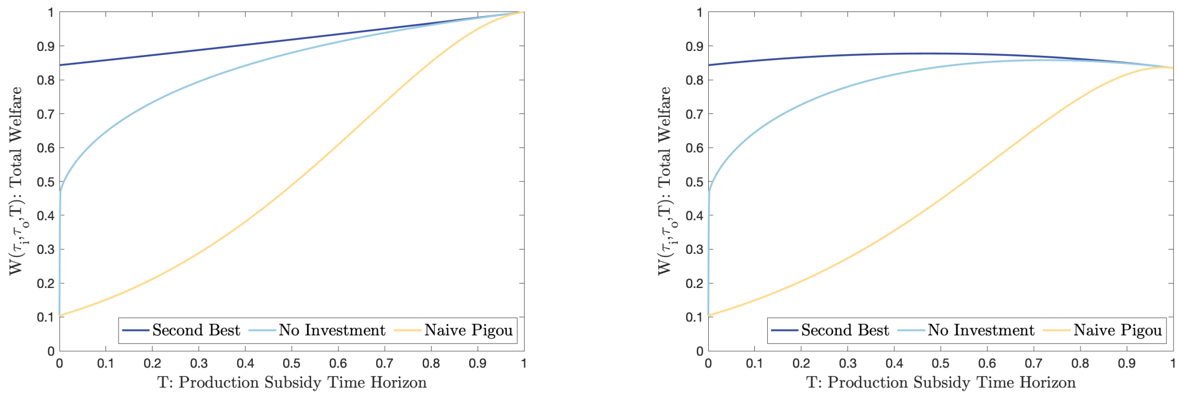
A two features of both graphs are worth noting for interpretation. First, note that all three functions are equal when $T = 1$. If there are no costs this is the efficient solution. Second, when $T = 0$ the three intercepts each have an economic interpretation. The welfare under the second-best at $T = 0$ reflects the best that the social planner could do with only an investment subsidy.[21] The other two intercepts reflect the unsubsidized equilibrium. The difference between the unsubsidized equilibrium and the first best allocation depends on the size of the externality, the costs, and returns to scale.

These panels reveal two main results. First, whenever there is a deadline, the second-best allocation strictly dominates the no investment allocation which in turn strictly dominates the naive Pigouvian allocation. This is true both with and without additional administrative costs of the policy. Note, however, that there are deadlines where the second-best policy is worse than the output-only and naive allocations at longer horizons. Although the order between these curves is fixed, the magnitude of the differences may change. For example, if the change in output, ΔQ, is larger the subsidy with no investment is much worse, falling closer to the naive Pigouvian policies.

The second result is that the qualitative sizes of differences in welfare may be large. In the reported simulations the unsubsidized equilibrium attains about 15% of the welfare possible in the first best. How does this compare to policies in practice? If dollar value of the PTC is close to the average externality for carbon offset by each MWh of wind energy, and if accelerated depreciation is not an effective subsidy, the current policy would be a close analogue to the naive Pigouvian subsidy. If the capital life of wind turbines is 20-25 years, and there is no investment subsidy, this simulation would suggest that the PTC's

[21] In other simulations the height of this intercept depends crucially on the Cobb Douglas parameters: a larger share of the variable input will lower the welfare under the pure investment subsidy.

Figure 2.4: Welfare Losses from Frictions and Suboptimal Policy



(a) No Administrative or Compliance Cost (b) With Administrative or Compliance Costs

Note: This figure shows what percent of total welfare is lost from suboptimal policy. The first-best option (without compliance costs) is normalized to one. The second-best policy chooses τ_i and τ_o optimally given T, the no investment policy chooses τ_o optimally given T and $\tau_i = 0$, and the naive Pigouvian policy chooses $\tau_o = \gamma$ and $\tau_i = 0$ for every value of T. Results from an (unapproximated) Cobb-Douglas calibration $q_t = x^a v^b$. Calibrated values: $a = 0.7$, $b = 0.2$, $c = 0.62$, $m = 0.3$, $\gamma = 0.5$ and in panel (b) $\phi(T) = T^2$.

deadline forgoes 50-60% of the potential social welfare. Interestingly, the ideal investment subsidy for the second best in this calibration is roughly 20%, suggesting that combining the PTC with a program like 1603 grants would have been reasonably close to the second best. Admittedly, the calibration is an oversimplification so not much weight should be put on the exact quantitative results, but the patterns are certainly striking.

In summary, these simulations show that there suboptimal policies can substantially reduce social welfare. The magnitude of these effects depend on which class of policy is pursued as well as on the market fundamentals that drive firm choices.

2.5.2 Implied Social Costs

Having explored welfare under inefficient policies, we conduct an inverse optimum exercise to determine what primitives could rationalize the existing policy as efficient. We focus on the social costs: the cost of raising revenue (λ) and marginal administrative costs ($\phi'(T)$). If the implied magnitudes of λ and $\phi'(T)$ are unreasonable, we can characterize what changes would increase welfare. Currently the production tax credit is an output subsidy worth \$25/MWh, and the PTC deadline of 10 year is 40% of the estimated 25-year lifespan of a wind turbine. The exact value of accelerated tax deprecation benefits as investment subsidies is hard to quantify, but we will assume that it is optimally calibrated for $T^* = 0.40$ to give the current policy the benefit of the doubt.[22]

First, assuming no social cost of raising revenue (i.e., $\lambda = 1$), we find that to justify the current policy, the external benefits of wind energy must be small and the administrative costs much be large. We present this case first because if $\lambda = 1$, we can assess policy optimally in general (i.e., without parameterizing the production function). Because the optimal output subsidy in this case is $\gamma = \tau_o$, optimally requires that $\gamma =$\$25/MWh. If 1MWh of wind energy reduces CO_2 emissions by 0.709 metric tons (as estimated by United States Environmental Protection Agency, 2022), a \$25/MWh benefit implies a social cost of carbon of \$36 per ton. This estimate is low relative to the EPA's estimate (\$51 per ton) and recent academic work (\$59 to \$99 in Cai and Lontzek, 2019); however, computing the true external value of wind is complicated by three considerations. first, there are external benefits to offsetting other pollutants besides CO_2 (raising γ); second, the average offset of 0.709 metric tons per MWh of wind may not reflect the marginal offset (likely decreasing γ in the short run Cullen, 2013); and third, there is heterogeneity across time and space in the amount of carbon offset by one MWh of wind energy (e.g., Hollingsworth and Rudik, 2019; Fell et al., 2021, causing an ambiguous effect). But despite these complications the current value of the PTC is probably too low for $\gamma = \tau_o$ to hold (on average).

In addition to the small implied social benefits of wind, justifying the 10-year deadline when revenue is raised costlessly, requires administrative costs of over \$130 M per year. From Section 3, recall that when $\lambda = 1$, the optimal deadline T is set such that the marginal administrative cost of a longer deadline, $\phi'(T)$, is equal to the marginal external value of increased production from a longer deadline, $\gamma \Delta Q$. If $\gamma =$ \$25 and $\Delta Q =$ 800 MWh for each month for each facility (see Table 2.1), then the marginal administrative cost is \$20,000 per facility per month. In other words, for the current policy to be optimal it must be the case that extending the deadline by one year would cost over \$240,000 per firm.[23] In 2020 578 firms received the PTC, suggesting administrative costs of roughly \$138 million,[24] more than 2.5% of the entire budget outlay for the PTC in 2020. It seems unlikely that this is the compliance burden on the IRS and firms.[25] If costs this large are implausible, the marginal benefits outweigh the marginal costs (even at $\gamma =$\$25), and the current deadline is too short.

In our second set of inverse optimum exercises, we explore the impact from social costs of raising revenue to fund the subsidy programs. Inverse optimum exercises in this context,

[22]Of course, this cannot be true for one policy at all parameter values but may be true at the true values.

[23]This linear approximation of $\phi(T)$ around $T = 0.4$ underestimates the administrative cost because $\phi(T)$ must be convex for an interior solution like $T = 0.4$ to be optimal when $\lambda = 1$.

[24]Whereas the convexity of $\phi(T)$ implies underestimates for extending the policy, it creates an ambiguity in the cross section since there is variation between 0 and 120 months.

[25]We also discuss how $\phi(T)$ may reflect costs of the political process, but as these are difficult to micro-found, we'll focus the political discussion on the cost of raising revenue for public projects.

however, require making assumptions about the shape, curvature, and complementarity of the production function. We use the calibrated Cobb Douglas production function presented in Section 2.5.1 to explore the implications of varying λ. Although this is a highly stylized model, it captures useful intuition about the direction effects will tend to go and plausible relative magnitudes.

For these calibration exercises, we report the implied social costs under a range of externalities and deadline-driven changes in production. We consider both λ, the marginal cost of public funds, and $\phi'(T)$, the marginal administrative costs of extending the subsidy deadline.[26] The implied λ comes from the subsidy size relative to the externality γ. Recall that conditional on the calibrations of the Cobb Douglas parameters a and b, τ_o is decreasing in λ for each value of the externally γ. Because we know the subsidy size, τ_o, we can recover the social cost of raising revenue for each possible γ. Once the revenue cost λ has been identified, the implied $\phi'(T)$ is the value that rationalizes the deadline, T, given the values of γ and λ.

Table 2.2 reports the results from these two exercises, revealing a large range of possible costs and also many combinations that cannot be rationalized at all. Panel A reports the implied values of λ for different values of γ (reported as the social value of one MWh of wind energy) and $\%\Delta Q$. The range of γ spans \$1 to \$100, with relative benchmarks listed to the left of the values. A given cell in Panel A reports the social cost of raising revenue λ that would justify a \$25 production tax credit given an external benefit, γ, and a deadline elasticity $\%\Delta Q$. For example, if $\gamma = \$25$ then the PTC would be optimal if $\lambda = 1$ (for any change in output). If $\gamma = \$35$ as used by the EPA and $\%\Delta Q = 6.5\%$ as estimated in our paper, then PTC would be optimal if $\lambda = 1.13$.

Panel A shows three important patterns about the social costs of taxation that would justify the PTC. First, if the social value of wind energy is less than \$25, then the current PTC is only justifiable if $\lambda < 1$, i.e., if taxation creates welfare gains. Second, the implied λ is increasing in γ and increasing in ΔQ (where $\gamma \geq \tau_o$). It is increasing in γ because if the benefits of subsidization are large, then a (relatively) small output subsidy can only be justified if there are large costs of raising the funds for higher subsidies. It is decreasing in ΔQ because if the responsiveness to τ_o is bigger, having τ_o be smaller can only be justified if there are larger costs of raising the funds. Third, at higher values of γ, such as the higher values in Cai and Lontzek (2019), the λ would need to be quite large in order to justify the current level of τ_o.

Panel B shows a similar set of relationships for the implied marginal administrative costs

[26] Because the relationship between γ and λ is linear conditional on a and b, this is equivalent to calibrating a range of λ and exploring the implied γ and $\phi'(T)$ as we did for the general case above.

that would rationalize $T = 0.4$, given γ, ΔQ and λ. This panel can be interpreted in the same way with one change. Rather than report $\phi'(T)$, which does not have very intuitive units in the calibration, we report the $\phi'(T)$ as a percent of total welfare. One way to think of this is given the slope, what percent of (average) yearly welfare would be lost to administrative costs by extending the output subsidy for an additional year. For example, if $\gamma = \$25$ then the PTC deadline would be optimal if extending T for a year would cost between \$25,000 -\$100,000 per firm in administrative costs and accounted for 0.4-1.7% of welfare.[27] If $\gamma = \$35$ as used by the EPA and $\%\Delta Q = 6.5\%$ as estimated in our paper, then PTC would be optimal if the administrative costs associated with extending the subsidy by a year represented 1.3% of welfare.

Just as some values of the external benefit γ that produced counter intuitive social costs like $\lambda < 1$, there are others that imply impossible $\phi'(T)$. This can happen in two ways: some values imply $\phi'(T) < 0$ and other imply that subsidies reduce welfare relative to the unsubsidized equilibrium. In either case the observed policy can't be justified. In practice, these implications bound the set of γ that rationalize the deadline. Specifically, given ΔQ in the range of our estimated confidence intervals, the set of possible external benefits is a subset of $\gamma \in [\$25, \$45]$. Values as high as $\gamma = \$45$ can only be rationalized when the change in quantity is large. This has two implications, depending on what values of γ one finds plausible. On one hand if the benefits from producing wind energy are really of this size, the rationalizing λs are fairly reasonable (if anything $\lambda \in (1.0, 1.3)$ is perhaps a little low), but the implied administrative costs are quite large (as high as \$255,000 per firm per year). On the other hand, if the benefits from producing wind energy are really as large as research like Cai and Lontzek (2019) suggests, then the subsidy size, τ_o, and the deadline length, T, are too low. In fact, there are no costs large enough to justify the combined policy.

Taken together these results suggest that it is unlikely that the existing combination of a production subsidy with a deadline and tax benefits for investment are achieving the efficient allocation. Even though the production response at the deadline suggests a relatively small elasticity that would justify a shorter deadline all else equal, the value of the externality is large enough that an longer deadline (or no deadline) would result in a better allocation. This finding is reminiscent of the insight from Aldy et al. (2021) that "output subsidies are more cost-effective than investment subsidies over a large range of output targets." We find that in this industry output subsidies with deadlines also underperform those without—despite the fact subsidies with no deadlines pay for many more inframarginal units.

[27] Assuming zero profits, the total welfare is the total quantity times the social value. On average facilities in the RD sample produce roughly 18,500 MWh/month, so we obtain this range by scaling that up to welfare of a year's production multiplied by the relevant percent.

Table 2.2: Inverse Optimum Exercises Require Large Social Costs to Rationalize Policy

Panel A: Social Cost of Raising Revenue		Change in Production:				
Reference Policy	Social Value of 1 MWh	$\%\Delta Q = 2.5\%$	$\%\Delta Q = 4.5\%$	$\%\Delta Q = 6.5\%$	$\%\Delta Q = 8.5\%$	$\%\Delta Q = 10.5\%$
Trump	\$ 1.00	0.84	0.76	0.69	0.64	0.60
PTC	\$ 25.00	1.00	1.00	1.00	1.00	1.00
	\$ 30.00	1.03	1.05	1.06	1.08	1.08
EPA	\$ 35.00	1.07	1.10	1.13	1.15	1.17
Min CL (2019)	\$ 40.00	1.10	1.15	1.19	1.23	1.25
	\$ 45.00	1.13	1.20	1.26	1.30	1.33
	\$ 50.00	1.16	1.25	1.32	1.38	1.42
	\$ 55.00	1.20	1.30	1.39	1.45	1.50
Max CL (2019)	\$ 70.00	1.30	1.46	1.58	1.68	1.75
	\$ 85.00	1.39	1.61	1.77	1.90	2.00
	\$ 100.00	1.49	1.76	1.96	2.13	2.25

Panel B: Social Cost Extending Deadline		Change in Production				
Reference Policy	Social Value of 1 MWh	$\%\Delta Q = 2.5\%$	$\%\Delta Q = 4.5\%$	$\%\Delta Q = 6.5\%$	$\%\Delta Q = 8.5\%$	$\%\Delta Q = 10.5\%$
Trump	\$ 1.00	-2.40%	1.23%	11.47%	32.16%	68.15%
PTC	\$ 25.00	0.41%	0.75%	1.08%	1.42%	1.76%
	\$ 30.00	0.52%	0.94%	1.38%	1.83%	2.29%
EPA	\$ 35.00	0.44%	0.84%	1.29%	1.79%	2.31%
Min CL (2019)	\$ 40.00	0.19%	0.44%	0.77%	1.20%	1.69%
	\$ 45.00	-0.19%	-0.27%	-0.23%	-0.07%	0.22%
	\$ 50.00	-0.71%	-1.30%	-1.80%	-2.18%	-2.40%
	\$ 55.00	-1.35%	-2.66%	-4.02%	-5.37%	-6.62%
Max CL (2019)	\$ 70.00	-3.92%	-8.93%	-15.85%	-25.68%	-39.84%
	\$ 85.00	-7.35%	-19.01%	-40.19%	-87.40%	-259.53%
	\$ 100.00	-11.51%	-33.84%	-90.64%	-451.70%	Negative Welfare

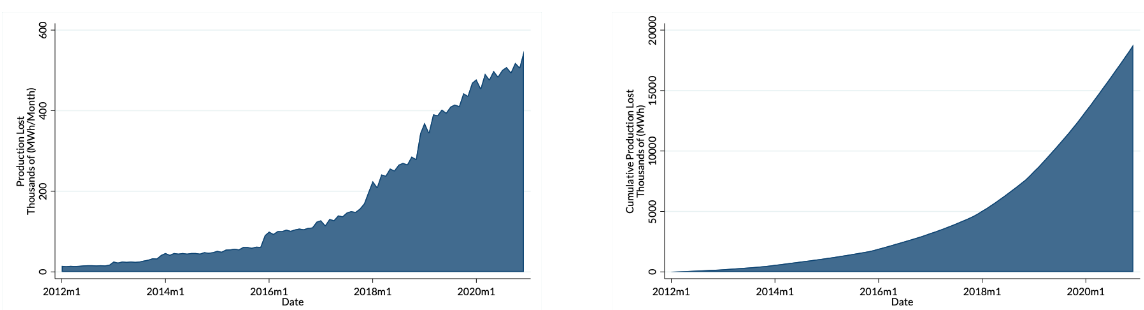
Note: This table reports calibration results calculating what social costs justify the current PTC policy. Panel A reports values of the the marginal cost of public funds that rationalize the \$25 production tax credit under different values of the external benefits of one MWh of wind (γ) and changes in production at the deadline ($\%\Delta Q$). These values are reported in terms of λ, the social cost (in dollars) per dollar of revenue raised. Panel B reports the administrative costs that would rationalize the 10-year deadline under different values of γ and $\%\Delta Q$ (and under the λ implied in Panel A). These costs are reported as the cost of running the subsidy for another year as a percent of total welfare.

2.5.3 Foregone Wind Energy Production

Having discussed the implications of the PTC deadline for optimal policy, we turn to also exploring its implications on energy markets. Wind facilities are a quickly-growing feature of energy markets in the United States, and between 2022 and 2026 there are over 28700 MW of wind capacity that will cross the PTC deadline and lose eligibility for the subsidies. This subsection explores the magnitude of deadline-induced changes in production on total energy generated by wind.

We quantify the dynamic production response attributable to the PTC deadline with a simple extrapolation exercise using the regression discontinuity estimates. For each month that firms produce after the deadline, we assume that their average capacity factor would have been 2.7 percentage points higher had the deadline not been in place. In doing so, assume that the changes in outputs come from the deadline. If anything, this will underestimate the total effect to the extent to which changes in slope also reflect causal effects of the deadline, (estimated with some imprecision in our paper, but with more clarity in Hamilton et al., 2020). The effects estimated in our paper capture the average effects in the years through 2021. We do not project these estimates forward in time because we find suggestive results that the response to deadlines is smaller in more recently constructed facilities (although there are more firms so the total effect may still be increasing).

Figure 2.5: Energy Production Foregone Because of the PTC Deadline



(a) Monthly Production

(b) Cumulative Production

Note: This figure shows the energy production that was lost from the PTC deadline. To calculate these estimates, we multiply facilities' nameplate capacity by the 2.7-percentage-point production response at the deadline and sum up the total effect each month. As such these estimates only capture the production lost along the intensive margin, not accounting for firm entry, exit, pr investment decisions.

Figure 2.5 reports the results. In the panel on the left we see that by December 2020, energy markets forgo 548,000 MWh/month of energy produced by wind. This corresponds to the power used by 613,000 homes (using the EIA's estimate that the average household uses 0.893 MWh/month United States Energy Information Administration, 2022) and a social externality value between \$13.7 and \$37.8 million dollars per month. The panel on the right shows the cumulative total foregone energy production from the PTC deadline. By the end of 2020, the energy market had 18 million MWh of foregone energy from wind, corresponding to between \$450 million and \$1.2 billion of lost external benefits.[28]

2.6. Conclusion and Policy Discussion

In this paper we show how output and investment subsidies can relate in the context of subsidies with deadlines, revealing that deadlines are often efficient deviations from Pigouvian policy. We also show that the change in production at the deadline must inform the optimal deadline length, and when estimating this change in production in the US wind energy industry, we find that the end of the production tax credit results in an 8% decrease in production—implying nearly 15 million MWh of foregone energy produced by wind in the last five years. Because only unreasonably large administrative costs of the program or surprisingly small values of the social externality can justify the current 10-year deadline, we conclude by considering ways that policy makers might restructure the PTC and how other subsidies may be designed in the future to better target efficient levels of alternative energy production. We focus on four main points: the current calibration of output subsidies, the current deadline lengths, the importance of investment subsidies when there are deadlines, and the implications of technological change over time for the subsidy changes.

Before making our recommendations, we want to emphasize that overall many features of the current set of subsidies reflect important aspects of the optimal theoretical results we derived. For example, the presence of both tax benefits for investment and an output subsidy with a deadline is exactly what the optimal policy with a deadline would require. We also think that given a 30%, rule-of-thumb social cost of raising tax revenue ($\lambda = 1.3$), a \$25 output subsidy may be close to appropriately calibrated—even if the social value of each MWh produced by wind is higher than \$25. Furthermore, the intuition of having a deadline when the production response at the deadline is small is reflected in practice with most of the deadlines focused in investment-intensive industries (with small production responses).

We make four recommendations for ways of improving future policy with research or in practice. First, the critical role that the external benefit has on the subsidies underscores the importance of quantifying externality sizes. Our results show that whether the subsidy has deadlines or not, the optimal output subsidy is always closely related to the externality. As such, whether considering a policy for subsidizing alternative energy, housing investment, or research and development, policymakers should carefully consider external benefits and

[28]Note that neither of these external benefit calculations are the welfare loss as there was no reduced tax expenditure, and firms earned fewer revenues but spent less in costs.

how they may vary. For example, rather than have uniform electric vehicle subsidies, policy makers might want to subsidize purchasing more in markets where electricity is produced from cleaner means (for reference see external benefits quantified in Holland et al., 2016).

Our second suggestion is the policy makers make externality-specific investment subsidies. Our theoretical results show the critical role that investment subsidies play in the presence of subsidies with deadlines. When an output subsidy has a deadline the only way to avoid under-investment is to couple it with an investment subsidy. Not even raising the size of the output subsidy can reach the efficient outcome. In the United States tax benefits subsidize investment in a sense, but are complex and prone to nullificaiton by other policies. For example because subsidies are given relative to a baseline of scheduled depreciation, it difficult to articulate the exact subsidy amount and to assess its optimality. Furthermore, common policies that offer these benefits to other industries remove the effective subsidy. For example, the 2020 CARES Act gave accelerated depreciation for almost all investments—essentially removing the marginal incentive to invest in capital for production positive-externality goods. One policy that could enable better synergy between the investment subsidies and subsidies with deadlines could be to replace tax benefits like accelerated depreciation with actual investment subsidies. Additionally, instead of forcing facilities to choose between output and investment subsidies (as with wind or geothermal energy) they should be allowed to claim both (as with subsidies for low income housing, healthcare or research and development). Even targeting revenue neutrality rather than efficiency, a change like this would facilitate welfare-improving adjustments with the three policy levers in the future.

Third, for the production tax credit, our results suggest that for many energy technologies should have longer—and different—deadlines. Given the nature of the market, and the costs of raising public funds through normal means, it is extremely likely that it is optimal for alternative energy subsidies to have deadlines. The fact that most sources of renewable resources are fueled by natural phenomena like wind, sun, waves, water, and geothermal heat means that most industries feature (relatively) small margins of to change production at the deadline—exactly the conditions under which deadlines tend to be justified. Our results suggest that a 10-year deadline is too short for wind; furthermore, we are skeptical that having deadlines which are the same length for both renewable energy resources and biomass-based resources is optimal. This would suggest that both industries have equal social externalities (even though burning biomass emits carbon and can emit particulate matter) and equal changes in production (even though biomass resources have fuel inputs that are not free or exogenous). We suggest that deadlines for alternative energy production be longer and that deadlines vary by externally size and response size.

Finally, we suggest that policy makers should consider to phasing out subsidies using the deadline rather than the output subsidy amount in industries where technical change leads to changes in the effects of the deadline over time. Our theoretical results show that when a deadline is optimal, the length of the deadline is increasing in the production response at the deadline. Empirically we showed that facilities using newer vintages of turbines may be less responsive to the deadline than older vintages (and results about pre-deadline changes in production suggest a similar pattern may emerge for future cohorts, see Hamilton et al., 2020). To the extent that technological innovations or changes in siting patterns drive these changes, it would be optimal to have longer deadlines in earlier years relative to later years. Current policies aimed at phasing out expenditures on the PTC reduce the amount of the subsidy τ_o. Our results suggest that an equivalent reduction in expenditures effected through shortening the time horizon T rather than τ_o would promote more welfare.[29] This same intuition could extend to other subsidies such as adding a deadline to the research an experimentation tax credit rather than the current practice of relying on the changing basis to avoid subsidizing inframarginal units.

These policy points show that while there are real stakes to implementing subsidies with deadlines, there are also substantial opportunities for gains. Subsidies with deadlines connect both output and investment subsidies, and the optimal deadline length is deeply related to the size of externalities and to the change in production at the deadline. In the US wind energy industry, there is a relatively small 8% decrease in production. Considering that deadlines matter in the wind industry, which is already a limiting case, we believe carefully considering deadlines will matter even more in other industries where the changes in production are likely much larger. With information about the size of the social externalities and estimates of firm responses, we can improve subsidies with deadlines to help them better target the goods they are designed to support.

[29] For this argument to hold in general, costs and external benefits must be constant across cohorts and tax benefits for investment must adjust to the shortening deadline.

CHAPTER III

Public Good Perceptions and Support: Evidence from Higher Education Appropriations (with Reuben Hurst and Andrew Simon)

3.0 Abstract

We conduct a survey experiment to understand how perceptions of public good provision affect support for public spending. Using US state spending on public higher education as a concrete example, we randomize taxpayers into three groups. One group receives information about their state's graduation rate, a second receives information about the rate and the state's rank relative to other states, and a third that receives no information. After receiving information, taxpayers prefer on average 5 percent more spending on higher education. The effect is driven by Republicans, by the elderly, and by taxpayers who learn that their state produces more graduates than they previously believed. As a result, information effectively and inexpensively decreases polarization.

3.1. Introduction

Taxpayers express their support for public spending on a variety of services, like healthcare, education, policing, infrastructure, and defense. If taxpayers trade off the value of the services and the cost to finance them, perceptions about the effective price of the services and return on investment will shape their expressed preferences. Variation in perceptions may create a wedge between efficient and implemented policies by distorting preferences and may also exacerbate polarization between groups. In this case, information could effectively and inexpensively correct suboptimal policies and reduce political division.

This paper studies how information about the provision of local public goods affects support for public expenditure. We consider public investment in higher education as a concrete example because it is the third largest budget area of US state spending (Laderman and Heckert, 2021).[1] We study how information about the state's graduation rates—in absolute terms and relative to other states—affects preferences for state spending on higher education. Since public colleges receive tax revenue from their state governments to produce a college-educated workforce, we report graduation rates, which reflect both the level of provision and the social return on investment. Additional information about the state's rank allows taxpayers to learn about their state's relative productivity and provision.

To understand how perceptions of public higher education provision affect support for public spending, we conduct a survey experiment with a national sample of taxpayers. After documenting the variation in taxpayers' perceptions, we shape those perceptions by randomly assigning taxpayers to one of two information treatments. The first presents taxpayers with the graduation rate in their state and the second presents them with the graduation rate and also tells them their state's rank relative to other states. Randomization allows us to estimate the causal effect of each type of information on preferences for public spending and how these effects vary across taxpayers based on their prior beliefs and characteristics like age and political identity. Since the effects depend on how taxpayers update their perceptions, our design also sheds light on how taxpayers reason about public good provision.

A simple price-theoretic model motivates our focus on the outputs of public good provision (e.g., graduation rates). Taxpayer demand for public spending is a function of the marginal benefits to them and the effective price of producing graduates—objects that the level of output directly informs. Providing information about the public good outputs is novel relative to existing experiments that usually provide information about inputs—for example education spending (Howell and West, 2009; Lergetporer et al., 2018; Lergetporer and Smarzynska Javorcik, 2019; Lergetporer et al., 2020; Giaccobasso et al., 2022) and general public spending (Cruces et al., 2013; Kuziemko et al., 2015; Haaland et al., 2020; Peyton, 2020; Roth et al., 2021).[2] However, the model shows that without information about outputs, spending levels convey ambiguous information because it conflates information about the effective price with information about the level of provision.

In the model we consider different ways information may affects preferences. Any change in preferred spending must come from shifting the perceived marginal cost or marginal benefit curves. The perceived price of raising the graduation rate determines perceived marginal cost curve, and perceptions about graduation rates in other states determine the perceived marginal benefit of having a more highly ranked state (often called "yardstick" competition,

[1] Medicaid and K-12 education are larger.

[2] Survey experiments have also been useful for understanding policy attitudes in general, like immigration (Alesina et al., 2018), foreign aid (Hurst et al., 2017), affirmative action (Haaland and Roth, 2021), the gender pay gap (Settele, 2019), and racial gaps (Alesina et al., 2021), among others.

Besley and Case, 1995). Information may affect both of these perceptions. Perceptions about the level of provision in the status quo, however, will not affect preferences as they represent movements *along* the cost and benefit curves. If preferences respond to information, the perceived marginal costs and marginal benefit functions must be shifting.

We find that correcting perceptions with information does change preferences, increasing support for spending on public higher education. On average taxpayers who learn only the graduation rate and those who learn both the graduation rate and their state's relative rank demand 5% higher expenditures ($600 per student). Furthermore, taxpayers who receive either information treatment are about 7.5% more likely to donate some of their wages to public colleges in their state ($p = 0.032$), and 16.5% more likely to write state representatives to request more spending on higher education ($p = 0.106$).

We identify two mechanisms in the model for how information may shift demand for spending: heterogeneity baseline perceptions, and differences in how taxpayers reason about the same information. First, since taxpayers have varied priors, the same graduation rate will send qualitatively different messages to different taxpayers. We show the importance of baseline perceptions by comparing the effects of information on taxpayers who underestimated and overestimated the graduation rate in their state. We find that receiving either information intervention increases support among taxpayers who underestimated graduation rates by 6-9% and has no effect on taxpayers who overestimated them. Similar patterns emerge when comparing individuals who underestimated and overestimated their state's rank. Interestingly, we reject symmetry for both treatments, revealing that information content is not the only mechanism through which information changes preferences.

The second mechanism is that certain taxpayers may reason differently about the same information. In fact the same information may lead different taxpayers to different conclusions about graduation rates, ranks, prices, and the provision in other states. For example, a taxpayer who learns that the graduation rate is higher than expected might reason that the price is lower than she originally thought or that the state is spending more than she originally thought. We also show that different groups of taxpayers do reason differently by estimating heterogeneous effects by taxpayer identity. Although changes in preferred spending are driven by Republicans and the elderly, on average their information content is almost identical with other taxpayers'. We also find strong heterogeneity in reasoning by political identify. Whereas Republicans' perceptions of yardstick competition respond to information, their perceptions of price often do not. On the other hand, Democrats' and Independents' perceptions of price do respond to information, but their perceptions of yardstick competition do not. These results complement a growing line of research seeking to understand how partisan differences affect how people reason about tax and expenditure programs (Gaines et al., 2007; Nyhan and Reifler, 2010; Stantcheva, 2020).

Regardless of the mechanism through which it operates this heterogeneity among groups demonstrates that providing information about public good provision reduces political polarization with respect to preferred spending levels. Among taxpayers in out sample Republicans and the elderly prefer low expenditures on education (and may actively oppose spending, see also Ehrenberg, 2008; Mettler, 2014; Taylor et al., 2020; Imlay, 2021). Information increases support, shrinking the partisan gap in preferred spending by 32% and eliminating the gap by age.[3] While polarized perceptions do alter policy views (Alesina et al., 2020), the fact that individuals often prioritize party over policy (e.g., Achen and Bartels, 2017; Barber and Pope, 2019) and engage in politically motivated reasoning (e.g., Epley and Gilovich, 2016; Su, 2022) produce pessimism about every overcoming polarization. We show, however, that providing credible information reduces polarization in perceptions and preferred policies.

Finally, our application on the political economy of higher education funding also informs broader conversations about education finance and fiscal federalism. States are not investing in public higher education despite the large private and social returns of increasing graduation rates (Bound and Turner, 2007; Bound et al., 2010; Moretti, 2004; Deming and Walters, 2017).[4] Economic explanations include budget crowd-out by Medicare (Kane et al., 2005) and decreased return on state investment due to increased graduate migration (Bound et al., 2019). Our results show that taxpayer support is, in fact, sensitive to perceived price and return on investment but that current perceptions are leading to inefficiently low levels of provision. This adds to a larger conversation on fiscal federalism (Oates, 1972, 1999) where other work has shown that inaccurate perceptions lead stated collective preferences to be different from socially efficient policies (Althaus, 1998; Gilens, 2001).

3.2. The Impact of Information on Preferences for Spending

This section presents a price-theoretic model to describe how taxpayers' preferences form. The model shows how preferences are shaped by perceptions and how changing perceptions about either the effective price of raising the graduation rate or the nature of yardstick competition may change preferences. It also discusses two main mechanisms for how taxpayers information may affect preferences for public good provision: baseline perceptions and heterogeneous reasoning.

As we discuss support for public goods, we focus on state colleges and use graduation

[3]These taxpayers are also less likely to state that they prefer \$0 in public higher education expenditure, implying information especially reduces extreme polarization.

[4]Bound and Simon (2021) also show that the private returns are not just for public college students since public investment affects the private college and labor markets as well.

rates as the measure of provision. Public colleges produce a skilled labor force for the state, promoting growth, raising students' wages, and generating tax revenue. Our goal is to consider a simple measure of public good "output" that is related to inputs (tax expenditures). We focus on the six-year graduation rate at public four-year colleges for three reasons. First, it reflects the level of provision, i.e., total number of graduates, in a way that is relatively straightforward for taxpayers to interpret. Second, state spending produces graduates by increasing enrollment and graduation rates (Bound et al., 2010; Deming and Walters, 2017). And third, for a given level of spending, the graduation rate reflects the average social return on investment: the number of graduates produced per dollar of tax revenue spent.[5] A higher rate indicates that more graduates are produced from a given level of spending and that the state is more productive and less wasteful.

3.2.1 A Model of Perceptions and Preference Formation

We begin by modeling preference formation. Assume agents have a concave utility function, $u_i(c, G, R)$, over consumption, c, and a public good composed of the graduation rate in their state, G, and their state's (percentile) rank of $R = F(G)$, for a differentiable distribution function $F()$. Consumption c is related to graduation rates through tax rates, implying a cost function $p(G, \omega)$ that captures tax liabilities, based on the state's productivity ω. Note that total spending is an aggregation of individual cost burdens, $S = \int_i p(G; \omega)$. Individuals have information set $\mathcal{I} = \{G, R, \omega, F\}$.

An individual's stated preference, $G^*(\mathcal{I})$, will satisfy the following first order condition with respect to G:

$$v_G(G^*(\mathcal{I})) + v_R(F(G^*(\mathcal{I})))f(G^*(\mathcal{I})) = u_c p'(G^*(\mathcal{I}); \omega) \tag{3.2.1}$$

where $v(c(G), G, F(G))$ is the indirect utility, partial derivatives are denoted by subscripts, $f(G) = F'(G)$ is the perception of the probability density function of G, and $p'(G; \omega)$ is the perceived marginal price of increasing G.

The two terms on the left hand side of the first order equation capture the marginal benefit from additional graduates. For any value of G the $v_G(G)$ term measures the direct marginal benefits of higher graduation rates, and $v_R(F(G))$ measures the marginal benefit of a higher rank. This $v_R(F(G))$ is scaled by $f(G)$ the perceived rank increase from a marginal increase in G. Individuals with higher marginal utilities of G or R (evaluated at a given point) will prefer more G^*, all else equal. Similarly, individuals who believe $F(G)$ is lower or $f(G)$ is higher will value a marginal increase in G change more. These terms trace out a

[5] A measure also used in Conzelmann et al. (2022) for the social return on investment.

marginal benefits curve over G.

On the right-hand side of the first order equation is one term that captures the marginal cost of producing graduates. This term is the product of the marginal utility of consumption $v_c(G)$ and the perceived marginal price of increasing the graduation rate $p'(G;\omega)$. This product captures the marginal utility of consumption forgone to pay taxes for a marginal increase in G. Individuals with higher marginal utility of consumption (evaluated at a given point) will prefer lower graduation rates, as will those those who think the effective price is high because the state is less productive. As with the marginal benefits, this side of the equation traces out a marginal cost curve that is increasing in G.

3.2.2 The Impact of Information on Perceptions

We consider how providing information affects perceptions and preferences, considering two mechanisms: (1) the role of providing information, and (2) the role of reasoning about information. We denote, prior and posterior perceptions of the public good process as $\tilde{\mathcal{I}} = \left\{\tilde{G}, \tilde{R}, \tilde{\omega}, \tilde{F}()\right\}$ and $\hat{\mathcal{I}} = \left\{\hat{G}, \hat{R}, \hat{\omega}, \hat{F}()\right\}$, noting that either may be inaccurate. Note that learning about provision will not affect policy preferences, but learning about productivity and provision in other states will.

Figure 3.1 depicts the role of information and reasoning on preferences. Consider an individual who learns that the graduation rate is higher than initially perceived ($G > \tilde{G}$). She will form a posterior perception $\hat{G} = G$, but this change could come through two channels. On one hand, she may conclude that the state is spending more tax revenue than anticipated. Panel (a) depicts this a movement *along* the cost function $p(G;\omega)$ since the function (indexed by ω) does not change. In this case information does not change her preferred graduation rate or spending level. On the other hand, she may learn that the state is more productive at producing graduates from tax revenue. Panel (b) depicts the resulting shift of the marginal cost curve. If this is the only movement, her preferred graduation rate increases (although the effect on spending level is ambiguous).

Figure 3.1 also depicts the role of learning about rank. In the example from the previous paragraph, the information about (only) G will also lead her to update her beliefs about R. On one hand she may conclude that the distribution of graduation rates is the same but that her state's rank (place in the distribution) is higher. This movement *along* the distribution function $F()$ is implicit in the depiction in Panel (a). Alternatively, she could conclude that all states' graduation rates are higher, a (first-order stochastically dominant) shift of distribution functions $F()$. Panel (c) depicts the resulting shift of the marginal benefit curve. If this is the only movement, her preferred graduation rate and spending level both increase. Information may change perceptions of the variance of $F()$ as well, which will make the marginal benefit curve flatter or steeper with ambiguous effects on preferred graduation rates, with an example shown in in Panel (d).

Because information may change marginal costs and marginal benefits, the causal effect of receiving information on preferences for spending reflects these changes in taxpayers' perceptions. For example, an individual who learns that the graduation rate is higher than initially perceived ($G > \tilde{G}$) will prefer a higher graduation rate if the perceived price is lower (downward shifted marginal cost curve)or if the perceived distribution of graduation rates is higher (upward shift in marginal benefits). But whereas a higher perceived distribution of graduation rates will increase the preferred spending level, a lower perceived price has an ambiguous effect. On the other hand, an individual who learns that the graduation rate is lower than initially perceived ($G > \tilde{G}$) will only prefer more spending if they are reasoning through the price channel.

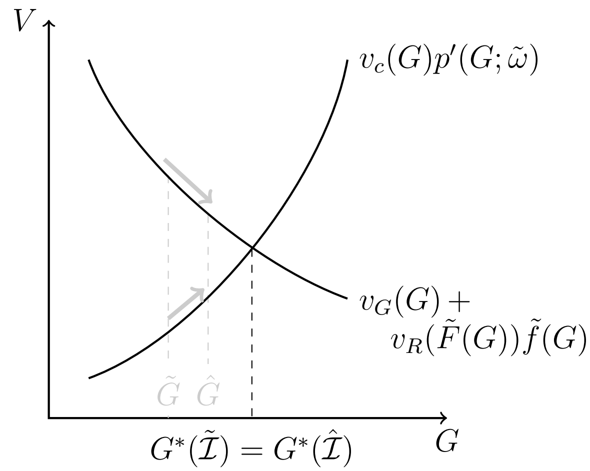
Individuals who receive information about both the graduation rate and their state's rank must update their perceptions. For example, if the a taxpayer who learns $G > \tilde{G}$ simultaneously learning the value of R, will change perceptions of $F()$. For example, learning $G > \tilde{G}$ and $R < \tilde{R}$ increases $F()$, increases G^*, and increases the preferred spending level all else equal; however, if she reasons that $R < \tilde{R}$ means the state is less productive or the variance changes, the overall effect will be ambiguous. All of the pieces are also ambiguous if $R > \tilde{R}$.[6]

The effects of information on posterior perceptions may be heterogeneous in two dimensions. First, since taxpayers have varied priors, the same graduation rate (and rank) communicates qualitatively different messages to different taxpayers. Second, depending on how taxpayers reason, even the same information may lead different taxpayers to draw different inferences about graduation rates, ranks, prices, and the provision in other states. The final effect on perceptions will be a result of the distribution of priors and the heterogeneity in reasoning. The final effect on preferences will also depend on idiosyncratic preferences for consumption and for the public good.

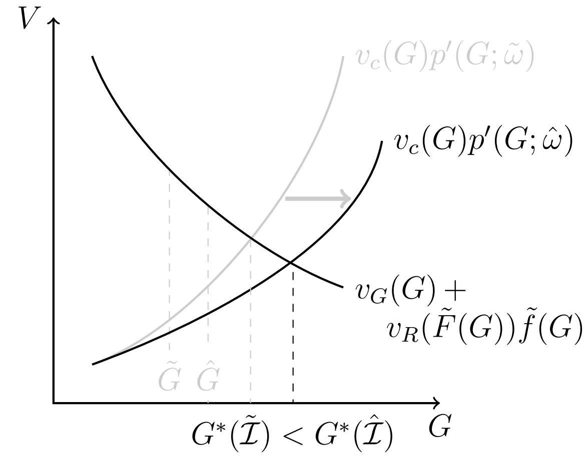
The overall effects of information on preferred spending depend on the extent to which prior perceptions vary and on any heterogeneity in how information affects their posterior perceptions of the public good. In the remainder of the paper, we seek to understand the relative magnitude of these effects across different types of taxpayers and to understand the mechanisms driving these effects. These comparisons are made possible by our survey experiment.

[6]Although the additional rank information has no direct bearing on the absolute value of $p()$, it certainly could affect perceptions about the distribution of $p_{-s}()$ in other states.

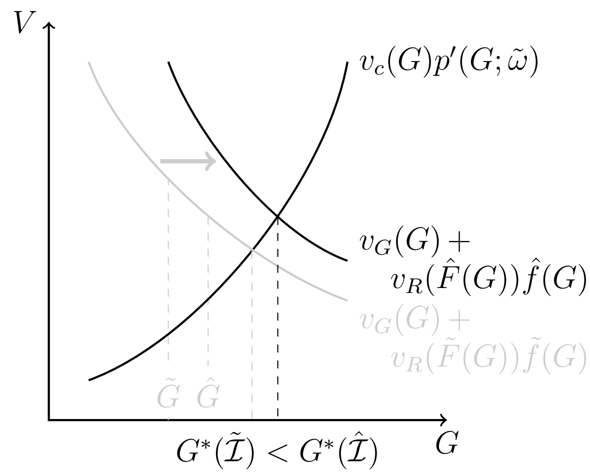
Figure 3.1: Information, Belief Updating, and Support



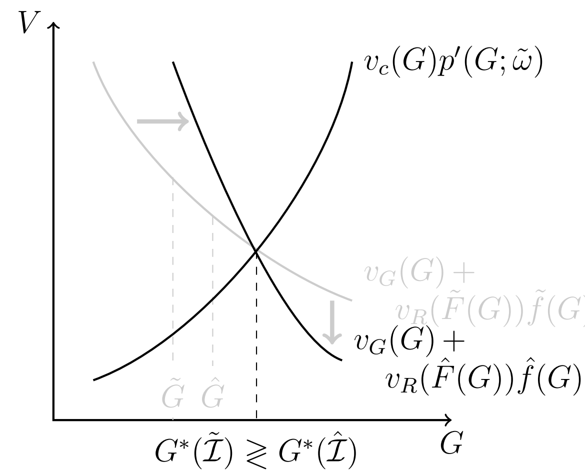
(a) Spending Channel: Moving Beliefs Along Marginal Cost and Benefit Curves



(b) Price Channel: Productivity Shifts Marginal Cost Curve



(c) Rank Channel: National Average Shifts Marginal Benefit Curve



(d) Dispersion Channel: Variance Across States Rotates Marginal Benefit Curve

3.3. Survey Experiment

In this section, we outline the design of our experiment. In the experiment we randomize taxpayers into receiving information about absolute and relative graduation rates from public four-year colleges. This allows us to measure the causal effect of information on support for state spending and to test the underlying mechanisms. The experiment and pre-analysis plan were registered with the AEA RCT registry.

3.3.1 Sample

We initially recruited 3715 participants through Qualtrics in May 2021. Specifically, we requested a sample pool that was broadly representative of the United States population in terms of political partisanship (a third of each Democrats, Republicans, and Independents), education level (41 percent high school equivalent or less, 29 percent some college or associate's degree, and 30 percent bachelor's degree or higher), and gender. We did not include participants from Washington, D.C. since it does not have a traditional public college system. However, since this initial sample was not representative on age, we recruited an additional 1004 participants in August 2021 using CloudResearch.

3.3.2 Experimental Procedure

After consenting to participate, participants answered demographic screening questions measuring political partisanship, gender, and state of residence.[7] Next, participants reported their beliefs regarding the US average six-year public college graduation rate, the graduation rate in their state, and the rank of this graduation rate relative to the other 49 states. Participants then additionally provided their prior beliefs of the tax expenditure spent per student on public higher education in their home state and its rank relative to other states.

After reporting their perceptions, participants were randomized into one of three treatment groups. The first treatment presented them with the actual six-year graduation rate at public colleges in their state, based on data from the National Center for Education Statistics. The second treatment presented them with the state's graduation rate and its rank relative to other states. In both information treatments, participants were also reminded of their prior beliefs and shown the difference between that prior and the truth. The third group was a

[7] We elicit age, employment status, and connection to the public higher education system at the end of the survey. To measure connection, we asked whether or not the respondent has had (or does have) children attending a public college in their state, whether they believe their children will attend a public college in their state in the future, and whether they themselves have attended or are currently attending a public college in their state.

control and did not receive any information. Appendix C.1 shows that our treatments are balanced on both taxpayer characteristics and beliefs.

After exposure to the information treatment, participants answered a range of survey questions. First, they were asked to indicate their preferred level of annual tax revenue per student at public four-year colleges in their state, "acknowledging the potential effects of the COVID-19." They were then asked to give their preferred level of spending "after the pandemic has passed." Since the pandemic presented exceptional challenges for governments, universities, and taxpayers, we focus on preferences for spending after the pandemic has passed.[8]

Participants also answered survey questions gauging their general trust in the way public four-year colleges use tax revenue, the extent to which students should bear the cost of funding public higher education, and whether or not the federal government should provide financial support for public four-year colleges. These questions were measured on a five-point Likert scale.

In addition to allowing taxpayers to state their preferred policies, they were given the opportunity to participate in two costly tasks to elicit revealed preferences. Although hypothetical and real responses on intensive margin decisions like the ideal level of spending tend to be very similar (List et al., 2006), we use the revealed preference measures as a robustness check against "hypothetical bias" (List and Gallet, 2001; List, 2001).[9] First, participants were invited to share written opinions about their state's spending on public four-year colleges with the state representatives of their choosing using an open-ended response format.[10] Spending time to send an optional message without added incentives is costly since it lowers effective wage for participating in the survey. This exercise also allows the researcher to observe taxpayers' first-order concerns (Ferrario and Stantcheva, 2022). Sending messages also has policy significance when politicians update policy positions to reflect voters' preferences (Sevenans, 2021).

As a second costly task, participants were given \$0.25 which they could keep or donate to a public college in their state. This amount is non-trivial. The median participant spent 6 minutes in total on the survey, so keeping the \$0.25 would increase their effective hourly

[8]Preferences across these two measures are highly correlated and the effects of information are similar. One exception is that Democrats and Independents state slightly lower preferences for spending during the pandemic than after it has passed, while Republicans ideal and COVID-19 spending levels are more similar. See Appendix C.2 for more details.

[9]Among our stated preference measures, we focus on how randomized information provision affects preferred spending levels. Responses are along the intensive margin because preferred spending is strictly positive for the vast majority of the sample.

[10]Possible representatives include the governor and the Republican and Democratic leaders in their state legislatures

wage by \$2.50. There is a strong relationship between our stated and revealed preference measures; taxpayers who are willing to donate or who tell their representatives to invest more in public higher education state higher preferred spending levels in the survey.[11]

3.3.3 Variable Definitions

We define our variables following the preanalysis plan posted to the AEA RCT Registry.

Belief Errors: We calculate a continuous variable measuring the belief error as the patricipants' priors minus the truth for their state's graduation rate and percentile rank as well as spending levels and percentile rank.[12] A negative error means the prior perceptions are lower than the truth. We observed the state graduation rates and spending levels from Snyder et al. (2019) and Laderman and Heckert (2021).

Main Survey Outcomes: The main survey outcome is the taxpayer's preferred spending level after the pandemic has passed. This outcome relates to the ideal spending level that taxpayers believe will produce $G^*(\hat{\mathcal{I}})$.

Behavioral Outcomes: For the writing an elected official task, we create an indicator for whether participants wrote to increase spending on or encourage investment in public colleges in the state. For the donation activity, we generate an indicator for whether or not they made a donation.

Additional Survey Outcomes: All Likert scale questions were converted to binary variables equal to one if the participants indicated agree or strongly agree. The question measuring trust that public colleges spend tax dollars well is our main measure of state productivity $\hat{\omega}$.

Demographics: Participants were coded as Republican if they indicated their partisanship as lean Republican, not very strong Republican, or strong Republican. We define the elderly as participants 65 and older. We also create a family attachment indicator for whether the participant or her child has attended a public-four year college in their state. Finally, a participant is said to follow the state's public college sport teams if they watch at least 2 games per year.

3.4. Perceptions of and Support for Public Higher Education

In theory, the impact of providing on taxpayers' perceptions depends on their initial perceptions. If they were perfectly informed, information would not alter perceptions and,

[11] Our two revealed preference measures are also similar to other recent experiments that have included charitable donations (Alesina et al., 2018), and petition signatures or other government notifications (Grigorieff et al., 2018; Haaland and Roth, 2020; Holz et al., 2020).

[12] In the survey, we elicit beliefs about the rank (1-50) but use the percentile rank for our analysis so both a higher rate and a higher rank both correspond to larger numbers.

therefore, should not affect their preferences for public spending. Panel (a) of Figure 3.2 shows that taxpayers in our sample have inaccurate perceptions about higher education. On average, perceptions of graduation rates are accurate but with large variance. Most taxpayers overestimate spending but underestimate their state's rank relative to others. Since our experiment focuses on the graduation rate and its rank, Panel (b) explores the relationship between these two belief errors. It shows that taxpayers who believe the graduation rate is higher than the truth, also tend to believe that the state's rank is higher than it is. The inaccuracy of these perceptions suggests that providing information may help taxpayers express their true preferences over policies.

The policy implications of altering perceptions depend on how they vary across the population. For example, are taxpayers who (do not) value public investment in higher education more likely to learn that the true graduation rate or percentile rank is lower or higher? What different taxpayers learn and how they reason affect both aggregate preferences for spending and polarization. First in Panel (a) of Figure 3.3 we test whether the value of public investment varies by taxpayer identity. Using the control group who do not receive any information, we estimate the following Poisson regression for a taxpayer's preferred spending level since some taxpayers state an preferred spending of \$0.:

$$\mathbb{E}(\text{Preferred Spending Level}_i) = \exp(\alpha + X\beta) \tag{3.4.1}$$

where X is a vector of taxpayer characteristics, as well as a linear probability model:

$$\mathbb{1}(\text{Trust Colleges to Spend Well}_i) = \alpha + X\beta + \varepsilon_i \tag{3.4.2}$$

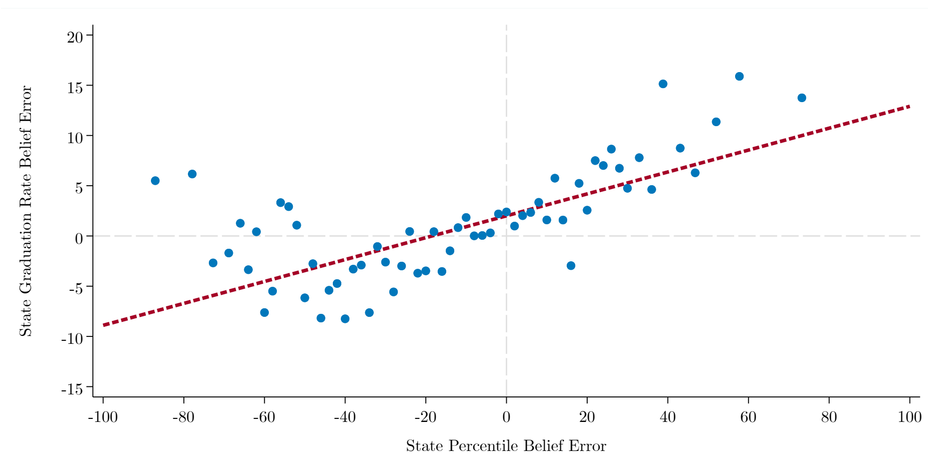
where the dependent variable is an indicator that is 1 if the taxpayer strongly agrees or agrees that she trusts public colleges in her state to spend tax revenue well.

Republicans and the elderly have much lower baseline preferences for spending and are less likely to say that the state spends tax revenue well on public higher education compared to Democrats and Independents, and the non-elderly, respectively. These gaps by partisanship and age are similar to results from two recent nationally representative Pew Research Center surveys on views of higher education (Parker, 2019) and support for free public college (Hartig, 2020). Doyle (2007) also found that Democrats are more concerned about "dropout" than Republicans, which suggests Democrats are willing to pay more in taxes to increase graduation rates. Our treatments allow us to understand to what extent these differences in support are driven by perceptions about the output of public goods and the effect of making perceptions more accurate.

Whether information will shrink or exacerbate these baseline gaps by identity in public

Figure 3.2: Perceptions of Public Higher Education

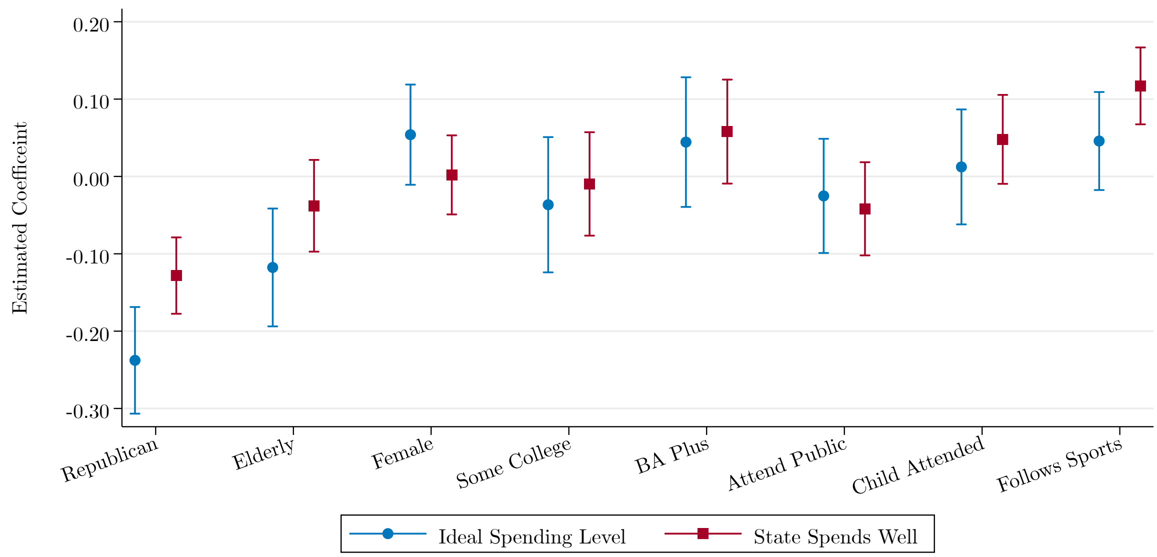
(a) Public Higher Education Belief Errors



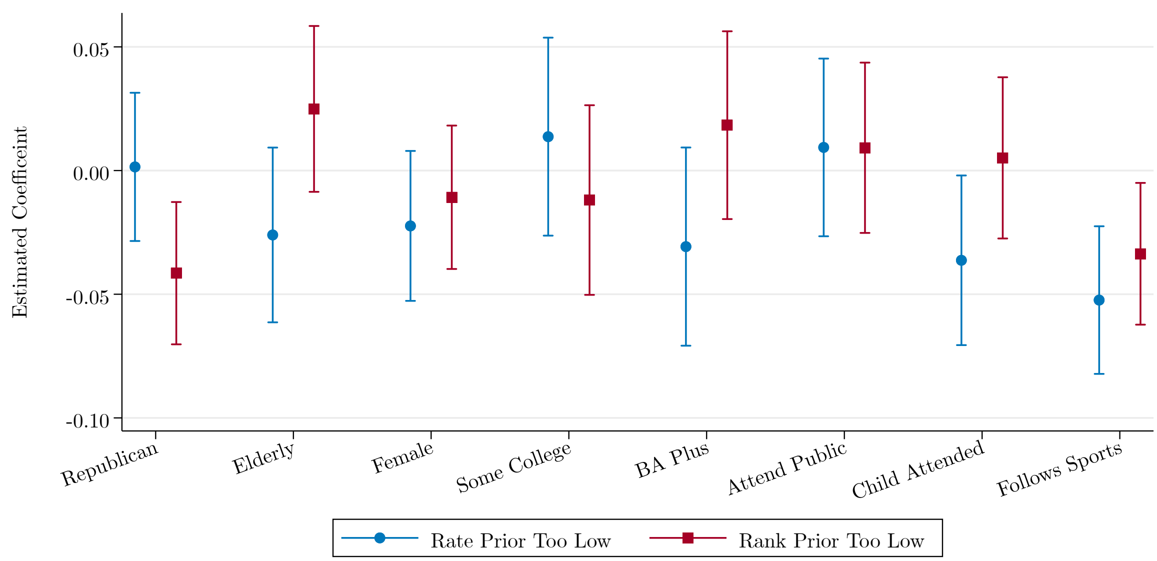
(b) Graduation Rate Belief Errors

Note: Panel (a) plots the distribution of belief errors for all elicited priors. The box plot presents the interquartile range with a line at the mean, and the wiskets show the 10-90 interquantile range. Means are marked with circles, and significant differences from no error on average are marked with asterisks. Errors are defined as the believed rate or percentile rank minus the truth. A value of zero means the participant expressed a perfectly accurate baseline belief. Panel (b) is a binned scatter plot to show the relationship between baseline beliefs about graduation rate from four year public colleges in the taxpayer's state and about the percentile rank of the state graduation rate relative to other states. Each bin plots the average graduation rate error for each percentile of the state rank error. The dashed line is the line of best fit from all responses. Both panels reflect responses of 4719 taxpayers recruited from Qualtrics and CloudResearch.
*** $p < 0.001$

Figure 3.3: Predictors of Perceptions and Support



(a) Preferences with Uncorrected Perceptions



(b) Perceptions about Graduation Rates and Rank

Note: The blue series in Panel (a) are estimate from a Poisson regression of the ideal spending level, and the red series are estimates from a linear regression on an indicator that the taxpayer strongly agrees or agrees that the state spends tax revenue well on public higher education. Panel (a) uses only the control group who do not receive any information. The blue series in Panel (b) plot the estimates from a linear regression of an indicator for prior rate belief - truth < 0 on indicators for several characteristics using the full sample of 4719 taxpayers. The red series plot the estimates from a similar specifications with an indicator for prior percentile rank belief - truth > 0 as the dependent variable. The omitted education group is high school graduate or equivalent, or no high school degree.

support depends on what types of biases different taxpayers have. To assess this, we estimate two linear probability models similar to Equation 3.4.2 where the dependent variables are indicators for whether the prior rate and prior rank are too low. A negative estimate shown in Panel (b) of Figure 3.3 indicates that members of a group are less likely than non-group members to believe the graduation rate or rank is lower than the truth. For example, the -0.04 point estimate on "prior rank too low" for Republicans indicates that Republicans are 4 percentage points less likely to believe that the graduation rank is less than the truth compared with Democrats and Independents; Republicans are relatively overoptimistic about the state's rank.

The results indicate that our treatments provide qualitatively different information to different types of taxpayers. Republicans and young taxpayers in the rate and rank treatment are less likely to learn that rank is high compared to similarly treated Democrats, Independents, and the elderly. We also find that parents of students who attended the in-state public college are more likely to believe the graduation rate is low even though they may have a more personal connection to the system. Finally, those who follow the college's sport teams are optimistic about absolute and relative graduation rate relative to those who do not follow sports, which may reflect state pride. These patterns suggest that within-group correlation in perceptions mean that information could have a meaningful impact on collective preferences and polarization. However, if different types of taxpayers update and respond differently to the same signals about productivity and provision, this need not be the case, leaving the overall effect part of the quantitative question we hope to answer.

3.5. Impact of Information on Preferences

We estimate the average effect of information about the graduation rate only, D_1, and both rate and rank, D_2, on taxpayers' preferred spending level from the following Poisson regression:

$$\mathbb{E}(\text{Preferred Spending Level}_i|D_i] = \exp(\tau_0 + \tau_1 D_{1i} + \tau_2 D_{2i} + \tau_3 X) \qquad (3.5.1)$$

where X is a vector of demographics, as before, plus an indicator that the participant was recruited from Qualtrics.[13]

We present the treatment effects on preferences for education spending in the top panel of Table 3.1. The remaining panels show the treatment effects from a linear probability model on the revealed preference measures and stated beliefs about productivity. In the full sample

[13]Controlling for whether the participant was recruited from Qualtrics or CloudResearch has a negligible effect on the estimates.

receiving information about the graduation rate or both the rate and rank increases taxpayers' preferred level of tax revenue collected per student by about 5 percent on average (first column). After learning the true graduation rate or the rate and rank, taxpayers are about 7.5% more likely to donate some of their wages to public colleges in their state ($p = 0.032$), and 16.5% more likely to write state representatives to request more spending on higher education ($p = 0.106$).[14] Taxpayers who receive rate and rank information also perceive their states as being more productive on average, reporting a belief that it spends tax revenue to fund higher education well.

In the context of the theory this implies that learning the graduation rate and state rank does not just change perceptions about the amount of spending or place in the distribution of states (moves along along $f()$ and $p()$), but it actually changes perceptions of state productivity and the full distribution of graduation rates.

3.5.1 Impact of Information by Prior Perceptions

The average effects from the previous subsection obscure important heterogeneity between taxpayers with different prior perceptions. Taxpayers who incorrectly believe that the graduation rate or rank are low receive a qualitatively different treatment from those who incorrectly believe it is high. We therefore divide participants based on their initial perceptions or priors. The second and third columns divide the sample based on whether an individual initially perceives the graduation rate to be lower or higher than the truth, while the fourth and fifth columns divide the sample based on the rank relative to other states. All participants are therefore in three columns, the full sample, one of the second or third, and one of the fourth or fifth. For each group, we estimate the treatment effects and report the results in the final four columns of each panel of Table 3.1. It is important to note that prior beliefs are not exogenous and so both qualitative treatment and sample both change across the columns. Figure 3.4 additionally illustrates the treatment effects of providing information over the distribution of belief rate and rank errors.

The largest effects of the graduation rate information are on taxpayers who initially believe that the rate is lower than the truth (second column). For these individuals, learning the graduation rate is higher increases preferred spending, and leads them to be more likely to donate to a public college in their state. We find these taxpayers who learn about the rate are about 7.6 percentage points more likely to say the state spends its tax revenue well than the control group.[15] Since these types of taxpayers express low levels of support without

[14] p-values from regressions that pool D_1 and D_2 into one variable. For separate tests see Table 3.1.

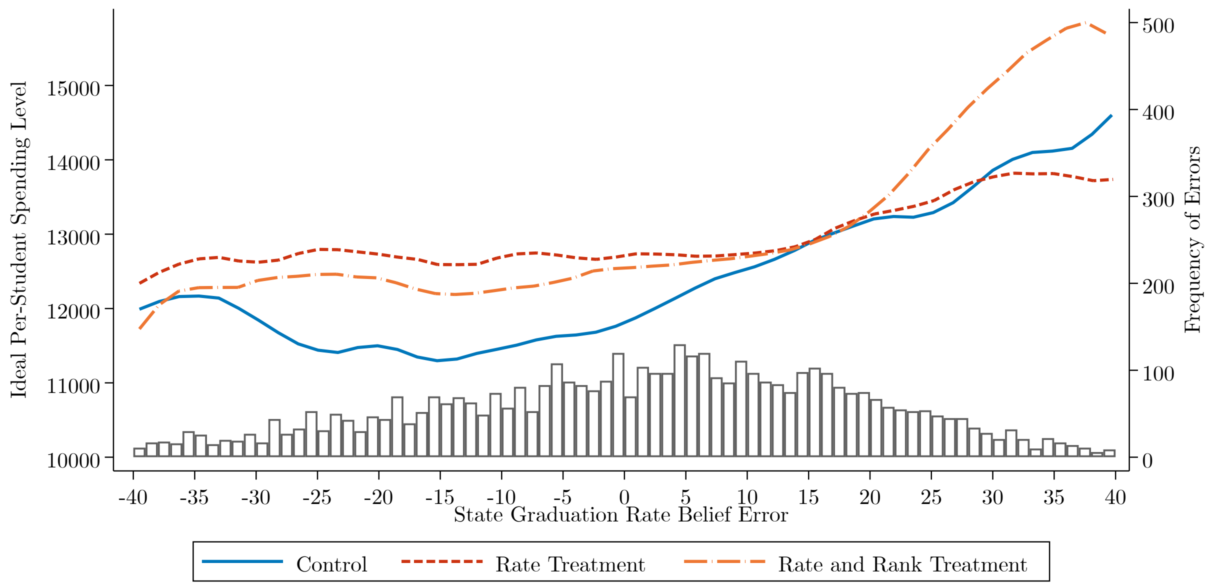
[15] Our results here are similar to Peyton (2020), who finds that experimentally induced increases in government mistrust reduce support for redistribution.

Table 3.1: The Effects of Information Vary by Prior Perceptions

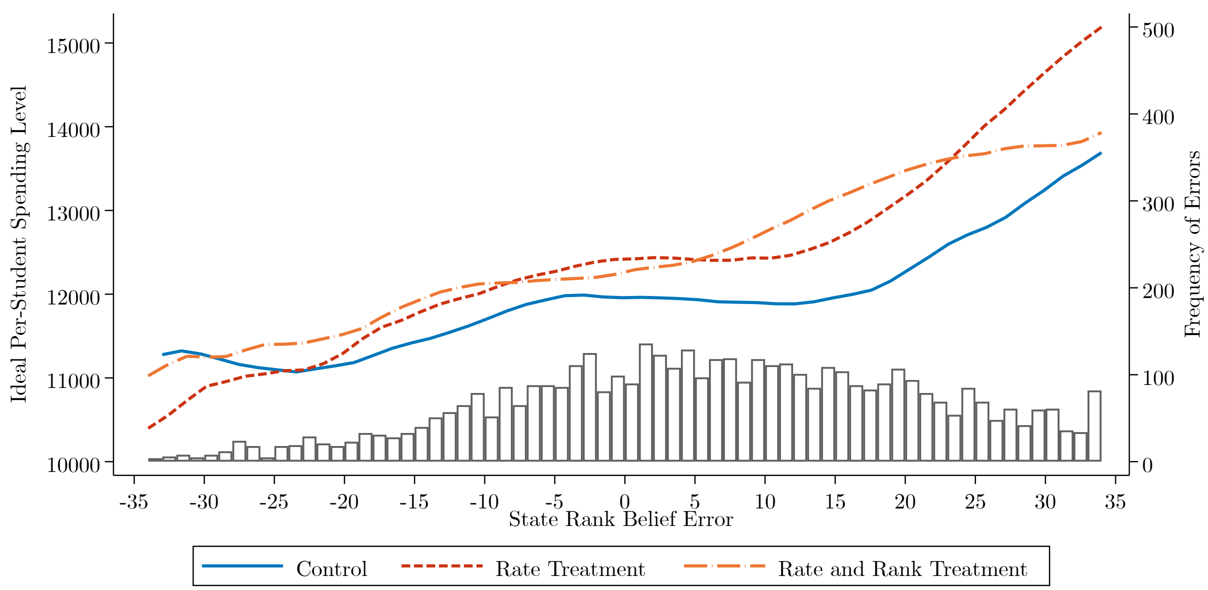
		Graduation-Rate Prior:		State Percentile Prior:	
	Full	Prior	Prior	Prior	Prior
Ideal State Spending Per-Student (% Change)	Sample	Too Low	Too High	Too Low	Too High
Rate Treatment	0.048*	0.087**	0.012	0.066*	0.013
	(0.022)	(0.033)	(0.029)	(0.026)	(0.038)
Rate and Rank Treatment	0.049*	0.064*	0.035	0.063*	0.019
	(0.021)	(0.032)	(0.028)	(0.026)	(0.036)
Control Mean	12,210	11,491	12,821	12,492	11,694
Donated to Public College					
Rate Treatment	0.032$^+$	0.051$^+$	0.013	0.037$^+$	0.019
	(0.017)	(0.026)	(0.024)	(0.022)	(0.030)
Rate and Rank Treatment	0.034$^+$	0.007	0.055*	0.029	0.041
	(0.017)	(0.026)	(0.024)	(0.022)	(0.030)
Control Mean	0.426	0.422	0.430	0.429	0.422
Wrote Representative for More Spending					
Rate Treatment	0.016	0.020	0.013	0.026*	-0.001
	(0.010)	(0.015)	(0.014)	(0.013)	(0.017)
Rate and Rank Treatment	0.012	0.023	0.002	0.023$^+$	-0.008
	(0.010)	(0.015)	(0.014)	(0.012)	(0.017)
Control Mean	0.085	0.082	0.087	0.082	0.091
Trusts State Colleges to Spend Money Well					
Rate Treatment	0.027	0.076**	-0.022	0.034	0.015
	(0.017)	(0.025)	(0.023)	(0.021)	(0.030)
Rate and Rank Treatment	0.037*	0.035	0.035	0.052*	0.010
	(0.017)	(0.024)	(0.024)	(0.021)	(0.030)
Control Mean	0.384	0.304	0.452	0.368	0.413
Average Rate Error	0.5	-16.4	14.5	-1.9	5.2
Average Percentile Error	-13.6	-18.8	-9.3	-31.7	20.7
Observations	4719	2134	2585	3086	1633

Note: This table reports the results from our main regression specifications. There are four sets of rows for the outcomes and five columns for different (sub)samples. Results are from OLS regressions with the exception of ideal spending which are from Poisson regressions. All regressions control for Republican identity, gender, age over 65, education, prior or family attendance in the college, state college sports following and whether the taxpayer was recruited by Qualtrics or CloudResearch. Standard errors are robust to non-identically distributed errors across observations.
*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, $^+p < 0.1$

Figure 3.4: The Impact of Information by Belief Errors



(a) The Impact of Information by Belief Rate Error



(b) The Impact of Information by Rank Error

Note: This figure displays the average preferred per-student spending by individuals with baseline rate (Panel (a)) and rank (Panel (b)) errors. In each figure preferences are plotted separately for individuals in the control, rate, and rank + rank assignments. Reported expectations are the smoothed values of a kernel-weighted mean estimated with an Epanechnikov kernel and a bandwidth of 6 (approximately 1/3 of standard deviation).

corrected perceptions, based on the control mean, the rate information treatment also helps reduce polarization by prior beliefs. This suggests that variance in perceptions exacerbate polarization in our setting. For taxpayers whose initial perceptions of the graduation rate are too low, additionally correcting perceptions of state rank may weakly decrease support compared to only providing the rate information, but may increase support through political participation.[16]

The treatment effects for taxpayers who learn that the state is more highly ranked are similar (fourth column). This is not unexpected as more than 73% of these taxpayers also initially perceive the graduation rate is lower than in reality, so there is substantial overlap in these samples. The only substantive difference in results is that whereas adding rank information in the second treatment reduces perceptions of state productivity for people with negative priors about the graduation rate (column 2), it increases perceptions of state productivity among those with negative priors about the state rank (column 4).[17] When we further divide these taxpayers by their rate prior, we find suggestive evidence that the rank information works in opposite directions based on the sign of the rate error.

Although information increases support for spending among taxpayers with pessimistic perceptions, we find no evidence that information leads overly optimistic taxpayers to reduce their preferences for spending (third and fifth columns). If anything the point estimates of the effect of information on ideal spending are positive. This asymmetry is striking[18] and suggests that who tend to have different prior perceptions also tend to reason differently about information. We find evidence of this in the effects. For example, there are no significant effects on trust in state productivity, suggesting that information leads taxpayers with pessimistic priors to update along the marginal cost function rather than shifting marginal costs. This combined with some learning about the distribution of graduation rates rationalizes the asymmetry.

3.5.2 Impact of Information by Taxpayer Characteristics

As shown in the previous subsection, the effect of providing information depends on the type of information taxpayers receive. In addition, the effect will depend on how taxpayers update their perceptions after learning the truth. This section explores the extent to which information can have different effects on different types of taxpayers based on their identities. If there are differences, providing information could exacerbate or reduce

[16] We cannot rule out that treatment effects are equivalent at the $p = 0.05$ level.

[17] The test that $\tau_2 - \tau_1$ is equal in among the two groups has $p = 0.005$

[18] We reject symmetry ($\tau_{k,low} = -\tau_{k,high}$ for $k = 1, 2$) at the 0.021 and 0.019 levels for the rate priors and 0.010 and 0.073 levels for the rank priors.

polarization. To explore these differences we now investigate the effect of our treatments by political partisanship, age, and family attachment to public higher education.

Table 3.2 shows that there are large differences in effects between groups. For example, information has a large positive effect on support for both Republicans and the elderly (first and third columns of Table 3.2).[19] This is true for the graduation rate treatment effect among both Republicans and the elderly. Among Republicans the additional rank information seems to reduce the effect of graduation rate information by almost half ($p = 0.077$), but among the other groups both effects seem quite similar

The differences between effects suggest that Republicans reason differently about information compared to Democrats and Independents. For Republicans, information increases the preferred level of spending, suggesting either an increase in the perceived marginal benefit or a decrease in the perceived marginal cost. Changes in perceptions of productivity are small and imprecise, suggesting no major changes in perceived marginal costs. In this case, we attribute remaining changes in preferred spending to learning about the full distribution of graduation rates across states, which changes the perceived marginal benefit. For Democrats and Independents on the other hand, there is no change in the preferred level of spending or changes in perceptions of productivity. This suggests that there should also be no changes in the perceptions about the distribution of graduation rates across states.

To explore whether changes in spending preferences arise from changes in perceptions of the distribution of graduation ranks, we compare the posterior perceptions of rank for 665 respondents who received the graduation rate treatment (Appendix C.3 details the data collection and descriptives). We find evidence that Republicans update their perceptions of the distribution of graduation rates (information produces changes across $F()$) whereas Democrats and Independents do not (information produces changes along $F()$). Figure 3.5 plots average priors and posteriors over the true graduation rank for each group. Each panel presents two cross-sectional relationships: the prior perception of state rank over and the posterior perception of state rank each over true graduation rates. It also reports a test of the null hypotheses that those perceptions do not change. We find that whereas the average perceptions of Republicans change, those of Democrats and Independents do not.

The elderly and younger taxpayers also seem to reason in different ways. The effects of information on preferences for spending among the elderly also suggest changes in the net marginal benefits of more graduates. The large effects on perceived productivity suggest that decreases in the perceived marginal cost of higher graduation rates is an important factor.

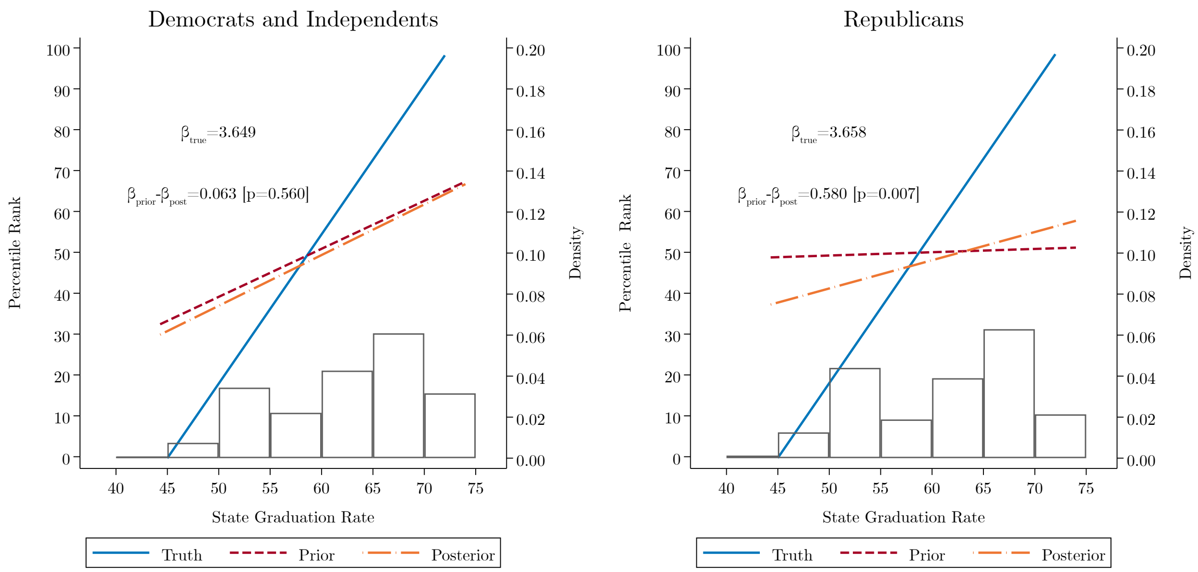
[19] The differences are significant at the $p = 0.004$ level for the graduation rate treatment among Republicans compared to Democrats and Independents, and the differences are significant at the $p = 0.034$ and $p = 0.005$ level for the two effects among elderly compared to younger taxpayers.

Table 3.2: The Impact of Information by Taxpayer Characteristics

	Republican	Independent or Democrat	Age 65 and Over	Under Age 65	Family Attachment to State College	No Family Attachment
Ideal State Spending Per-Student						
Rate Treatment	0.138***	0.003	0.090**	-0.004	0.073*	0.026
	(0.040)	(0.026)	(0.029)	(0.032)	(0.032)	(0.030)
Rate and Rank Treatment	0.071^{+}	0.036	0.101***	-0.017	0.055^{+}	0.043
	(0.039)	(0.025)	(0.029)	(0.031)	(0.031)	(0.029)
Control Mean	10,417	13,293	12,520	12,759	12,077	12,323
Donated to Public College						
Rate Treatment	0.048^{+}	0.023	0.058*	0.001	0.058*	0.008
	(0.029)	(0.022)	(0.023)	(0.026)	(0.026)	(0.024)
Rate and Rank Treatment	0.016	0.045*	0.034	0.035	0.047^{+}	0.022
	(0.028)	(0.022)	(0.023)	(0.027)	(0.026)	(0.024)
Control Mean	0.381	0.454	0.424	0.476	0.444	0.412
Wrote Representative for More Spending						
Rate Treatment	0.012	0.020	0.029*	-0.000	0.014	0.019
	(0.012)	(0.015)	(0.013)	(0.017)	(0.017)	(0.013)
Rate and Rank Treatment	-0.002	0.021	0.024^{+}	-0.004	0.028^{+}	-0.002
	(0.010)	(0.015)	(0.013)	(0.016)	(0.016)	(0.013)
Control Mean	0.040	0.127	0.086	0.106	0.111	0.081
Trusts State Colleges to Spend Money Well						
Rate Treatment	0.035	0.023	0.045*	0.002	0.045^{+}	0.013
	(0.027)	(0.022)	(0.022)	(0.027)	(0.025)	(0.023)
Rate and Rank Treatment	0.025	0.045*	0.043^{+}	0.029	0.056*	0.021
	(0.027)	(0.022)	(0.022)	(0.027)	(0.026)	(0.023)
Control Mean	0.297	0.436	0.382	0.434	0.387	0.381
Average Rate Error	0.6	0.5	0.8	0.2	1.1	0.0
Average Percentile Error	-11.4	-14.5	-15.3	-11.2	-12.7	-14.3
Observations	1,740	2,979	2,712	2,007	2,167	2,552

Note: This table reports the results from our main regression specifications. There are four sets of rows for the outcomes and five columns for different (sub)samples. Results are from OLS regressions with the exception of ideal spending which are from Poisson regressions. Except when omitted to condition, all regressions control for Republican identity, gender, age over 65, education, prior or family attendance in the college, state college sports following and which sampling method generated the observation. Standard errors are robust to non-identically distributed errors across observations.
*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, $^{+}p < 0.1$

Figure 3.5: Differences in Updating by Political Identity



Note: This figure displays the average relationship between perceptions of state rank and the true graduation rate. Linear approximations of the relationship with priors, posteriors, and the truth are plotted separately by political identify. p-values are from a test that the slopes are equal in a linear regressions on the 665 individuals with the main regression controls and robust standard errors.

Like Democrats and Independents, younger taxpayers do not show increased preferences for spending or changes in perceptions of costs. We cannot detect differences in updating about the full distribution of graduation rates across states.[20]

The final two columns of Figure 3.2 estimate the treatment effects separately for taxpayers with and without a family attachment. Recall that we define a taxpayer's family attachment as whether she or her child attended a public college in her state. Since higher education provides both a public good and a private benefit to those who attend, information may have a different effect for those with a personal attachment, and private benefits, than those without one. We find that our effects of information are larger for those with a personal attachment. This may be because the perceived marginal return on investment is higher for them or because the information is more salient.[21] Those without a family attachment may still respond somewhat positively, suggesting that information does not create backlash among taxpayers who do not have private benefits from public expenditure.

These results are not consistent with a simple model where effects are driven by taxpayers

[20] Since the posteriors were collected from the younger Cloud Research Sample, we are underpowered in testing whether there are differences in rank updating between the elderly and non-elderly.

[21] Similar to these mechanisms, Chatterji et al. (2018) find that legislators who attended their states' public colleges spend more on higher education.

with ill-informed perceptions. If there were no differences in how taxpayers from different groups reason, group heterogeneity could reflect differences in the strength of their priors. But the results suggest this is not the case. The groups with the most exposure to the state higher education system should have the strongest priors, and groups with strong priors would have smaller effects. If this were the case, the large effects on Republicans, who are less likely to be college educated, and the elderly, who were educated longer ago, seem quite intuitive, but the effects on individuals with a family connection to the public higher education system are the opposite of this prediction.

3.6. Discussion

3.6.1 Implications for Polarization

The heterogeneity we document shows that information increases demand for spending more among individuals with lower baseline preferred spending. This pattern of heterogeneity is reflected in results by prior perceptions and by observable characteristics. This subsection explores the role of information in reducing polarization, a relevant goal for understanding political economy and for implementing optimal public policy.

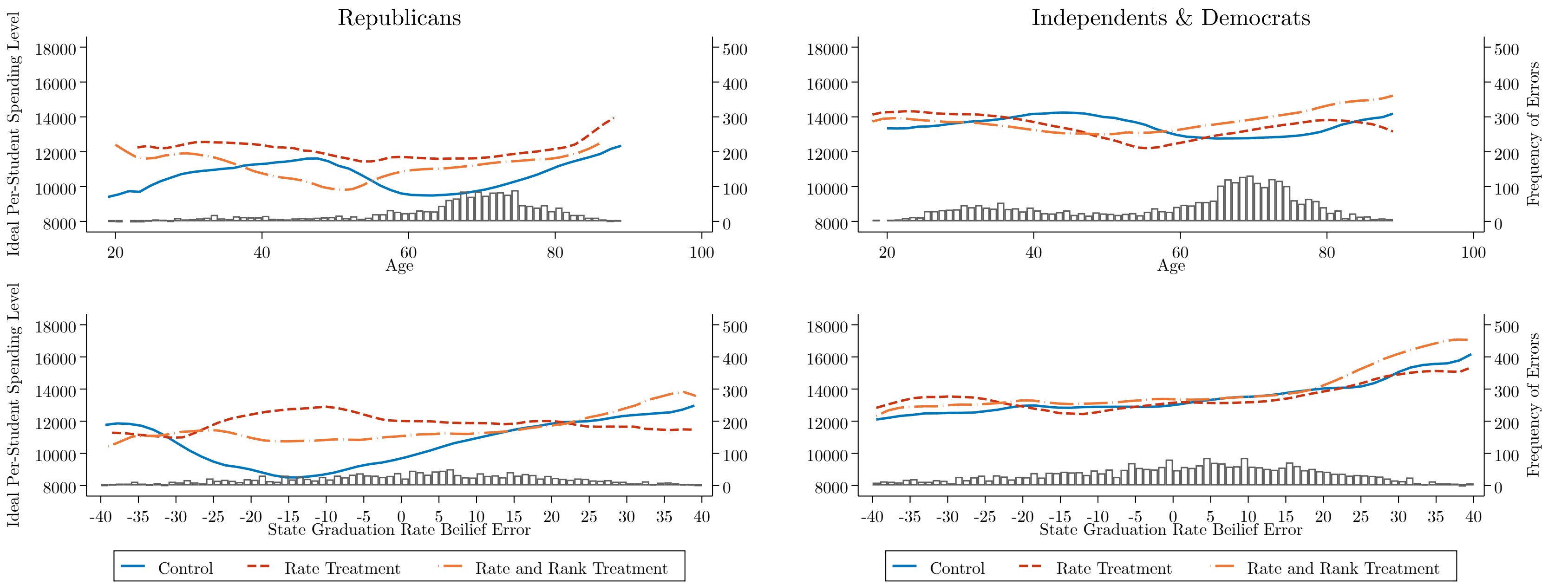
The large effects of information on Republicans and the elderly in Table 3.2 are striking because these groups have the lowest baseline preferences for spending on higher education. Providing information to them simultaneously increases average support and greatly reduces polarization. The political partisan gap in preferred spending falls by 32 percent, and the gap between the elderly and non-elderly is eliminated. Part, but not all, of the effect comes from a 2.4 ($p = 0.02$) percentage point decrease for Republicans and a 1.3 ($p = 0.06$) percentage point decrease for the elderly in the likelihood they state that they prefer $0 in public higher education expenditure after receiving one of the information treatments.[22]

To better understand the relationship between political and age polarization, the top two panels of Figure 3.6 plot the average preferences for spending over the age distribution separately by political affiliation and treatment. Information has the largest effect on elderly Republicans, who prefer especially low levels of spending when their perceptions are left uncorrected.[23] We find much less of an age divide in preferences for spending for Independents and Democrats both with and without corrected perceptions.

[22]These regressions pool both treatment effects but are otherwise similar to other linear regression specifications. Note that quantile regressions suggest that rather than reducing polarization on both ends of the distribution, information increases preferences at all places of the distribution (except for Republicans at high percentiles where learning both rate and rank may actually reduce preferred spending).

[23]Our results also suggest a very large effect on the very young, but there are very few participants in our sample younger than 30.

Figure 3.6: Impact of Information on Polarization



Note: This figure displays the average preferred per-student spending by individuals with different characteristics, priors, and treatment assignments. In each figure preferences are plotted separately for individuals in the control, rate, and rate and rank assignments. The top two panels show the average preference for spending over the distribution of age, with Republicans on the right and with Democrats and Independents on the left. The bottom two panels show the average preferences for spending over the distribution of belief errors about the graduation rate, again with Republicans on the right and with Democrats and Independents on the left. Reported expectations are the smoothed values of a kernel-weighted mean estimated with an Epanechnikov kernel and a bandwidth of 6 (approximately 1/3 of a standard deviation).

The bottom two panels of Figure 3.6 further explore how perceptions and information affect political polarization by plotting the average preferences for spending over the belief error distribution separately by political affiliation and treatment. Preferences for spending are largely increasing in graduation rate belief error for Republicans who did not receive information.[24] Those who believe the graduation rate is high, i.e., those with a positive belief error, are more likely to support spending, perhaps because they associate a high graduation rate with a high return on investment, less waste, or greater efficiency,[25] or because they already think the state is spending higher amounts of money. Information about the rate and rank dramatically flatten the slope and raise average preferences for spending. This implies that inaccurate perceptions played a large role in shaping taxpayers' stated preferences. Only Republicans who perceived the graduation rate as extremely high respond negatively, on average, to information. Those who initially believe it is high are significantly more likely to have a family attachment to the public college system, so the negative information may signal a decrease in the marginal return as well as lower productivity.

3.6.2 Interpreting Revealed Response Outcomes

Although the measures of stated and revealed preferences reveal similar patterns, the effects on the stated measure of preferences do not always exactly align with the revealed measures of preferences and the mechanisms. This section connects the four main outcomes with the model to see whether the differences are revealing interesting variation in preferences or updating.

Trust to Spend Well. In the model, only information about the graduation rate should affect perceptions of absolute state productivity. In practice, we observe the graduation rank information also affecting whether taxpayers express trust that the state's higher education system will spend tax dollars well. We interpret this as evidence that information can also affect what taxpayers are willing to count as "spending well." In this case information about the distribution of graduation rates also communicates information about the distributions of state productivity and state higher education spending.

Political Engagement. The current measure of political engagement is the most closely connected to our theory. Essentially it captures whether a taxpayer wants spending to increase. But whereas someone's preferred spending level independent of her perceptions of the current level of expenditures, whether she tells representatives to increase support

[24]We find a decrease in support moving from -40 to -15 of belief rate errors, but there are relatively few participants with perceptions in that range.

[25]Doyle (2010) finds that conservative US senators are more likely to vote in favor of issues regarding the efficiency of colleges and universities, suggesting these concerns may be especially important for determining Republican taxpayer preferences.

will not be. To the extent to which individuals are moving along the perceived cost $p()$, the distance between their preferred and perceived spending may be shrinking, reducing the need to engage in political participation.

Donation Activity. Whereas the stated preference measure depends on where marginal cost and marginal benefit of graduates are equal, the donation activity also captures information about the marginal cost and marginal benefit in the status quo. Taxpayers will only donate when their perceived marginal benefit of public spending is greater than the marginal cost, which in the donation activity is their marginal utility of consumption. For the majority of the paper, we argued that shifts along the perceived costs, $\hat{p}(G)$ and perceived distribution of graduation rates $\hat{F}(G)$ would not affect preferred spending, but that will not be the case for the donation activity. For example, learning that the graduation rate is lower will increase the marginal utility of having a higher graduation rate if utility is increasing and concave in G. This is consistent with the fact that the effect on donations is the largest for taxpayers who get bad news about both the rate and the rank.

3.7. Conclusion

Although public investment in higher education improves student outcomes like graduation rates (Bound and Turner, 2007; Bound et al., 2010; Deming and Walters, 2017) and higher education provides a broad public good (Moretti, 2004), states have not prioritized its funding. The level of public higher education services may be lower than the efficient level, in part, because of taxpayers' perceptions about the marginal costs and benefits of the public good. We find that taxpayers with overly pessimistic perceptions of state graduation rate are more likely to prefer less spending and are more likely to express their misinformed preferences to elected officials. Providing information can therefore help correct the inefficiently low level of provision.

We find that providing taxpayers with information about the graduation rate increases averages preferences for public expenditure. Taxpayers who learn about their state's graduation rate increase their preferred level of tax revenue per student by about 5 percent. This effect seems to be driven by those learn that the graduation rate is higher than previously perceived. This change is consistent with learning that the return on investment is higher than expected, and these taxpayers are more likely to say that the state spends its tax revenue well on public higher education. Additional information about relative provision from the state's graduation rank has little additional effect on average preferences for spending. However, the effects of the two information provision treatments vary importantly by initial perceptions. Our treatments also have especially large effects on Republicans and the elderly, two groups who historically oppose high spending on higher education. Information therefore greatly reduces political polarization in preferences for public expenditure.

Although different groups tend to receive qualitatively different information, we show evidence that different groups also tend to reason differently about the information they receive. For example, Democrats and Independents are much more likely to adapt their perceptions of state productivity and much less likely to change their perceptions about the distribution of graduation rates across states. On the other hand Republicans tend to do the opposite; they adapt their perceptions about the distribution of graduation rates across states much more than their perceptions of state productivity.

By focusing on information about the outputs of the public good provision process, our research also opens new questions about the role of information about inputs like spending. In fact, differential learning about perceptions about the distribution of productivity and spending across states could be driving some of the less intuitive patterns we see in our revealed preference measures of state spending. Future research should combine our insights about the importance of output information with the literature about providing information about costs to delve into this more.

While our results highlight the value of information about public higher education, the initial distribution of taxpayer beliefs and how taxpayers reason about new information will determine how providing information shapes preferences. Altering perceptions in other settings may not have the same benefits. Another promising avenue of research and policy practice could be investigating how taxpayers respond to information about the level of provision of other public goods such as health care, public safety, infrastructure, defense, and welfare. Although there is vast heterogeneity in preferences for many of these programs, our results suggest that information plays a crucial role in the formation and expression of welfare-maximizing preferences for the provision of public goods.

APPENDICES

APPENDIX A

Appendix to Chapter 1

A.1. Data Appendix

Data for this project largely come from the Michigan Education Data Center (MEDC) which houses administrative education data collected by the Center for Educational Performance Information (CEPI) and the Michigan Department of Education (MDE). The two main datasets for my analysis are a student-year level data set of K-12 enrollment and a student-year level data set of assessments, both spanning records from the 2001-2002 school year until the 2018-2019 school year.

The K-12 data contain longitudinal records for each student enrolled in Michigan public schools between the 2001-02 and 2018-19 school years. These rich data contain reported student characteristics including sex (Male/Female) and race (Black non-Hispanic/White non-Hispanic/Hispanic/Native American, Alaskan, Hawaiian, or Pacific Islander/Asian/Other), administrative poverty status (any of the following in school year: free or reduced price lunch, SNAP or TANF recipient, homeless, migrant, or in foster care), and Census block group. They also include scholastically relevant scholastic variables including grade, school and district attended, assigned district, attendance rate, and detailed special education service receipt. Unfortunately, the grade variable has difficulty separating out developmental and traditional kindergarten, but after 2014 it does indicate whether students were in a separate developmental kindergarten classroom.

I sample the 1,874,778 students from the enrollment data who entered kindergarten in Michigan public schools between fall 2002 and fall 2018. I assume students who turn five between March through December of a given year to make their decisions based on that year's cutoff and that students who turn five in January or February to make decisions relative to the recently passed cutoff. For example students who turned five on March 1, November 1, and January 1 effectively choose between entering in 2013 and waiting to enter in 2014. This assumption implies that all students face one relevant cutoff. For example, the 2013 cohort used as the main sample is comprised of children who turned five between March 1, 2013 and February 28, 2014. A possible violation of these assumptions could be if academic redshirting is so prolific that students who turn five in the winter act as if they face the cutoff in the coming fall rather than the relevant cutoff behind them. Empirically this does not seem to be an issue. The redshirting rate reaches almost zero by the spring, and continues to decrease moving towards the winter. In practice I focus on students born in closer to the cutoff dates, which means these assumptions have very little bite.

I then restrict my sample. I keep students facing the December 1, 2002 cutoff through the September 1, 2015 cutoff because these are the cutoffs for which I observe all the relevant students and scores. I also drop 20,933 (1.1%) students who enter in years other than the two that they should be choosing between based on the assigned cutoff. While these students are recorded as starting in a year that would not be allowed based on their birthdate, it is unclear how much of this is due to true (and in most cases illegal) choices as opposed to moving into the state or measurement error in birthdates. This restriction leaves me with 1,549,314 students who enter kindergarten between 2002 and 2016.

The assessment data contain raw, scaled, and standardized scores for yearly assessments. Although students take assessments almost every year beginning in third grade, having third grade test scores is important because they represent the nearest-term outcomes to kindergarten entry. I use the scaled score for my analyses because it is psychometrically calibrated to be compatible within grade across years. After merging the assessment data onto to my sample, I standardize the scaled math scores among the students facing each cutoff.

A.2. Implications of Strategic Selection for Efficiency and Equity

Using birthday cutoffs to identify selection around birthday recommendations (both selection in levels and selection on gains) will allow me to answer the positive questions about how parents engage in strategic selection, but does not say anything about the normative implications of selection for efficiency and equity. This section presents my concepts of efficiency and equity. Let an allocation, $\mathcal{W} = \{W_1, ...W_N\}$, be the set of waiting decisions made for each of the N children. Because my interest is comparing policies and allocations, I define relative measures of efficiency and equity.

I define one allocation as being more efficient than another if two conditions are met:

the allocation implements choices that are revealed preferred to families, *and* the allocation results in higher average test scores. This concept implies that social welfare is more than just scores. The difference in indirect utility between starting and waiting matters because there are real costs associated with waiting. Whereas the test score optimizing allocation would be to make all children wait to start kindergarten, efficiency takes into account the heterogeneity in costs and in family preferences for gains. At the same time, this concept of efficiency also recognizes that families may or may not fully internalize the social benefits of waiting to their children. These externalities are the motivation for including the average scores in the comparison. An intuitive way to think of this criterion is that a more efficient outcome is one that would be preferred by a social planner with any relative weights on students and parents. A more efficient allocation may not be a Pareto improvement because some children may not prefer it, even though it is revealed preferred. Also note that because one allocation is more efficient than another if parents prefer it *and* it results in higher scores, there may be some ambiguous comparisons (for example an allocation that is revealed preferred but with low average scores versus one that is not preferred but has higher scores), so allocations are only partially ordered.

This conceptualization of efficiency has definite strengths, but it does impose some restrictions on what is allowed to contribute to families' indirect utility. First, this efficiency criterion requires there to be no spillovers in utility among families. A violation of this could be if families would rather not have their children wait because it is costly but they do so because they expect other families to strategically select and do not want their children to be disadvantaged. In this setting replacing a recommendation with a requirement would increase the well-being of all families, but the revealed preference argument would miss that. While occasionally surfacing in popular media, this arms-race model of relative-age is not validated empirically. For example, Appendix Figure A.6 shows the share of always takers at a given date is the same regardless of the cutoff date (Cook and Kang, 2018, produce a similar result).

A second restriction imposed by this model of welfare is its focus the social benefits on average test scores. This focus has two main implications. First, it assumes that test scores capture the social benefits not present in family utility. This assumption rules out examples where a child grows up and regrets the forgone year of earnings despite being higher achieving in school or where she experiences life-long returns from noncognitive gains from waiting beyond what was captured in test scores.[1] A second, more subtle implication is that this

[1]To the extend to which the relationship between test score gains noncognitive gains mostly over u and x, estimating effect could allow a social planner to appropriately reweight average testing gains into a fully welfare-relevant statistic.

assumption limits the extent to which waiting generates spillovers in test scores. It does not rule out the possibility of spillovers (in fact classrooms do benefit from peers who wait, see Peña, 2017); however, it limits the roll of spillovers to operate only through the channel of average achievement not other outcomes or higher-order moments. In most cases this is reasonable, although there is evidence of small increases in crime among students who wait which are not captured by test score gains (Cook and Kang, 2016).

Because social planners and policy makers may care about the outcomes of different groups, a utilitarian concept of efficiency is likely insufficient for discussing welfare. To address this concern I also introduce a simple criterion for equity. Equity for students of type $X = x$ is measured in average differences in realized test scores $\mathbb{E}[Y|X = x] - \mathbb{E}[Y|X \neq x]$, and allocations can be compared by measuring the resulting change in achievement gaps. As such, an allocation is more equitable for students of type x if this gap is smaller. Changes in the size of the gaps arise because different allocations change who waits and because different types of children x have different baseline likelihoods of waiting and different expected gains from doing so. Note that whereas the definition of efficiency included both indirect utility and student achievement, the concept of equity is focused only on achievement. This is consistent with how inequities are defined in income, tax incidence, labor market outcomes, and other settings.

These ideas of efficiency and equity feature in policy conversations surrounding kindergarten entry recommendations and requirements—although rarely using those words. For example, the argument that there should be requirements that force children who turn five after the cutoff has both an efficiency and and equity argument. Strategic selection on this margin lowers average scores, and if the children who would benefit the most from “the gift of time” are less likely to wait, these losses could be large.[2] Furthermore, if these families come from systematically disadvantaged groups, forgoing gains by choosing to start when recommended to wait would widen inequities. These rationales may be behind the policies in 15 US states and many school districts that have strict requirements to wait for children with birthdays after the cutoff.

On the other hand, conversations about academic redshirting often feature both equity and efficiency arguments. When discussing requirements that would force children who turn five before the cutoff to start and not wait (denying treatment to would-be always takers), it is often pointed out that this type of selection is driven by highly educated and often high income families. The assertion is that allowing these families to select around the recommendation to start exacerbates persistent gaps is an equity-based justification for an

[2]Although for the requirement to be socially efficient the social cost of those foregone gains must be larger than the cost imposed on parents by forcing their children to wait.

entry requirement rather than a recommendation. On the other hand, another common assertion is that because these children tend to come from privileged backgrounds, they would do well no matter what. This is an argument about selection that implies that there are not test-score gains from allowing the strategic selection around the recommendation which is related to the ideas of efficiency and average scores.[3]

With equity and efficiency defined, it becomes clear that the ability to extrapolate is critical for measuring both. Although most of the strategic selection happens near the cutoffs, without extrapolation I only assess equity and efficiency at a given cutoff. I would not be able to determine whether scores or gaps would change on average—only among the subsample with birthdays at the cutoff. It is using the marginal treatment effects to extrapolate away from the cutoff that allows me to compare the effects of policies on the whole population. It is the extrapolation that will allow me to assess the merits of the arguments surrounding recommendations, requirements, and strategic selection around them.

A.3. Comparing My Framework to the "Early," "Late," and "On-Time"

Answering the dual questions of how individuals select into treatment in the context of kindergarten entry and identifying heterogeneous treatment effects on student achievement requires a clear definition of treatment. This section proposes a new definition of treatment, "waiting to enter kindergarten," and explores its relationships to prior work answering both descriptive and causal questions.

The key insight behind my conceptualization of treatment is that each student faces the choice between two entry years. A student with a given birth date can either enter this year (usually at or just below five years old) or wait to enter kindergarten next year (usually at or just below six). I define the treatment decision $D_i \in \{0,1\}$ in these terms. Treatment is waiting to enter kindergarten, $D_i = 1$. A student's potential outcomes $Y_i(1)$ and $Y_i(0)$ are the third-grade math scores (in standard deviations), if she waits to enter or does not wait, and her realized scores are $Y_i = D_i \cdot Y_i(1) + (1 - D_i)Y_i(0) \equiv Y_i(0) + D_i \cdot \tau_i$ where τ_i is the effect waiting would have on her.

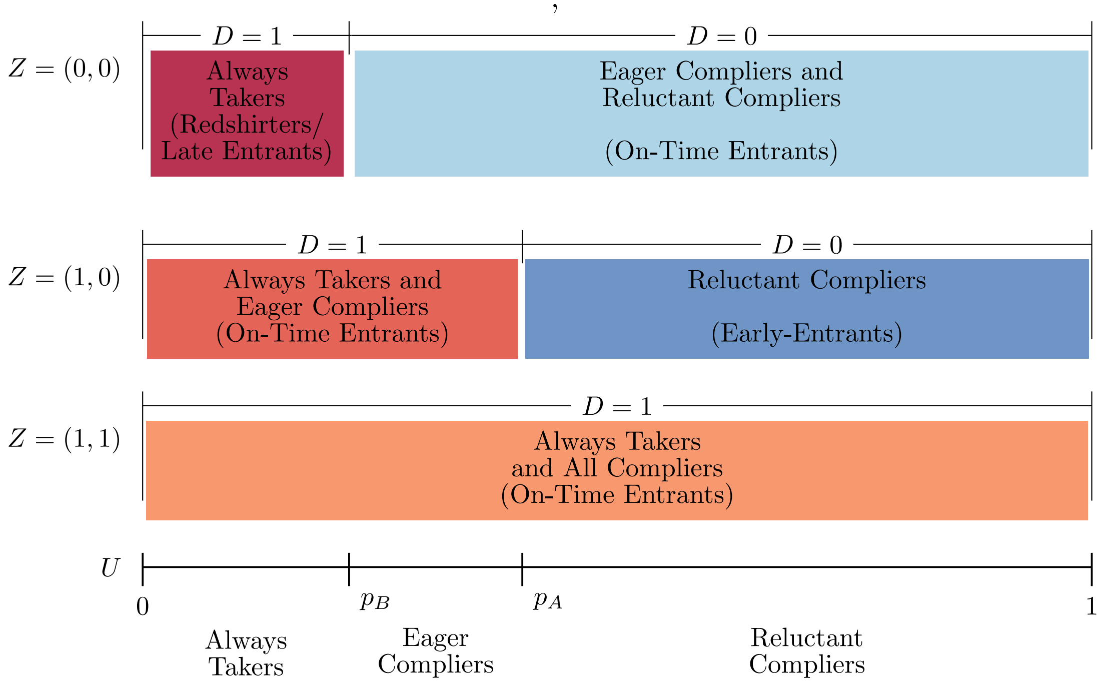
This definition of treatment and the resulting potential outcomes work better than using the traditional entry groups (like those used in Bassok and Reardon, 2013; Fortner and Jenkins, 2017, for example). Using "early" and "late" entry as treatments or potential outcomes relative to "on-time" entry is problematic because the groups are dependent on assignment. Consider a student who turned five on November 1 and is choosing between entering in 2013,

[3]A careful reader will notice that there is a tension between these two narratives. If always takers would do well no matter what (i.e., the treatment effect of waiting is small) then the decision to wait or not cannot be exacerbating persistent gaps between racial and other demographic groups.

as assigned, or waiting to enter in 2014 (redshirting). There is no material difference between her choice and the choice of a student who turned five on November 2 and who was assigned to wait. If they wait, both students would enter in the same year at the same absolute and relative age. This comparison makes it clear that redshirting (waiting when assigned to start) is not a distinct treatment from complying with an assignment to wait; rather the difference is in the selection into waiting.[4] Figure A.1 illustrates the inconvenient implications of the early, on-time, late framework. None of the groups are mutually exclusive. In fact on-time entrants are a combination of always takers, eager compliers who wait, eager compliers who enter, reluctant compliers who wait, and reluctant compliers who enter. Furthermore, not all always takers are late entrants nor are all reluctant compliers early entrants because of randomization.

Figure A.1: Early, Late, and On-Time Entry Do Not Fully Capture the True Student Types

,



Note: This figure depicts the way that groups in the noncompliance framework map into the unobserved willingness to wait to enter kindergarten and connects those groups with the traditional groups of early, on-time, and late entry. The values $p_{B,}$ and p_A represent the probability of waiting to enter before the cutoff and after the cutoff respectively.

Defining treatment as waiting a year to enter kindergarten also serves as a better treatment variable than using continuous "entry age." There are intuitive and econometric reasons for

[4]The characterization that students only have two choices is, in fact, accurate. Early entry for students assigned not to wait would mean enrolling three-year-old children who are about to turn four. Likewise redshirting for students assigned to wait would mean enrolling at almost seven years old (in violation of the compulsory schooling laws in 36 states).

this. The intuitive reason is that waiting or not waiting is the choice that decision-makers face. Knowing the causal effect (while interesting, academically perhaps) is not behaviorally relevant because parents and policy makers cannot manipulate it separately from testing age, relative age, and their interactions. For them the effect of interest is not that of being a year older (separate from everything that comes with it), but of waiting (along with everything that comes with it).[5]

Econometrically, my definition of treatment overcomes two problems with using "entry age." The first is an issue of fundamental unidentifiablilty, reflected from the intuition. Just as decision-makers generally can not separately manipulate entry age, econometiricians generally cannot separately identify the effects of entry age from testing age (Angrist and Pischke, 2008).[6] The second econometric issue is a failure of monotonicity. When there is noncompliance, using continuous age with either a month (or day) of birth instrument or the variation around the birthday cutoff violates the monotonicity assumption necessary to identify the school-entry LATE (Barua and Lang, 2016). A binary treatment variable like waiting to enter kindergarten overcomes these issues, something Black et al. (2011); Dhuey et al. (2019) apply in their respective settings.

A.4. Econometric Appendix

A.4.1 Recovering Control Complier Means

Consider the following intuition: Among children who do not wait to enter kindergarten we know that the average scores just after November 1 are $\mu_{0,rc}(r_1)$ (since only reluctant compliers enter early). At the same time, the average score just before November 1 must be a weighted average of both eager and reluctant compliers. If I let $p_g(r) = P(G_i = g|R_i = r)$, I can write these limits as

$$\lim_{r \to r_1^+} \mathbb{E}[Y_i|D_i = 0, r] = \mu_{0,rc}(r_1)$$

$$\lim_{r \to r_1^-} \mathbb{E}[Y_i|D_i = 0, r] = \frac{p_{ec}(r_1)}{p_{ec}(r_1) + p_{rc}(r_1)}\mu_{0,ec}(r_1) + \frac{p_{rc}(r_1)}{p_{ec}(r_1) + p_{rc}(r_1)}\mu_{0,rc}(r_1)$$

[5] This does make treatment a black box of sorts. Although I cannot disentangle the effects of absolute age, relative age, human capital acquisition in the intervening time, and other moving pieces in this paper, recognizing their dependence and finding estimable, policy-relevant treatment effects is the first order concern. Understanding the effects of these individual mechanisms insofar as it is possible seems to be an interesting and important area of future research.

[6] Furthermore, the effect of confounders like testing age and relative age likely varies with entry age—making each even more challenging to separately identify.

where the limit from the left is a weighted average over both complier groups and the limit from the right is only for reluctant compliers. Because the probabilities and expectations are estimable from their sample analogues, the first equation identifies $\mu_{0,rc}$ and together the two identify $\mu_{0,ec}$—the control complier mean at the cutoff.

The limits of the expectations can be estimated by conditioning on D, and I can estimate $p_g(r_1)$ from the limits of $\mathbb{E}[D_i|r]$ at r_1 as long as $p_g(r)$ is continuous for all g at November 1. With $\mu_{0,rc}(r_1)$ defined, $\mu_{0,ec}(r_1)$ can be recovered by algebraic manipulation.

Specifically, the share of students who wait right before the cutoff, $p_{at}(r_1)$ can be estimated as $p_{at}(r_1) = \lim_{r \to r_1^-} \mathbb{E}[D_i|r]$, the share of eager compliers is the share of students induced to wait at the cutoff is $p_{ec}(r_1) = \lim_{r \to r_1^+} \mathbb{E}[D_i|r] - \lim_{r \to r_1^-} \mathbb{E}[D_i|r]$, and the remaining students are reluctant compliers $p_{rc} = \lim_{r \to r_1^+} \mathbb{E}[D_i|r]$. The algebreic definition of $\mu_{0,ec}$ is

$$\mu_{0,ec}(r_1) = \frac{p_{ec} + p_{rc}}{p_{ec}} \lim_{r \to r_1^-} \mathbb{E}[Y_i|D_i = 0, r] - \frac{p_{rc}}{p_{ec}} \lim_{r \to r_1^+} \mathbb{E}[Y_i|D_i = 0, r]$$

A.4.2 Functional Form Test for Negative Selection of Always Takers

My third approach to test for selection in levels exploits variation in average achievement of students who enter without waiting as the share of redshirters increases. As the birthdays approach to November 1 (from the left), the share of redshirters increases monotonically as "marginal" always takers are induced to redshirt.[7] As the probability of waiting increases, the composition of students who enter without waiting changes, and so the slope of average scores among children who enter ($\mathbb{E}[Y_i|D = 0, r]$) captures both the causal effect of being one day older on scores and the changing composition. This has implications for selection in levels. If the "marginal" always takers are positively selected in levels (relative to the students who enter), inducing them to wait will reduce the average scores among the remaining students who enter. In this scenario, the composition changes more as r increases, implying that the average scores should fall more quickly than the causal effect of r. On the other hand, if marginal always takers are negatively selected in levels, the compositional change will increase the average, implying a slope greater than the causal effect of r. The challenge is that the true effect of r and the compositional change are not separately identified in general.

Child development offers a theoretical insight that I turn into a test for negative selection. Because students enter kindergarten at young ages, the effect of an additional month of "entry age" is not thought to be uniform. In fact, the theory suggests that the additional time

[7]Formally these "marginal" always takers are children with $U \in [0.00, 0.18)$ who do not wait to enter given their observed birthday but who would redshirt if they turned five on November 1.

should be more valuable to younger students than to older students: A month of maturity and experience is relatively more to a five year old than to a six year old (Deming and Dynarski, 2008).[8] In my context theory implies two things. First, the slope of $\mathbb{E}[Y_i(0)|r]$ (the causal effect) should always be negative—students with later birthdays enter younger and perform worse—and second, the slope of $\mathbb{E}[Y_i(0)|r]$ should be more negative closer to the cutoff—younger students would benefit more from a given change in absolute age.

The theory suggests that the slope of $\mathbb{E}[Y_i|D = 0, R = r]$ will only be higher closer to the cutoff in the presence of negative selection. I test this theory by comparing the slope of $\mathbb{E}[Y_i|D = 0, R = r]$ for students who turn five between March 1 and July 15 with those who turn five between August 15 and November 1, pooling across years.

$$
\begin{aligned}
Y_i = {} & b_0 + b_1 \mathbb{1}\left(\text{bday}_i \in [\text{Mar 1, July 15}]\right) \\
& + b_2 \text{bday}_i \cdot \mathbb{1}\left(\text{bday}_i \in [\text{Mar 1, July 15}]\right) + b_3 \text{bday}_i \cdot \mathbb{1}\left(\text{bday}_i \in [\text{Aug 15, Nov 1}\right) + e_i
\end{aligned}
$$

From this, I can test the theoretically predicted null hypothesis $b_2 \geq b_3$. With n selection in levels or positive selection in levels, the test will fail to reject the null because the slope at b_2 will be greater (less negative) than the slope of b_3.

I find that "marginal" always takers are negatively selected relative to compliers. Table A.1 shows the results. Both b_2 and b_3 are negative, as suggested by the theory, but the effect in March-July (-0.0008)[9] is *not* greater than the effect in August-November (-0.0005). I reject the null that $b_2 \geq b_3$ at $p = 0.002$ level in the full sample and at similar levels for most subgroups, implying that "marginal" always takers are negatively selected in levels relative to compliers. Interestingly, the selection in levels between always takers and eager compliers is not the same as between eager and reluctant compliers. For example, here we cannot reject the null of no selection in levels among boys who redshirt, and we have somewhat stronger statistical evidence of that low-income children negatively select into redshirting. As in the previous section, the selection results for black students are suggestive but not significant at conventional levels.

[8]Although the first explicit assertion of this theory I have found is in 1997 (Morrison et al., 1997), the intuition behind this idea is visible in age childhood assessments like the Peabody Picture Vocabulary Test (1981, 1997, 2018). Deming and Dynarski (2008) make the clearest argument.

[9]Because there is so little noncompliance in the March-July region, the slope of $\mathbb{E}[Y_i|D = 0, R = r]$ should be a relatively good estimate of $\mathbb{E}[Y_i(0)|R = r]$ in that region. Interestingly, the other region with very little noncompliance (December-February) has a similar slope for treated outcomes (-0.0007). This is what the developmental theory would suggest, as a relative age effect would be similar for a March student who enters and a February student who waits. The fact that it is largely stable on the December-July window also suggests that the parallel trends assumption employed in Section 4 to compare the November and December LATEs is quite reasonable. In fact, it may well be true along the whole support.

Table A.1: Always Takers are Negatively Selected on Average and in Most Subgroups

Effect of Birthday on Third Grade Math Scores ($\frac{SD}{1000}$)			
	March 1 - July 15	August 15 - November 1	Difference
All Students	-0.784***	-0.494***	-0.290
N= 771,869	(0.036)	(0.098)	[$p = 0.002$]
Low SES	-0.662***	-0.469***	-0.198
N=339,517	(0.049)	(0.114)	[$p = 0.054$]
Higher SES	-0.834***	-0.301***	-0.533
N=417,475	(0.047)	(0.130)	[$p = 0.000$]
Black	-0.867***	-0.679***	-0.187
N=157,368	(0.0700)	(0.160)	[$p = 0.143$]
White	-0.788***	-0.132	-0.656
N= 528,034	(0.0435)	(0.113)	[$p = 0.000$]
Girls	-0.930***	-0.310*	-0.620
N= 387,432	(0.0481)	(0.123)	[$p = 0.000$]
Boys	-0.637***	-0.651***	0.014
N=384,094	(0.0523)	(0.136)	-

Note: This table compares the slope of test scores for students who do not wait to enter kindergarten over different ranges of birthdays. Estimates come from a liner regression over two on disjoint samples with uniform weighting. Standard errors allow for arbitrary variance-covariance structure within schools. The sample includes students who meet the following criteria: entering kindergarten in Michigan public schools in the 2002-02 to 2014-15 school years; turning five between March 1 and July 15 or between August 15 and November 1, 2002-2013; and taking state math exams in third grade. Hypothesis tests are one sided tests that the slope in the March to July period is less negative than the slope in the August to November period.

A.5. Empirical Robustness Checks

A.5.1 Regression Discontinuity Robustness Checks

This section explores robustness of the regression discontinuity results to potential pitfalls. Table A.2 shows that the first stage, reduced form, and fuzzy RD relationships are not sensitive to the regression specification. In both panels column (1) reports the main specification from the paper. Columns (2) and (3) change the bandwidth. Shrinking the bandwidth reduces power, but gives similar (possibly larger) results. Widening the bandwidth (to the relevant side) increases precision slightly and gives similar (possibly smaller) results. Columns (4) and (5) show that the uniform kernel gives similar results as a triangular and epanichokov kernels. Colums (6) through (8) explore other polynomial approximations of the conditional expectation of scores over birthdays. Using levels shrinks the estimates considerably, but only if done without shrinking the bandwidth. Using a quadratic term decreases power but suggests fairly similar (possibly larger) effects.

Since all of the regressions in the paper are performed without controls, Table A.3 shows

Table A.2: Treatment Effects Are Constant Over Alternative Specifications

Panel A: Eager Compliers Around November 1	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
First Stage	0.225***	0.221***	0.237***	0.222***	0.225***	0.315***	0.242***	0.218***
	(0.015)	(0.021)	(0.013)	(0.016)	(0.016)	(0.009)	(0.015)	(0.022)
Reduced Form	0.073*	0.091	0.071**	0.074*	0.070*	0.043**	0.076*	0.075
	(0.032)	(0.047)	(0.026)	(0.036)	(0.035)	(0.016)	(0.033)	(0.050)
Fuzzy RD	0.325*	0.413	0.299**	0.331*	0.313*	0.138**	0.314*	0.342
	(0.144)	(0.215)	(0.112)	(0.162)	(0.155)	(0.050)	(0.134)	(0.229)
Bandwidth Around Cutoff	[-30,30]	[-15,15]	[-90,30]	[-30,30]	[-30,30]	[-30,30]	[-7,7]	[-30,30]
Kernel	Uniform	Uniform	Uniform	Triangular	Epanichokov	Uniform	Uniform	Uniform
Polynomial	Linear	Linear	Linear	Linear	Linear	Levels	Levels	Quadratic
Observations	15,066	7,535	31,775	15,066	15,066	15,066	3,669	15,066

Panel B: Reluctant Compliers Around December 1	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
First Stage	0.478***	0.460***	0.513***	0.468***	0.471***	0.540***	0.481***	0.454***
	(0.014)	(0.019)	(0.009)	(0.015)	(0.014)	(0.009)	(0.014)	(0.020)
Reduced For,	0.111**	0.126**	0.078**	0.120**	0.117**	0.054**	0.102**	0.132**
	(0.034)	(0.048)	(0.028)	(0.038)	(0.036)	(0.017)	(0.036)	(0.050)
Fuzzy RD	0.232**	0.274**	0.153**	0.256**	0.248**	0.099**	0.213**	0.290**
	(0.071)	(0.105)	(0.054)	(0.081)	(0.078)	(0.031)	(0.076)	(0.112)
Bandwidth Around December 1	[-30,30]	[-15,15]	[-30,90]	[-30,30]	[-30,30]	[-30,30]	[-7,7]	[-30,30]
Kernel	Uniform	Uniform	Uniform	Triangular	Epanichokov	Uniform	Uniform	Uniform
Polynomial	Linear	Linear	Linear	Linear	Linear	Levels	Levels	Quadratic
Observations	14,873	7,401	31,156	14,873	14,873	14,873	3,330	14,873

Note: This table compares estimates of the average effect of waiting using different RD specifications. The sample includes students who meet the following criteria: entering kindergarten in Michigan public schools in the 2013-14 or 2014-15 school years; turning five within the bandwidth of the relevant cutoff; and taking state math exams in third grade.

Table A.3: Treatment Effects Are Fairly Constant With Controls

Panel A: Eager Compliers	All Students			Low-SES Students			Higher-SES Students		
First Stage	0.237***	0.234***	0.233***	0.312***	0.310***	0.314***	0.143***	0.138***	0.133***
	(0.013)	(0.013)	(0.013)	(0.017)	(0.017)	(0.017)	(0.020)	(0.019)	(0.018)
Reduced Form	0.071**	0.060*	0.060**	0.046	0.043	0.041	0.088*	0.079*	0.058
	(0.026)	(0.023)	(0.023)	(0.033)	(0.032)	(0.032)	(0.035)	(0.033)	(0.035)
Fuzzy RD	0.299**	0.255*	0.256*	0.148	0.138	0.131	0.618*	0.569*	0.435
	(0.112)	(0.100)	(0.100)	(0.107)	(0.102)	(0.102)	(0.248)	(0.252)	(0.266)
Controls		✓	✓		✓	✓		✓	✓
School Fixed Effects			✓			✓			✓

Panel B: Reluctant Compliers	All Students			Low-SES Students			Higher-SES Students		
First Stage	0.478***	0.475***	0.463***	0.503***	0.499***	0.498***	0.443***	0.441***	0.415***
	(0.014)	(0.014)	(0.013)	(0.016)	(0.016)	(0.016)	(0.022)	(0.022)	(0.020)
Reduced Form	0.104***	0.078**	0.088***	0.086*	0.086**	0.092**	0.068	0.066	0.076*
	(0.028)	(0.025)	(0.025)	(0.034)	(0.033)	(0.035)	(0.037)	(0.036)	(0.038)
Fuzzy RD	0.218***	0.163**	0.191***	0.171*	0.172**	0.186**	0.154	0.15	0.184*
	(0.058)	(0.052)	(0.055)	(0.068)	(0.065)	(0.069)	(0.085)	(0.081)	(0.092)
Controls		✓	✓		✓	✓		✓	✓
School Fixed Effects			✓			✓			✓

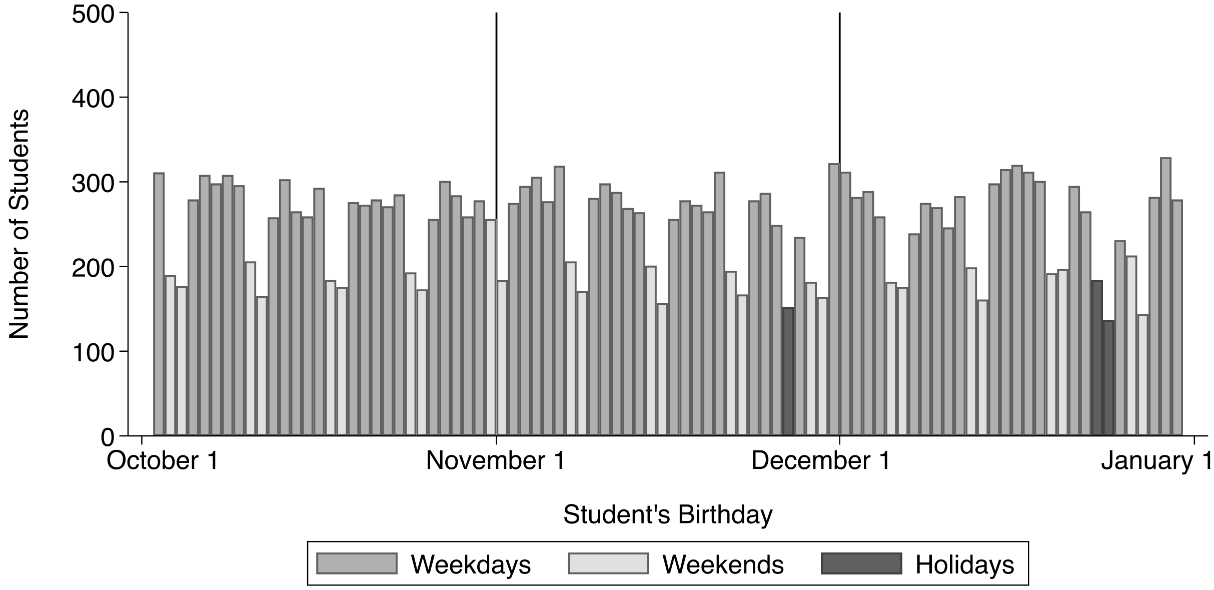
Note: This table compares estimates of the average effect of waiting using different RD specifications. The sample includes students who meet the following criteria: entering kindergarten in Michigan public schools in the 2013-14 or 2014-15 school years; turning five within 90 days of either cutoff; and taking state math exams in third grade.

that the first stage, reduced form, and fuzzy RD relationships are not sensitive to the regression specification. The table focuses on the low-income high-income results to demonstrate that the selection on gains looks similar persists when controls are added. Controls included race, sex, poverty, English language learner, enrollment by school of choice, neighborhood characteristics (percent black, percent Hispanic/other-nonwhite, employment rate, median household income, percent with no high school degree, percent with a bachelors or more), and an interaction of low-income and sex. Including these controls did not do much to the effects, maybe shrinking them toward zero. Including (kindergarten) school fixed effects may reduce the eager complier LATE a little and possibly increases the reluctant complier LATE, but the differences are imprecise.

A key assumption in the regression discontinuity framework is that potential outcomes are continuously distributed around the cutoffs. One violation of this assumption could be if different types of students tend to have birthdays on either side of the cutoff. Because assignment is not enforced it seems unlikely that a significant number of parents strategically plan births to fall on one side of the cutoff or another; however, other factors may influence birth timing. To explore this Figure A.2 reports the counts of students with each birthday.

It is immediately apparent from Figure A.2 that birthdays are not uniformly distributed. For example, children are less likely to be born on weekends and holidays. If these patterns are unrelated to baseline family characteristics, this non-uniformity would not be a problem;

Figure A.2: Birthdays Are Consistent (if not Uniform) around the Cutoffs

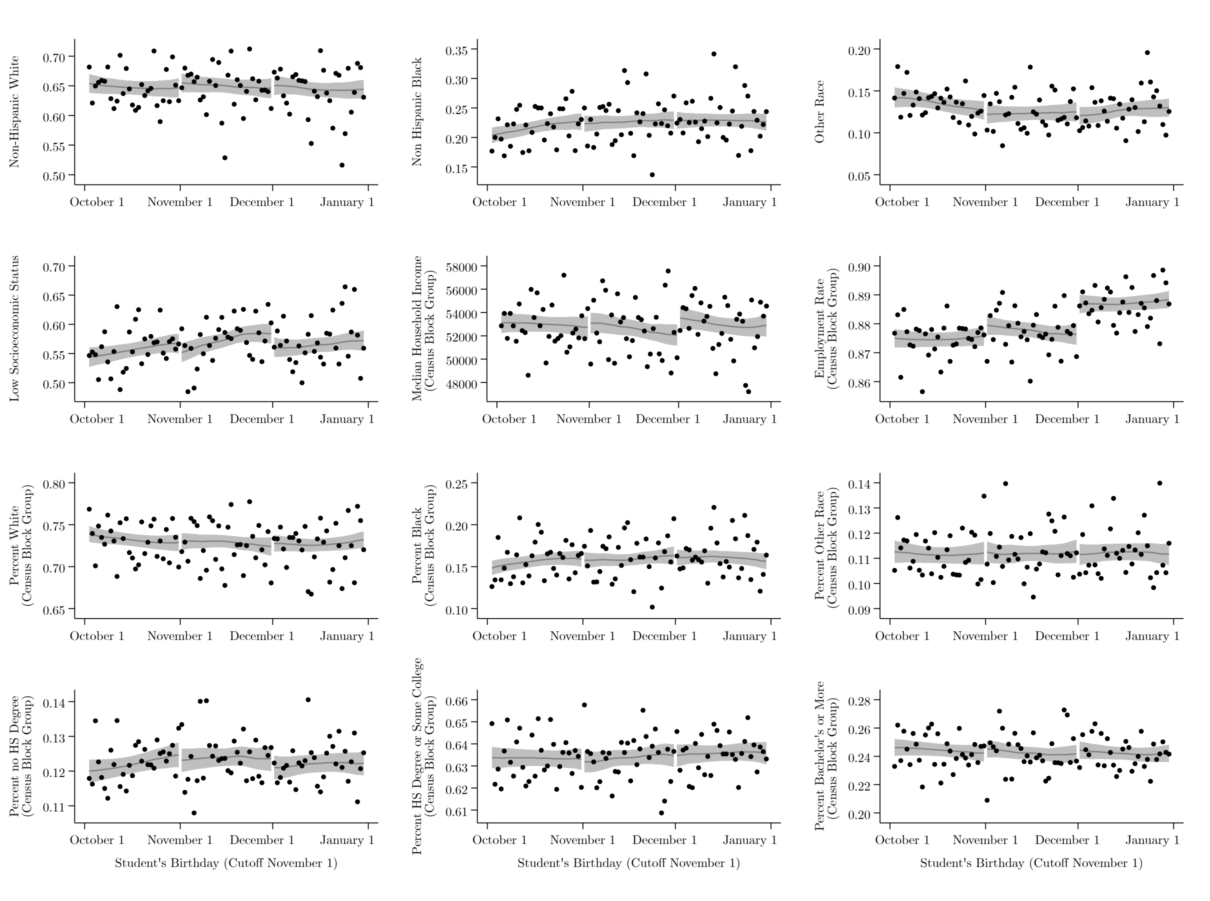


Note: This figure reports the count of students with each birthday. It also shows whether the student was born on a weekend or national holiday. The sample includes students who meet the following criteria: entering kindergarten in Michigan public schools in the 2013-14 or 2014-15 school years; turning five within 30 days of either cutoff; and taking state math exams in third grade.

however, it is well known that more affluent families are more likely to have their children on weekdays (mostly because of schedule inductions). In my main analytical sample students born on Saturdays are 3.9 percentage points more likely to be low-income; students born on Sundays are 5.6 percentage points more likely to be low-income; and students born on Thanksgiving, Christmas Eve, or Christmas are 6.9 percentage points more likely to be low-income—all relative to students born on non-holiday Mondays (0.54). This is concerning since neither cutoff falls in the middle of the week. As a result the limit could be biased towards an unrepresentative sample. Fortunately, Table A.3 already suggested this was not a problem in practice, and Table A.4 shows that the results are not sensitive to the exclusion of students born on weekends.

A second way to explore the assumption of continuity at the cutoffs is to explore covariates. Figure A.3 displays control variables plotted over the support of birthdays. Overall these comparisons suggest that students born right before and after each cutoff are very similar. When run as regressions only two of the twenty four comparisons are significantly different at the $p = 0.05$ level: low-income at the December cutoff and and census block group employment rate at the December cutoff. Both of these differences are visible in Figure A.3. I do not know why children who turn five right after December 1 have 1.1 percentage point

Figure A.3: Most Covariates are Fairly Smooth around the Cutoffs



Note: This figure plots the conditional expectations of covariates over birthdays separately for each side of the cutoffs. Local polynomials are shown together with daily averages for the characteristics. Local linear regressions reveal very similar results.

higher employment rates in their census block groups, but the difference is not present or significant in the low-income or higher-income subsamples, suggesting that conditional on the split there are no differences and that in the whole sample if any differences are present, they are small.

The two biggest apparent threats to identification, however, come not from *ex ante* differences in the population, but from other programs that could create discontinuities in the potential outcomes at the threshold. In the IV setting, one could think about these as violations of the exclusion restriction whereas the day-of-the-week birthday issues were more issues of independence.

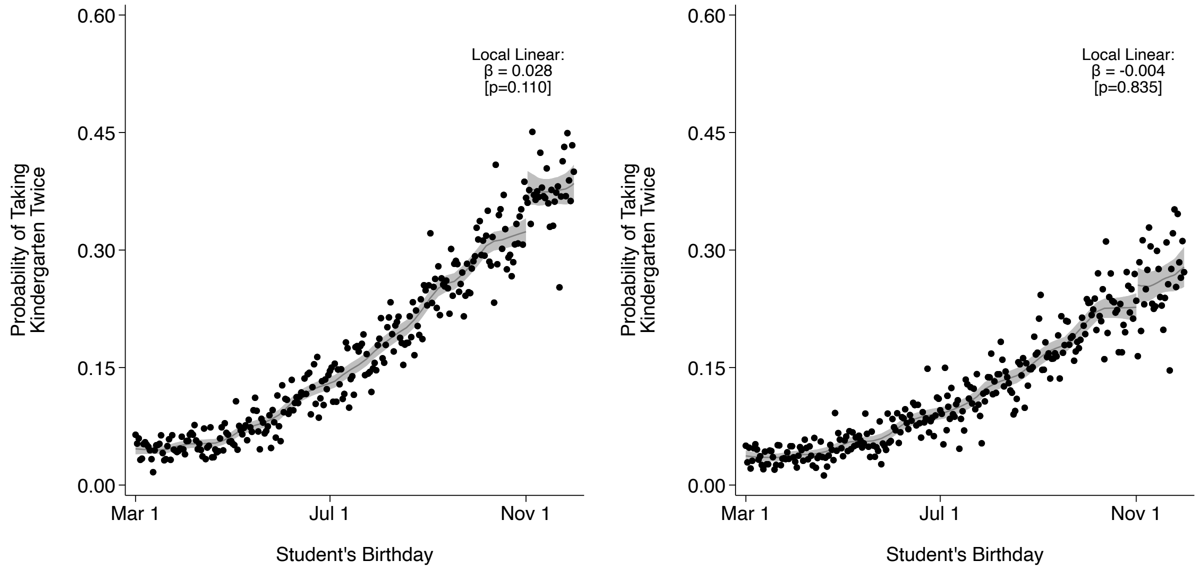
The first threat from from Michigan's Great Start Readiness Program (GSRP). The state of Michigan operates GSRP as a program for four-year-old children with special needs, who come from lower-income or divorced families, or who meet other risk criteria. The threat to identification is that for my main analysis cohort students were eligible for GSRP in 2012 if they turned five between November 2, 2012 and December 1, 2013 (older children could have participated in 2011). The threat to identification is that if GSRP takeup is discontinuous at November 1 *and* GSRP affects third grade test scores, the potential outcomes would not be comparable on either side of the cutoff. Indeed, participation in GSRP increases by about five percentage points at the November 1 cutoff (among low- and higher-income subgroups), but it does not jump at the December 1 cutoff as would be expected. Table A.4 shows that dropping GSRP participants does not erase the treatment effect in the full sample. Interestingly, conditioning on GSRP and income does not change the selection on gains among high-income or the lack thereof among low-income (if anything low-income GSRP participants select negatively on gains into waiting).

The other potential problem could be developmental kindergarten programs. Developmental kindergarten programs are "kindergarten" classes in public schools that are intended to be the first of two years of kindergarten before a student enters first grade. If families are completely free to enroll their students in developmental kindergarten programs, they would only be a mechanism for the effect, not a confounder; however, anecdotally I am aware of some schools only allow early entrants to enroll in developmental programs (in other words while districts are required by the state to accept early entrants, they can require early entrants to enroll in developmental programs rather than in traditional programs in their first year. This would violate exclusion if reluctant compliers who would not have taken kindergarten twice had they turned five by November 1 are then forced to take kindergarten twice, increasing their third grade achievement conditional on entry decision discontinuously at the cutoff.

This is particularly concerning because Table A.7 shows that reluctant compliers are indeed more likely to take kindergarten twice. Note that taking kindergarten twice is not a violation of the exclusion/continuity assumption. In fact, it may be a mechanism through which reluctant compliers achieve higher test scores than eager compliers. It is *forcing* reluctant compliers to take kindergarten twice after the cutoff when they would not have done so before the cutoff that would be problematic. If this happened, it would bias the estimates of the eager complier LATE upward and could lead to erroneous rejections of effect homogeneity between eager and reluctant compliers.

Unfortunately, there is no comprehensive record of which schools and which districts had requirements like this. I do however know which schools offer official developmental kindergarten programs starting in 2014. This is a superset of the set of schools that force early entrants to take developmental kindergarten. Figure A.4 shows that restricting the sample to schools that do not have these programs is sufficient to drive the discontinuity in taking kindergarten twice to zero (which could be over controlling if eager and reluctant

Figure A.4: Schools with Official Developmental Kindergarten Programs May Create Discontinuities in Repetition



Note: This figure shows the rates at which students take kindergarten twice. Both figures show local polynomial approximations of the conditional expectation with seven-day bins, and they daily averages for the rate of taking kindergarten twice among the subsample of students who enter kindergarten in 2013 without waiting. The panel on the left shows all schools and the panel on the right is restricted to schools that do not have official developmental kindergarten programs in 2014.

compliers do have different propensities to take kindergarten twice). Table A.4 reports the regression results from dropping these schools. Both the eager complier and the reluctant complier LATEs shrink a little in this specification. Since about 20% of the students in my sample attend these schools, the estimates are also a good deal less precise. The subgroup analyses are not reported, but show the same trends as in all the other specifications.

Taken together these results suggest that both sets of Local Average Treatment Effects are well identified and well estimated—both for the whole population and for the relevant subpopulations.

A.5.2 Testing Parallel Trends

Above I demonstrated that the effects estimated at the different cutoffs are significantly different from one another for students from higher-income families. Interpreting these differences as positive selection on gains hinges on the parallel trends assumption proposed in Section 1.2 that for a given type of child (characterized by u and x) changes in scores over birthdays would have the same slope for those who start as for those who wait. If this

Table A.4: Treatment Effects Constant Dropping Possible Problems

Panel A: Eager Compliers Around November 1	(1)	(2)	(3)	(4)	(5)
First Stage	0.225***	0.240***	0.226***	0.176***	0.227***
	(0.015)	(0.017)	(0.015)	(0.018)	(0.018)
Reduced Form	0.073*	0.083*	0.072*	0.068^{+}	0.060
	(0.032)	(0.036)	(0.033)	(0.037)	(0.037)
Fuzzy RD	0.325*	0.346*	0.318*	0.389^{+}	0.263
	(0.144)	(0.152)	(0.144)	(0.212)	(0.162)
Restrictions	None	Drop weekends and holidays	Day-of-week and holiday FE	Drop GSRP	Drop Young Fives Schools
Observations	15,066	11,567	15,066	11,552	11,606

Panel B: Reluctant Compliers Around December 1	(1)	(2)	(3)	(4)	(5)
First Stage	0.478***	0.479***	0.482***	0.487***	0.465***
	(0.014)	(0.016)	(0.014)	(0.016)	(0.016)
Reduced Form	0.111**	0.086*	0.094**	0.087*	0.105**
	(0.034)	(0.040)	(0.035)	(0.041)	(0.038)
Fuzzy RD	0.232**	0.180*	0.195**	0.179*	0.226**
	(0.071)	(0.083)	(0.073)	(0.083)	(0.082)
Restriction		Drop weekends and holidays	Day-of-week and holiday FE	Drop GSRP	Drop Young Fives Schools
Observations	14,873	11,262	14,873	10,677	11,573

Note: This table compares estimates of the average effect of waiting using different RD specifications. The sample includes students who meet the following criteria: entering kindergarten in Michigan public schools in the 2013-14 or 2014-15 school years; turning five within 90 days of either cutoff; taking state math exams in third grade; and the additional restrictions for each column.

parallel trends assumption is not met, then differences estimated at the different cutoffs could be attributable to differences between the compliers at each cutoff or to heterogeneity in effects over birthdays. This subsection evaluates on testable implication of this parallel trends assumption.

To assess the plausibility of the parallel trends assumption for third-grade math scores, I explore whether test scores in third grade have a similar slope over birthdays as do test scores in fourth grade. This comparison arises from the insight that parallel trends imply that the slope of third grade test scores over birthdays should be the same for students whether they had waited or started. In other words, they should be parallel for students who take tests at different ages (e.g., third grade at age eight vs age nine). But if this assumption is true, scores should also be parallel for students in different grades (e.g., third grade at age eight vs fourth grade at age nine). Unlike potential third-grade test scores, test scores in third grade and fourth grade are both directly observable.

To estimate whether there are differences in the slopes of test scores, I test the a null hypothesis related to parallel trends in potential third grade achievement:

$$\mathcal{H}: \frac{\partial \mathbb{E}[Y_{G3}|u,x]}{\partial r} = \frac{\partial \mathbb{E}[Y_{G4}|u,x]}{\partial r}$$

The partial derivative $\frac{\partial Y_W}{\partial r}$ reflects the direct effect of being one day younger when taking an exam—a well-documented negative relationship. Because changes in birthdates r change the probability of waiting $p_z(r)$, this partial derivative is not identified by within-sample changes in test scores over birthdays; however, when children are required to wait $z = 2$, there is no change in $p_2(r)$, and any change in test scores stems directly from differences in age. Because of this I estimate $\frac{\partial \mathbb{E}[Y_G|u,x]}{\partial r}$ for among the sample of students with birthdays after December 1, for grades 3, 4, 6, and 8 (to hold the sample constant I restrict the sample to student who start kindergarten before 2008 and who have eigth grade test scores).

Table A.5 demonstrates that the trends are close to parallel. The first four rows of Table A.5 show how a one-week change affects scores. Unsurprisingly, all coefficients are negative and of very small magnitude. The last two columns show tests of the null hypotheses that the slopes of third and fourth grade are equal and that the slopes in all grades are equal. The comparison between third and fourth grade is the most relevant to comparing third grade scores of students who either started kindergarten at five or waited an additional year. Column five shows that in the population and for most groups I can't reject the hypothesis that grade 3 and grade 4 are different, and comparing columns one and two reveals that any difference is quite small.

There do seem to be some deviations from the parallel trends in later grades, but the

Table A.5: Deviations from Parallel Trends Are Extremely Small

	Change in Test Scores from Being One Week Younger					
	Third Grade	Fourth Grade	Sixth Grade	Eighth Grade	Test of $\beta_{G3} = \beta_{G4}$	Test of all $\beta_G = \beta$
All Students	-0.007***	-0.006***	-0.005***	-0.001	[$p = 0.171$]	[$p = 0.000$]
N= 95,403	(0.001)	(0.001)	(0.001)	(0.001)		
Higher-Income	-0.010***	-0.007***	-0.005**	-0.001	[$p = 0.023$]	[$p = 0.000$]
N= 57,438	(0.002)	(0.002)	(0.002)	(0.002)		
Low-Income	-0.005**	-0.006**	-0.005**	-0.003	[$p = 0.885$]	[$p = 0.291$]
N= 37,169	(0.002)	(0.002)	(0.002)	(0.002)		
Black	-0.004	-0.004	-0.001	0.002	[$p = 0.855$]	[$p = 0.098$]
N= 18,637	(0.003)	(0.003)	(0.003)	(0.003)		
White	-0.007***	-0.006***	-0.005**	-0.001	[$p = 0.128$]	[$p = 0.000$]
N= 67,175	(0.002)	(0.002)	(0.002)	(0.002)		
Girls	-0.009***	-0.008***	-0.005**	-0.002	[$p = 0.466$]	[$p = 0.000$]
N= 47,292	(0.002)	(0.002)	(0.002)	(0.002)		
Boys	-0.006**	-0.004*	-0.004*	0.000	[$p = 0.244$]	[$p = 0.000$]
N= 48,093	(0.002)	(0.002)	(0.002)	(0.002)		

Note: This table reports a test of whether test scores really do evolve in parallel for students of different ages. It reports the slope of test scores over birthdays (reported in weeks) and tests of equality among the coefficients. The sample is restricted to students with birthdays after December 1 who entered kindergarten in Michigan public schools between 2002-2008 and for whom I observe test scores in third, fourth, sixth, and eighth grade. $^{+}p < 0.1$, $*p < 0.05$, $**p < 0.01$, $***p < 0.001$.

deviations are such that they would bias me against finding the type of heterogeneity I document. I reject the null hypothesis of parallel trends when considering all four grades (and for higher-income families between third and fourth grade). If anything the slopes are becoming less steep over time—likely because being a week younger means less in eighth grade than it does in third grade. These deviations from parallel trends are small in magnitude, however. For example, a two-month change in birthday could only explain about 0.01 standard deviations of the heterogeneity in treatment effects. The deviations from parallel trends also go in the opposite direction as the heterogeneity: if the scores of younger students are more negative, then treatment effects measured at October or November should be smaller than those measured at December all else equal. In that case, the violation of parallel trends leads me to under-estimate the true amount of positive selection on gains—especially among higher-income families who have the most dramatic change in slope between grades. This bounding logic is similar to the approach of Rambachan and Roth (2022) for bounding differences in differences estimators for the largest plausible deviation from parallel trends.

A.5.3 Bounding the Effect on Always Takers

A.5.3.1 Monotonic Selection

As an alternative to assuming that the achievement of third graders who enter kindergarten without waiting evolve linearly over the unobserved cost of waiting, U, this subsection explores a weaker monotonicity assumption:

$$m_0(u, r) \leq \mu_0(u', r) \iff u \leq u'$$

Monotonicity assumptions are increasingly common in the MTE literature since Mogstad et al. (2018) (Kowalski, 2022a, see for example). Monotonicity is attractive because it does not impose a functional form assumptions on the untreated outcomes, but still yields informative bounds on the treatment effects under certain conditions. I weaken these monotonicity assumptions by assuming monotonicity only in the expectation across groups at the cutoff rather than for all U.

As we know that $\mu_{0,ec}(r_1) < \mu_{0,rc}(r_1)$, the bite of this assumption is in bounding $\mu_{0,at}(r_1)$ below $\mu_{0,ec}(r_1)$. This inequality is not strong enough to identify a point estimate of the τ_{at}, but it can identify an informative bound on the average treatment effect for always takers:

$$\tilde{\tau}_{at} \equiv \mu_{1,at} - \mu_{0,ec}$$

Because I assume $\mu_{0,at} \leq \mu_{0,ec}$, the statistic $\tilde{\tau}_{at} = 0.64$ is a lower bound on τ_{at}[10] Because $\tilde{\tau}_{at} > 0$, it is an informative bound.[11]

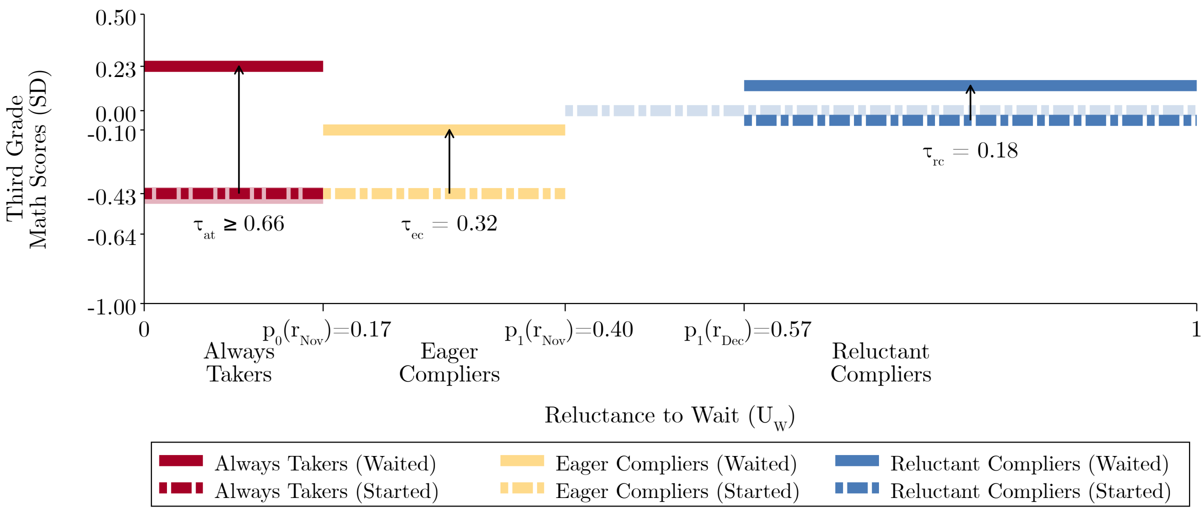
Figure A.5 illustrates these estimates graphically. It plots the upper bound on untreated outcomes over the range of always takers at $\mu_{0,at} \leq -0.42$. This means that we assume that had they not waited, redshirted students at the cutoff would have scored *at least* two fifths of a standard deviation below average. Figure A.5 also shows the implied lower bound on the treatment effects $\tau_{at} \geq \tilde{\tau}_a t = 0.64$. This is a lower bound on the effect, and it is still very large.

To test the null of effect homogeneity between always takers and eager compliers under this weaker assumption I use a test for heterogeneity from in Kowalski (2022a). This test determines whether the eager-complier LATE, τ_{ec}, falls within the range bounded by $\tilde{\tau}_{at}$. The one-tailed test statistic returns a 1 if it rejects treatment effect homogeneity (i.e., whether τ_{ec}

[10] This is because $\mu_{0,ec}$ is an upper bound on the untreated outcomes for always takers so what ever the difference between the bound and the treated outcomes, it is less than the true effect.

[11] More generally, assuming monotonicity across groups will produce an informative bound about τ_{at} under either of two conditions: (1) $\mu_{1,at} - \mu_{0,g} > 0$ and $\mu_{0,g} - \mu_{0,g+1} < 0$ or (2) $\mu_{1,at} - \mu_{0,g} < 0$ and $\mu_{0,g} - \mu_{0,g+1} > 0$.

Figure A.5: Selection on Gains Robust to Weaker Monotonicity Assumption



Note: This figure graphically illustrates the smallest possible effect on always takers under monotonicity. This bound is recoverable from the average of student outcomes by intervention assignment (whether assigned wait to enter kindergarten) and treatment status (whether actually waited) and the auxiliary assumption of weakly monotonic expected untreated outcomes. The sample is comprised of 15,081 students who meet the following criteria: enter kindergarten in Michigan public schools in the 2013-14 or 2014-15 school years; turn five within thirty days of November 1, 2013; and take standard third-grade math tests. Average outcomes for treated and untreated compliers are backed out of observed data and choice probabilities. For bootstrapped standard errors see table 1.5

is outside of the bounds from $\tilde{\tau}_{at}$) and a 0 otherwise:

$$\mathbb{1}\left[(\tilde{\tau}_{at} - \tau_{ec}) \cdot \mathcal{B}_{0,ec,rc} < 0\right]$$

Again for inference I use a nonparametric block bootstrap and report the percent of failures to reject as the p-value. Under this second assumption I reject homogeneity at the $p = 0.011$ level. Full results are reported in Table 1.5.

A.6. Appendix Tables and Figures

Table A.6: Special Education Outcomes Suggest that Always Takers Are Negatively Selected

	Sample Mean	Always Takers	Compliers Wait	Compliers Enter	Difference	LATE
Sample Shares		20.3%	77.3%			
Detailed Special Education:						
Reported Cognitive Impairment	0.003	0.011	0.000	0.001	0.011***	-0.001**
	(0.000)	(0.002)	(0.001)	(0.000)	(0.002)	(0.001)
Reported Emotional Impairment	0.001	0.002	0.001	0.000	0.001	0.001**
	(0.000)	(0.001)	(0.000)	(0.000)	(0.001)	(0.000)
Reported Speech or Language Impairment	0.072	0.097	0.056	0.068	0.041***	-0.013***
	(0.001)	(0.005)	(0.003)	(0.002)	(0.006)	(0.003)
Reported Early Childhood Developmental Delay	0.009	0.022	0.004	0.006	0.019***	-0.002*
	(0.000)	(0.002)	(0.001)	(0.001)	(0.003)	(0.001)

Note: This sample shows the average near-term outcomes for always takers, eager compliers, and reluctant compliers. The sample is comprised of students who started kindergarten in Michigan public schools in the fall 2002-2012 and turned five within thirty days of December 1. Note that in this table I do not restrict to students who took non-accommodated third-grade math exams. Block bootstrapped standard errors for estimated means and differences are given in parentheses, blocking by school with 1000 replications.

Table A.7: Early Elementary School Outcomes for 2013 Sample

	Sample Mean	Always Takers (Wait)	Eager Compliers Wait	Eager Compliers Start	Reluctant Compliers (Start)	(τ_{ec}) $\mu_{1,at} - \mu_{1,ec}$	$\mu_{1,ec} - \mu_{0,ec}$	$\mu_{0,ec} - \mu_{0,rc}$
Sample Share		18%	11%	11%	60%			
Testing Outcomes:								
No Third Grade Math Test	0.047	0.090	0.052	0.036	0.036	0.038***	0.017	0.000
		(0.010)	(0.008)	(0.008)	(0.003)	(0.015)	(0.011)	(0.010)
Took Alternative Test	0.017	0.052	0.018	0.017	0.007	0.034***	0.002	0.010**
		(0.007)	(0.005)	(0.004)	(0.001)	(0.011)	(0.007)	(0.005)
Kindergarten Outcomes:								
School of Choice in Kindergarten	0.231	0.153	0.219	0.200	0.268	-0.066***	0.019	-0.067***
		(0.012)	(0.013)	(0.020)	(0.012)	(0.019)	(0.021)	(0.019)
Kindergarten Attendance Rate	0.938	0.938	0.924	0.931	0.945	0.014***	-0.007	-0.010***
		(0.003)	(0.004)	(0.003)	(0.001)	(0.005)	(0.005)	(0.004)
Repeat Kindergarten	0.240	0.009	0.022	0.219	0.379	-0.014***	-0.197***	-0.160***
		(0.003)	(0.003)	(0.022)	(0.014)	(0.005)	(0.023)	(0.029)
Double Promotion	0.002	0.003	0.006	0.000†	0.001	-0.003	0.008***	-0.003**
		(0.002)	(0.002)	(0.001)	(0.000)	(0.003)	(0.002)	(0.001)
Special Education in Kindergarten	0.124	0.206	0.142	0.146	0.087	0.064***	-0.004	0.060***
		(0.013)	(0.011)	(0.012)	(0.004)	(0.020)	(0.015)	(0.015)
Detailed Special Education Outcomes:								
Cognitive Impairment	0.003	0.014	0.001	0.004	0.001	0.013***	-0.003	0.003*
		(0.003)	(0.002)	(0.002)	(0.000)	(0.005)	(0.003)	(0.002)
Emotional Impairment	0.001	0.002	0.000	0.002	0.000	0.002	-0.002	0.002**
		(0.001)	(0.001)	(0.001)	(0.000)	(0.002)	(0.001)	(0.001)
Speech or Language Impairment	0.088	0.103	0.099	0.109	0.074	0.004	-0.011	0.036***
		(0.008)	(0.008)	(0.011)	(0.004)	(0.014)	(0.014)	(0.014)
Early Childhood Developmental Delay	0.014	0.040	0.015	0.013	0.006	0.024***	0.003	0.007
		(0.006)	(0.005)	(0.004)	(0.001)	(0.009)	(0.006)	(0.004)

Note: This sample shows the average outcomes for always takers, eager compliers, and reluctant compliers facing the November 1, 2013 cutoff. The sample is comprised of 15,990 students who started kindergarten in Michigan public schools in the 2013 or 2014 and turned five within thirty days of November 1. Note that in this table I do not restrict to students who took non-accommodated third-grade math exams. Block bootstrapped standard errors for estimated means and differences are given in parentheses, blocking by school with 1000 replications.
† This cell had a control complier mean slightly below zero because of sampling error.

Table A.8: Early Elementary School Outcomes By Subgroup

Panel A: All Students	Sample Mean	Always Takers	Compliers Wait	Compliers Enter	Difference	LATE
Sample Shares		20.3%	77.3%			
Special Education (K)	0.099	0.170	0.069	0.086	0.101***	-0.018***
	(0.001)	(0.006)	(0.003)	(0.002)	(0.008)	(0.004)
Special Education (3)	0.147	0.210	0.114	0.138	0.096***	-0.025***
	(0.001)	(0.007)	(0.004)	(0.003)	(0.009)	(0.005)
No Third Grade Math Score	0.110	0.154	0.083	0.116	0.071***	-0.033***
	(0.001)	(0.006)	(0.003)	(0.003)	(0.008)	(0.004)
Accommodated Test in Third Grade	0.008	0.013	0.006	0.006	0.008***	0.000
	(0.000)	(0.002)	(0.001)	(0.001)	(0.002)	(0.001)

Panel B: Low-SES Students	Sample Mean	Always Takers	Compliers Wait	Compliers Enter	Difference	LATE
Sample Shares		13.1%	85.0%			
Special Education (K)	0.116	0.279	0.078	0.099	0.201***	-0.022***
	(0.002)	(0.014)	(0.004)	(0.004)	(0.016)	(0.006)
Special Education (3)	0.178	0.330	0.137	0.171	0.193***	-0.035***
	(0.002)	(0.015)	(0.005)	(0.004)	(0.018)	(0.007)
No Third Grade Math Score	0.115	0.223	0.085	0.119	0.138***	-0.035***
	(0.001)	(0.012)	(0.004)	(0.004)	(0.015)	(0.006)
Accommodated Test in Third Grade	0.012	0.033	0.008	0.009	0.024***	-0.001

Panel C: Higher-SES Students	Sample Mean	Always Takers	Compliers Wait	Compliers Enter	Difference	LATE
Sample Shares		26.4%	71.0%			
Special Education (K)	0.088	0.128	0.059	0.074	0.069***	-0.015***
	(0.001)	(0.007)	(0.004)	(0.003)	(0.010)	(0.005)
Special Education (3)	0.123	0.167	0.089	0.108	0.078***	-0.019***
	(0.001)	(0.007)	(0.005)	(0.004)	(0.011)	(0.006)
No Third Grade Math Score	0.105	0.127	0.080	0.110	0.047***	-0.030***
	(0.002)	(0.006)	(0.004)	(0.004)	(0.009)	(0.006)
Accommodated Test in Third Grade	0.005	0.006	0.003	0.003	0.003	0.000
	(0.000)	(0.002)	(0.001)	(0.001)	(0.002)	(0.001)

Panel D: Black	Sample Mean	Always Takers	Compliers Wait	Compliers Enter	Difference	LATE
Sample Shares		7.7%	87.6%			
Special Education (K)	0.091	0.238	0.065	0.082	0.174***	-0.017***
	(0.003)	(0.022)	(0.005)	(0.005)	(0.024)	(0.007)
Special Education (3)	0.153	0.317	0.115	0.161	0.202***	-0.046***
	(0.002)	(0.025)	(0.006)	(0.006)	(0.028)	(0.009)
No Third Grade Math Score	0.127	0.265	0.100	0.131	0.165***	-0.031***
	(0.002)	(0.022)	(0.005)	(0.005)	(0.024)	(0.008)
Accommodated Test in Third Grade	0.011	0.031	0.009	0.010	0.023**	-0.002
	(0.001)	(0.009)	(0.002)	(0.002)	(0.010)	(0.002)

Panel E: White	Sample Mean	Always Takers	Compliers Wait	Compliers Enter	Difference	LATE
Sample Shares		25.4%	72.8%			
Special Education (K)	0.105	0.161	0.072	0.093	0.088***	-0.020***
	(0.001)	(0.007)	(0.004)	(0.003)	(0.009)	(0.005)
Special Education (3)	0.148	0.194	0.115	0.132	0.080***	-0.017***
	(0.002)	(0.007)	(0.005)	(0.004)	(0.010)	(0.006)
No Third Grade Math Score	0.094	0.128	0.066	0.098	0.061***	-0.031***
	(0.001)	(0.006)	(0.004)	(0.003)	(0.008)	(0.005)
Accommodated Test in Third Grade	0.008	0.012	0.006	0.004	0.006**	0.001
	(0.000)	(0.002)	(0.001)	(0.001)	(0.003)	(0.001)

Panel F: Girls	Sample Mean	Always Takers	Compliers Wait	Compliers Enter	Difference	LATE
Sample Shares		16.6%	80.8%			
Special Education (K)	0.065	0.121	0.047	0.059	0.074***	-0.012***
	(0.001)	(0.008)	(0.003)	(0.003)	(0.010)	(0.004)
Special Education (3)	0.101	0.154	0.081	0.099	0.073***	-0.018***
	(0.001)	(0.009)	(0.004)	(0.004)	(0.011)	(0.006)
No Third Grade Math Score	0.106	0.163	0.077	0.114	0.086***	-0.037***
	(0.001)	(0.009)	(0.004)	(0.004)	(0.011)	(0.005)
Accommodated Test in Third Grade	0.005	0.010	0.004	0.003	0.006**	0.000
	(0.000)	(0.002)	(0.001)	(0.001)	(0.003)	(0.001)

Panel G: Boys	Sample Mean	Always Takers	Compliers Wait	Compliers Enter	Difference	LATE
Sample Shares		23.8%	74.0%			
Special Education (K)	0.132	0.203	0.092	0.115	0.111***	-0.023***
	(0.002)	(0.009)	(0.005)	(0.004)	(0.012)	(0.006)
Special Education (3)	0.191	0.246	0.149	0.180	0.097***	-0.031***
	(0.002)	(0.010)	(0.006)	(0.005)	(0.014)	(0.007)
No Third Grade Math Score	0.114	0.148	0.089	0.118	0.059***	-0.029***
	(0.001)	(0.007)	(0.004)	(0.004)	(0.009)	(0.006)
Accommodated Test in Third Grade	0.011	0.016	0.008	0.009	0.008**	-0.001
	(0.000)	(0.003)	(0.002)	(0.001)	(0.004)	(0.002)

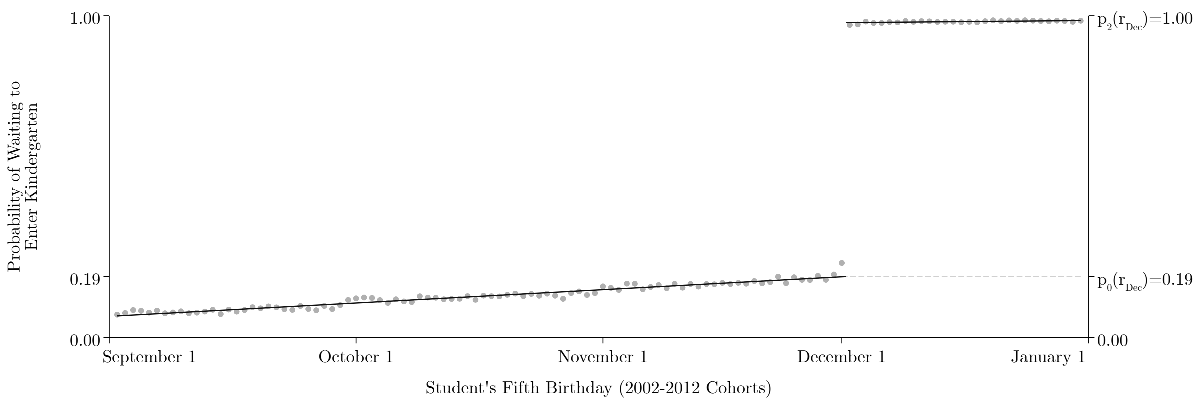
Note: This sample shows the average near-term outcomes for always takers, eager compliers, and reluctant compliers. The sample is comprised of 15,990 students who started kindergarten in Michigan public schools in 2002-13 and turn five within 90 days of December 1. Note that in this table I do not restrict to students who took non-accommodated third-grade math exams. Block bootstrapped standard errors for estimated means and differences are given in parentheses, blocking by school with 1000 replications.

Table A.9: Different Measures of ATEs

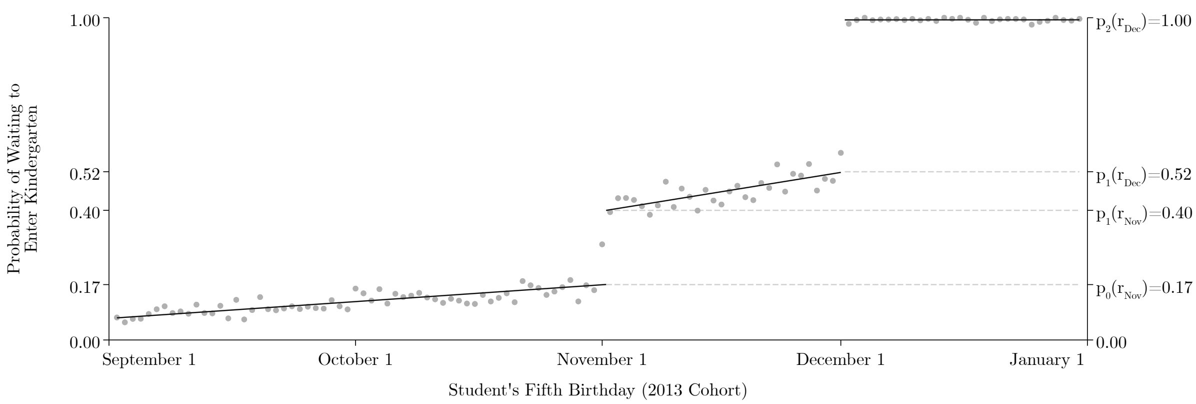
	Reluctant Complier Effect	Eager Complier Effect	Always Taker Effect	Average Effect (RC Bound)	Average Effect (EC Bound)
All Students	0.230***	0.333***	0.838***	0.361***	0.373***
	(0.071)	(0.134)	(0.198)	(0.082)	(0.090)
Higher-SES Students	0.156*	0.596***	1.179***	0.481***	0.550***
	(0.084)	(0.238)	(0.302)	(0.128)	(0.153)
Low-SES Students	0.170***	0.153	0.168	0.164**	0.163**
	(0.069)	(0.106)	(0.154)	(0.071)	(0.074)
Female	0.204**	0.201	0.940***	0.345***	0.345***
	(0.087)	(0.136)	(0.189)	(0.094)	(0.100)
Male	0.230***	0.423***	0.959***	0.369***	0.397***
	(0.071)	(0.175)	(0.222)	(0.086)	(0.102)
White	0.206***	0.386***	0.999***	0.412***	0.435***
	(0.072)	(0.136)	(0.177)	(0.083)	(0.093)
Black	0.166*	0.262	0.397	0.207*	0.215*
	(0.093)	(0.194)	(0.280)	(0.109)	(0.117)

Note: This table shows estimated effects of the ATE under different assumptions of what the effect of waiting is on students who are reluctant compliers at November 1 but eager compliers at December 1. It shows that average treatment effects are stable across a wide range of assumptions—in large part because these students make up such a small portion of the population.

Figure A.6: Visualization of Policy Change



(a) 2002-2012 Cohorts: Only December Cutoff



(b) 2013 Cohort: November and December Cutoffs

Note: This figure presents the first stage of the regression discontinuity by depicting the probability that students with different birthdays wait to enter kindergarten and how this changes at the birthday cutoffs move. The graph shows both a scatter plot of the probability of waiting to enter kindergarten by birthday and the associated lines of best fit (with uniform weights). Note that the limits of these lines identify the unconditional probability of waiting on either side of the cutoff $p_{z(r)}$ and $p_{z'}(r)$. The sample is comprised of first-time kindergarteners who turned five between October 1 and December 31 2013 and for whom I observe third-grade test scores.

APPENDIX B

Appendix to Chapter 2

B.1. RD Robustness

Appendix Figure B.1 shows the estimated treatment effect for different bandwidths and for different samples (extending the entering cohorts included). The left figure shows the treatment effect when the bandwidth around the cutoff is varied but the RD sample remained constant, using only firms with at least 60 months of data following the PTC deadline. The right figure shows how the estimates vary when the RD sample varies from including firms with at least 36 months post-PTC deadline to only including firms with 84 months. The effect remains negative and stable across bandwidths and samples, though the standard errors increase as the number of months required to be in the sample increases and therefore the number of firms in the RD sample decreases.

Regarding RD specification, Panel A in Appendix Table B.1 presents alternative estimates from a using a triangular or Epanechnikov kernel and Panel B displays estimates from alternative polynomial orders using a uniform kernel. For all specifications, the estimated treatment remains roughly similar with estimates ranging from 1.2 to over 2 percentage points.

B.2. Proofs for the Optimal Tax Model

In order to prove the main results in proposition 1, we first prove a helpful lemma.

Lemma 1. Under assumption 1, the marginal increase in the firm's variable input (v_2^f) with respect to a marginal change in a policy parameter is equal to the marginal increase in the capital input (x^f) scaled by the ratio of the second derivatives of the production function.

Table B.1: Treatment Effect Sensitivity to Alternative RD Setups

Panel A: Kernel Specifications

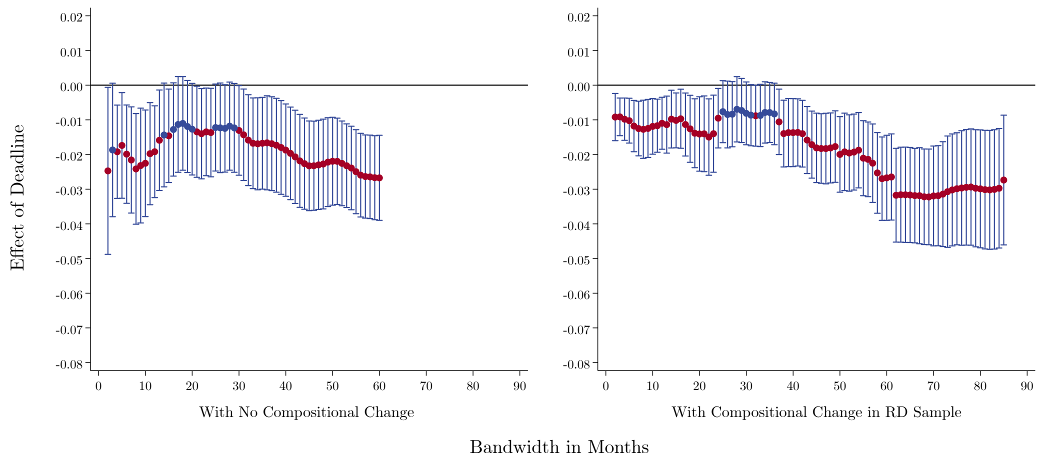
	Uniform	Triangle	Epanechnikov
Effect of Deadline	-0.027**	-0.021**	-0.023**
	(0.006)	(0.006)	(0.006)

Panel B: Polynomial Order

	Zero Order	Linear	Quadratic
Effect of Deadline	-0.015**	-0.027**	-0.011
	(.005)	(.006)	(.008)

Note: This table reports alternative specifications of RD estimation. Panel A shows the results with different kernel specifications. Panel B shows the results with different polynomial orders. All regressions use capacity factor as the dependent variable, include observations within 60 months of the deadline in the RD sample, and include non-RD observations to estimate the state-by-month-year fixed effects.
$**p < 0.01$.

Figure B.1: Effect Sizes are Consistent Across Bandwidths and Sample Compositions



Note: This figure shows how estimates of the production response at the production tax credit deadline change with the RD bandwidth. Panel A on the left shows how the effect changes when the RD bandwidth changes holding constant the sample of firms that produce at least 60 months after the deadline. It shows that among this sample of facilitates the effects are constant. Panel B on the right shows how the effect changes when both the sample and the RD bandwidth change—here longer bandwidths imply fewer identifying observations. The slight downward slope is the identifying variation of the vintage heterogeneity mentioned in Section 1.4. Both regressions use the same set of control firms for the fixed effects (removing observations that would be treated in the case of Panel B).

$$\frac{\partial v_2^f}{\partial \tau_i} = -\frac{q_{xv}(x^f, v_2^f)}{q_{vv}(x^f, v_2^f)} \frac{\partial x}{\partial \tau_i}$$
$$\frac{\partial v_2^f}{\partial \tau_o} = -\frac{q_{xv}(x^f, v_2^f)}{q_{vv}(x^f, v_2^f)} \frac{\partial x}{\partial \tau_o}$$
$$\frac{\partial v_2^f}{\partial T} = -\frac{q_{xv}(x^f, v_2^f)}{q_{vv}(x^f, v_2^f)} \frac{\partial x}{\partial T}$$

Proof. Under assumption 1, if there exists an interior solution to the firms problem for a given choice of policy parameters (τ_i, τ_o, T), the firm's solution can be defined as the choice of x^f,v_1^f,v_2^f such that $F = \begin{pmatrix} f_1 \\ f_2 \\ f_3 \end{pmatrix} = 0$.

$$
\begin{aligned}
f_1(x^f, v_1^f, v_2^f; \tau_i, \tau_o, T) =& q_v(x^f, v_1^f)(1+\tau_o) - m \\
f_2(x^f, v_1^f, v_2^f; \tau_i, \tau_o, T) =& q_v(x^f, v_2^f) - m \\
f_3(x^f, v_1^f, v_2^f; \tau_i, \tau_o, T) =& T q_x(x^f, v_1^f) + (1-T) q_x(x^f, v_2^f) + T\tau_o q_x(x_1^f, v_1^f) - c(1-\tau_i)
\end{aligned}
$$

Using the implicit function theorem we can define a function g such that

$$
\begin{pmatrix} x^f \\ v_1^f \\ v_2^f \end{pmatrix} = g \begin{pmatrix} \tau_i \\ \tau_o \\ T \end{pmatrix}
$$

and furthermore

$$
\frac{\partial g}{\partial \tau} = -[J_{f,y}(\tau, g(\tau))]^{-1} \frac{\partial f}{\partial \tau}
$$

Evaluating the expressions for $\frac{\partial x^f}{\partial \tau}$ and $\frac{\partial v_2^f}{\partial \tau}$ proves the lemma.

Proof. Proof of Proposition 1 The optimal investment and output subsidies are derived by choosing τ_o and τ_i to maximize welfare (equation 2.3.2). Differentiating welfare with respect to τ_i and τ_o lead to the following first order condition:

$$
\begin{aligned}
\frac{\partial W}{\partial \tau_i} = -cx^f + \gamma \Big[T \big[q_x(x^f, v_1^f) \frac{\partial x^f}{\partial \tau_i} + q_v(x^f, v_1^f) \frac{\partial v_1^f}{\partial \tau_i} \big] + (1-T) \big[q_x(x^f, v_2^f) \frac{\partial x^f}{\partial \tau_i} + q_v(x^f, v_2^f) \frac{\partial v_2^f}{\partial \tau_i} \big] \Big] + \\
\lambda \Big[-c \frac{\partial x^f}{\partial \tau_i} \tau_i + cx^f - T\tau_o \big[q_x(x^f, v_1^f) \frac{\partial q}{\partial x^f} + q_v(x^f, v_1^f) \frac{\partial v_1^f}{\partial \tau_i} \big] \Big] = 0
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial W}{\partial \tau_o} = & -Tq(x^f, v_1^f) + \gamma \Big[T \Big(q_x(x^f, v_1^f) \frac{\partial x^f}{\partial \tau_o} + q_v(x^f, v_1^f) \frac{\partial v_1^f}{\partial \tau_o} \Big) + (1-T) \Big(q_x(x^f, v_2^f) \frac{\partial x^f}{\partial \tau_o} + q_v(x^f, v_2^f) \frac{\partial v_2^f}{\partial \tau_o} \Big) \Big] \\
& + \lambda \Big[-c \frac{\partial x^f}{\partial \tau_o} \tau_i + T \Big[q(x^f, v_1^f) - \tau_o \big(q_x(x^f, v_1^f) \frac{\partial x^f}{\partial \tau_o} + q_v(x^f, v_1^f) \frac{\partial v_1^f}{\partial \tau_o} \big) \Big] \Big] = 0
\end{aligned}
$$

To further simplify these expressions, we use the following Taylor approximations:

$$
\begin{aligned}
q_v(x^f, v_1^f) &\approx q_v(x^f, v_2^f) + \Delta v \, q_{vv}(x^f, v_2^f) \\
q_x(x^f, v_1^f) &\approx q_x(x^f, v_2^f) + \Delta v \, q_{xv}(x^f, v_2^f)
\end{aligned}
$$

Using the firm's first order condition, we find that

$$\Delta v = \frac{m\tau_o}{q_{vv}(x^f, v_2^f)(1+\tau_o)}$$

Using these approximations, the result from lemma 1, and assuming there are no efficiency costs of raising tax revenue ($\lambda = 1$) the first order conditions simplify to the following pair of equations

$$\begin{aligned}\tau_i =& \frac{(\gamma - T\tau_o)\big(q_x - q_v\frac{q_{xv}}{q_{vv}}\big)}{c} \\ 0 =& T(\gamma - \tau_o)\Big[q_v\frac{m}{q_{vv}(1+\tau_o)^2} + \Delta v\big(\frac{m}{(1+\tau_o)^2}\big)\Big]\end{aligned}$$

The term inside $[\cdot]$ is equal to $q_v(x^f, v_1^f) * \frac{m}{q_{vv}(1+\tau_o)^2}$. By assumption $q_v > 0$ and $q_{vv} < 0$ so as long as $m \neq 0$ then this term is nonzero. Thus the only way for the second condition to hold is for $\tau_o = \gamma$.

Setting $\tau_o = \gamma$ in the first expression, it simplifies to

$$\tau_i = \frac{\gamma(1-T)\big(q_x - q_v\frac{q_{xv}}{q_{vv}}\big)}{c}$$

Proof. Proof of Proposition 2 The optimal deadline length, T, is found by differentiating equation 2.3.2 with respect to T. Use the Taylor approximations, lemma 1, and $\lambda = 1$ to simplify, the first order condition becomes

$$\frac{\partial W}{\partial T} = -\Delta q(1+\gamma) - m\Delta v - \phi'(T)$$

Here $\Delta q = q(x^f, v_1^f) - q(x^f, v_2^f)$ is used to denote the change in output at the end of the output subsidy. Using a Taylor approximation for the production function $\Delta q = \Delta v q_v(x^f, v_2^f) = m\Delta v$ then the first order condition simplifies further and the optimal T is defined by

$$\phi'(T) = -\gamma\Delta q$$

B.3. Continuous Time Model

In this appendix, we prove that the two stage model is isomrophic to a continuous time model with discounting. In a continuous time framework, the firm discounts profits at a rate β and chooses capital x and variable inputs v_t for $t \in [0,1]$ to maximize the continuous time firm's problem:

$$\max_{x,v_t} \int_0^T \exp\{-\beta t\}\Big[q(x,v_t)(1-\tau_o) - mv_t\Big]dt + \int_T^1 \exp\{-\beta t\}\Big[q(x,v_t) - mv_t\Big]dt - cx(1+\tau_i)$$

Lemma 2. The Firm's optimal variable inputs v_t will be a constant piece-wise function for $t < T$ and $t \geq T$ defined by

$$\begin{cases} q_v(x,v_1^f) - \frac{m}{1-\tau_o} = 0 & t < T \\ q_v(x,v_2^f) - m = 0 & t \geq T \end{cases}$$

Proof. For any t, the first order condition with respect to v_t:

$$\exp\{-\beta t\}\Big[q_v(x,v_t)(1-\tau_o) - m\Big] \cdot \mathbb{1}(t < T) + \exp\{-\beta t\}\Big[q_v(x,v_t) - m\Big] \cdot \mathbb{1}(t \geq T) = 0$$

The above equation holds with equality for v_t defined by

$$\begin{cases} q_v(x,v_1^f) - \frac{m}{1-\tau_o} = 0 & t < T \\ q_v(x,v_2^f) - m = 0 & t \geq T \end{cases} \tag{B.1}$$

Lemma 3. The Firm's optimal investment decision is defined by

$$\tilde{T} q_x(x^f, v_1^f)(1-\tau_o) + (1-\tilde{T})q_x(x^f, v_2^f) - \tilde{c}(1+\tau_i) = 0 \tag{B.2}$$

Where $\tilde{T}$ and $\tilde{c}$ are defined as

$$\tilde{T} = \frac{1-\exp\{-\beta T\}}{1-\exp\{-\beta\}}$$
$$\tilde{c} = \frac{\beta}{1-\exp\{-\beta\}}c$$

Proof. From lemma 2 we can write the firm profits as

$$\max_x \left\{ \Big(\frac{1-e^{-\beta T}}{\beta}\Big)\Big[q(x,v_1^f)(1-\tau_o)-mv_1^f\Big] \right.$$
$$\left. +\Big(\frac{e^{-\beta}(e^{\beta(1-T)}-1)}{\beta}\Big)\Big[q(x,v_2^f)-mv_2^f\Big]-cx(1+\tau_i)\right\}$$

The first order condition is therefore

$$\Big(\frac{1-e^{-\beta T}}{\beta}\Big)\Big[q_x(x,v_1^f)(1-\tau_o)\Big]+\Big(\frac{e^{-\beta}(e^{\beta(1-T)}-1)}{\beta}\Big)\Big[q_x(x,v_2^f)\Big]-c(1+\tau_i)=0$$

Multiplying through by $\frac{\beta}{1-e^{-\beta}}$ the first order conditions becomes

$$\tilde{T}q_x(x^f,v_1^f)(1-\tau_o)+(1-\tilde{T})q_x(x^f,v_2^f)-\tilde{c}(1+\tau_i)=0$$

Proposition 3. The two-stage model is isomorphic to the continuous time model with exponential discounting.

Proof. From lemmas 2 and 3 the firm's problem in continuous time defined by equations B.1 and B.2 and is identical to the two-stage firm's problem with T and c replaced with $\tilde{T}$ and $\tilde{c}$. The optimal τ_o and τ_i are therefore

$$\tau_i^* = -\frac{\gamma(1-\tilde{T})\frac{\mathrm{d}q}{\mathrm{d}\tau_i}}{\tilde{c}\frac{\partial x^f}{\partial \tau_i}} \tag{B.3}$$

$$\tau_o^* = -\gamma \tag{B.4}$$

The optimal subsidy deadline T is set to maximize

$$\begin{aligned}W =& q(x^f,v_2^f)(1-\tau_o^*)\frac{1-\exp\{-\beta T\}}{\beta}-mv_1^f\frac{1-\exp\{-\beta T\}}{\beta}+\big(\exp\{-\beta\}\frac{\exp\big\{\beta(1-T)\big\}-1}{\beta}\big)\\ &+\gamma\big[q(x^f,v_1^f)\frac{1-\exp\{-\beta T\}}{\beta}+q(x^f,v_2^f)\exp\{-\beta\}\frac{\exp\big\{\beta(1-T)\big\}-1}{\beta}\big]\\ &+\lambda c x^f\tau_i^*+\lambda\tau_o^* q(x^f,v_1^f)\frac{1-\exp\{\beta T\}}{\beta}\end{aligned}$$

Multiplying through by $\frac{\beta}{1-\exp\{-\beta\}}$ yields a monotonically transformed welfare function $\tilde{W}$ which gives rise to an isomorphic expression for the optimal $\tilde{T}$.

$$\phi'(\tilde{T}^*) = \gamma \Delta q$$

B.4. Cobb Douglass Solution

Let $q_t = x^a v_t^b$. Then plugging the firm's first order conditions into the equation for welfare yields

$$\begin{aligned}
W = &\left(\frac{b}{m}\right)^{\frac{b}{1-a-b}} \left[\frac{a}{c(1+\tau_i)} \left(T(1+\tau_o)^{\frac{1}{1-b}} + (1-T)\right)\right]^{\frac{a}{1-a-b}} \\
&\cdot \Bigg[(1+\gamma)\left(T(1+\tau_o)^{\frac{b}{1-b}} + (1-T)\right) + (1-\lambda)\tau_o T(1+\tau_o)^{\frac{b}{1-b}} \\
&- \left(b + \frac{a(1+(1-\lambda)\tau_i)}{1+\tau_i}\right)\left(T(1+\tau_o)^{\frac{1}{1-b}} + (1-T)\right)\Bigg] - \phi(T)
\end{aligned}$$

with the FOC

$$\begin{aligned}
\frac{\partial W}{\partial \tau_o} :& a(1+\gamma)\frac{T(1+\tau_o)^{\frac{b}{1-b}} + (1-T)}{T(1+\tau_o)^{\frac{1}{1-b}} + (1-T)} + a\frac{(1-\lambda)\tau_o T(1+\tau_o)^{\frac{b}{1-b}}}{T(1+\tau_o)^{\frac{1}{1-b}} + (1-T)} \\
&= (1-b)\left(b + \frac{a(1+(1-\lambda)\tau_i)}{1+\tau_i}\right) - (1-a-b)\left[\frac{\left(1+\gamma+(1-\lambda)\tau_o\right)b}{1+\tau_o} + (1-\lambda)(1-b)\right]
\end{aligned}$$

$$\begin{aligned}
\frac{\partial W}{\partial \tau_i} :& (1+\gamma)\,\frac{T(1+\tau_o)^{\frac{b}{1-b}} + (1-T)}{T(1+\tau_o)^{\frac{1}{1-b}} + (1-T)} - \frac{(1-\lambda)\tau_o T(1+\tau_o)^{\frac{b}{1-b}}}{T(1+\tau_o)^{\frac{1}{1-b}} + (1-T)} \\
&= b + \frac{(1-a-b)\lambda + a(1+(1-\lambda)\tau_i)}{(1+\tau_i)}
\end{aligned}$$

$$\begin{aligned}
\frac{\partial W}{\partial T} :& \left(\frac{b}{m}\right)^{\frac{b}{1-a-b}} \left[\frac{a}{c(1+\tau_i)}\left(T(1+\tau_o)^{\frac{1}{1-b}} + (1-T)\right)\right]^{\frac{a}{1-a-b}} \left((1+\tau_o)^{\frac{1}{1-b}} - 1\right) \\
&\left(\frac{a\,(1+\gamma)}{1-a-b}\frac{T(1+\tau_o)^{\frac{b}{1-b}} + (1-T)}{T(1+\tau_o)^{\frac{1}{1-b}} + (1-T)} - \frac{1-b}{1-a-b}\left(b + \frac{a}{1+\tau_i}\right) + (1+\gamma)\,\frac{(1+\tau_o)^{\frac{b}{1-b}} - 1}{(1+\tau_o)^{\frac{1}{1-b}} - 1}\right) = \phi'(T)
\end{aligned}$$

B.4.1 Optimal Subsidies if $\lambda = 1$

If $\lambda = 1$, combining the FOC for τ_i and τ_o gives us

$$1+\tau_o = 1+\gamma$$
$$1+\tau_i = \frac{1-b}{(1+\gamma)\,\frac{T(1+\gamma)^{\frac{b}{1-b}}+(1-T)}{T(1+\gamma)^{\frac{1}{1-b}}+(1-T)} - b}$$

And combining FOC for either τ_i or τ_o and T and letting $\tau_o = \gamma$, reveal

$$(1+\gamma)\,\frac{(1+\gamma)^{\frac{b}{1-b}}-1}{(1+\gamma)^{\frac{1}{1-b}}-1} - b = \phi'(T)$$

Note that both this quantity and ΔQ are functions of the ratio $\frac{(1+\gamma)^{\frac{b}{1-b}}-1}{(1+\gamma)^{\frac{1}{1-b}}-1}$, and Δv is a function of b. In the Taylor expansion the $\Delta v q_v$ cancels with the $1\cdot\Delta Q$, but because this solution is unapproximated, second order terms lead to slight differences.

B.4.2 Optimal Subsidies if $\lambda > 1$

By combining the FOC for τ_o and τ_i, they simplify and can be solved symbolically by Wolfram. As long aslong as $b(1-\lambda) \neq b-(1-\lambda)(1-a-b)$, $-\lambda \neq \gamma$, $z \neq a(1-\lambda)$, and $b \neq 1$, the optimal subsidies are

$$\tau_o = \frac{-\gamma-(1-\lambda)\frac{1-a-b}{b}}{-\lambda+(1-\lambda)\frac{(1-a-b)}{b}}$$
$$\tau_i = \frac{z-(1-a-b)\lambda-a}{a(1-\lambda)-z}$$

where

$$z = (1+\gamma)\,\frac{T\left(\frac{-\lambda-\gamma}{-\lambda+(1-\lambda)\frac{(1-a-b)}{b}}\right)^{\frac{b}{1-b}}+(1-T)}{T\left(\frac{-\lambda-\gamma}{-\lambda+(1-\lambda)\frac{(1-a-b)}{b}}\right)^{\frac{1}{1-b}}+(1-T)} - \frac{(1-\lambda)\frac{-\gamma-(1-\lambda)\frac{1-a-b}{b}}{-\lambda+(1-\lambda)\frac{(1-a-b)}{b}}T\left(\frac{-\lambda-\gamma}{-\lambda+(1-\lambda)\frac{(1-a-b)}{b}}\right)^{\frac{b}{1-b}}}{T\left(\frac{-\lambda-\gamma}{-\lambda+(1-\lambda)\frac{(1-a-b)}{b}}\right)^{\frac{1}{1-b}}+(1-T)} - b$$

As we couldn't find an interpretable closed form solution for T, so we solve for T^* computationally in the calibrations.

APPENDIX C

Appendix to Chapter 3

C.1. Balance

Table C.1 shows the average belief errors and characteristics of participants assigned to each treatment arm: control, graduation rate, and rate and rank. It also reports the p-values of t tests of equality between groups.

C.2. Preferred Spending During and After the Pandemic

Our main results focus on taxpayers' preferences for spending on public four-year colleges after the pandemic has passed. However, since the pandemic has created extraordinary circumstances for government tax collection and society as a whole, preferences for spending (on higher education) during the pandemic may be different.

Figure C.1 shows the relationship between taxpayers' preferred spending during and after the pandemic, overall and by group. Panel (a) plots the two measures and shows that they are very highly correlated. The estimated dashed line, which is the estimated linear fit, is close to the dotted 45 degree line. The average taxpayer expresses similar preferences for spending during and after the pandemic. However, we find some suggestive evidence that taxpayers with high preferences for spending during the pandemic prefer lower spending when it has passed, while the reverse is true for households with low preferences for spending during the pandemic.

Although the overall difference between preferred policies is small, we now investigate how the difference varies across types of taxpayers. We estimate a regression of a regression of preferred spending after the pandemic has passed minus preferred spending during the

Table C.1: Balance on Beliefs and Taxpayer Characteristics

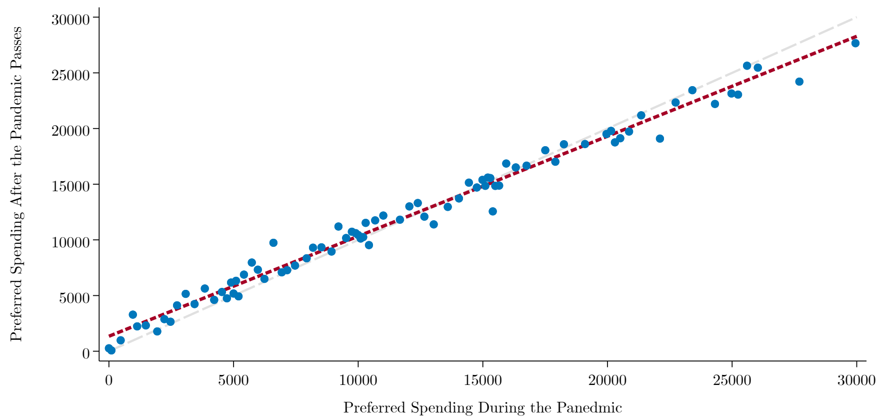
Panel A: Belief Errors	Mean	Control	Rate	Rate and Rank	Control - Rate	Control - Both
US Graduation Rate	1.081	0.798	2.168	0.252	-1.369^{+}	0.547
					[p=0.055]	[p=0.458]
State Graduation Rate	0.518	0.206	1.498	-0.174	-1.292^{+}	0.380
					[p=0.100]	[p=0.633]
State Graduation Rank	7.685	7.322	7.551	8.184	-0.230	-0.863
					[p=0.724]	[p=0.191]
State Spending per Student	3349	3174	3256	3618	-82	-443
					[p=0.776]	[p=0.122]
State Spending Rank	3.423	2.970	3.481	3.814	-0.511	-0.844
					[p=0.491]	[p=0.269]

Panel B: Respondent Characteristics	Sample Mean	Control Group	Level	Level and Rank	Control - Level	Control - Both
Female	0.513	0.513	0.512	0.513	0.002	0.001
					[p=0.923]	[p=0.965]
Republican	0.406	0.414	0.396	0.408	0.018	0.006
					[p=0.352]	[p=0.760]
Never Attended College	0.409	0.400	0.412	0.414	-0.012	-0.014
					[p=0.535]	[p=0.471]
Four-Year College Degree	0.305	0.307	0.299	0.310	0.008	-0.003
					[p=0.653]	[p=0.891]
Age	67.0	67.0	67.0	67.0	0.014	0.054
					[p=0.975]	[p=0.907]
Attended in State	0.262	0.269	0.262	0.255	0.006	0.013
					[p=0.716]	[p=0.447]
Child Attended in State	0.273	0.272	0.277	0.269	-0.005	0.003
					[p=0.765]	[p=0.869]
Likely that a Child Will Attend	0.689	0.676	0.703	0.687	-0.027	-0.011
					[p=0.146]	[p=0.547]
Family Attachment	0.793	0.796	0.798	0.785	-0.002	0.011
					[p=0.885]	[p=0.496]
Follows College Sports	0.425	0.440	0.423	0.411	0.017	0.029
					[p=0.403]	[p=0.147]
Observations	3,715	1,225	1,259	1,231	2,584	2,556

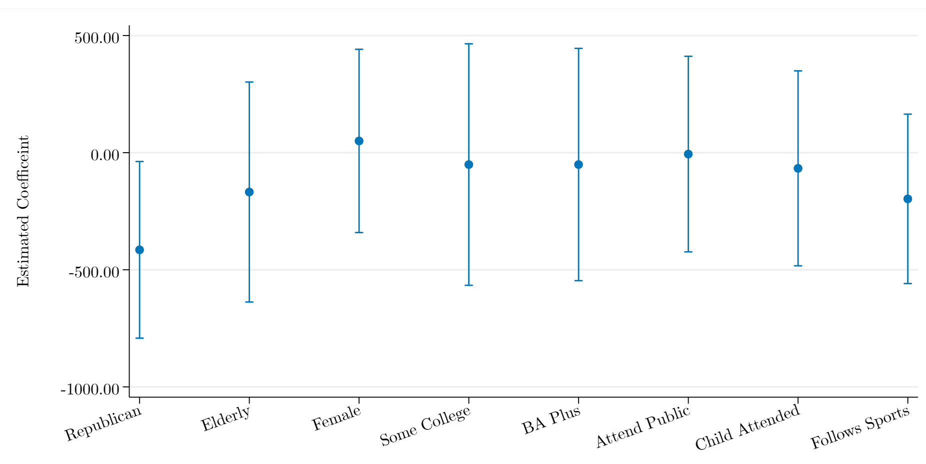
Note: Each row presents the overall sample mean, mean by treatment group, and the difference between the two treatments and control. The standard deviation of continuous variables are also listed in parentheses in the overall mean column.

pandemic on a set of demographic indicators. Panel (b) plots the coefficients. We find important differences by political affiliation. While Republicans' pandemic and post-pandemic preferred policies barely differ, Democrats and Independents prefer less spending during the pandemic.

Figure C.1: Preferred Spending During and After the Pandemic



(a) Ideal vs Covid Preferred Spending



(b) Differences in Ideal and Covid Preferred Spending By Groups

Note: Panel (a) presents the relationship between preferred spending during and after the pandemic by percentile of preferred spending during the pandemic. The dashed gra line shows the 45 degree line for reference. Panel (b) estimates a regression of preferred spending after the pandemic has passed minus preferred spending during the pandemic on a set of demographic indicators. The omitted education group is high school graduate or equivalent, or no high school degree. Both panels only use taxpayers in the control group.

C.3. Rank Updating Details

We explore the what extent to which changes may represent learning across public good parameters. That is, to what extent to taxpayers learn about the rank when they receive information about the rate? We measure this effect at the end of our Cloud Research experiment with about 665 participants who do not receive the rank information as part of the experiment itself. For those in the control or graduation rate only conditions, we give (again) the true rate and ask for their updated beliefs about the rank.

Unconditionally, individuals in our sample living in a state with a 1 percentage point higher graduation rate are in a state that rank 3.6 percentiles better. Their priors of this relationship are attenuated: on average living in a state with a 1 percentage point higher graduation rate only predicts a prior rank 0.9 percentiles better (or 0.80 places conditioning on their prior of the graduation rate). When exposed to information about the rate, individuals update their beliefs, but only 7 percent of the gap in between the prior and the truth (an additional 0.20 percentiles per percentage point increase in the graduation rate). These averages obfuscate important heterogeneity in who updates as described in Figure 3.5.

BIBLIOGRAPHY

BIBLIOGRAPHY

Abadie, A. (2002): "Bootstrap tests for distributional treatment effects in instrumental variable models," Journal of the American Statistical Association, 97, 284–292.

Abrell, J., S. Rausch, and C. Streitberger (2019): "The economics of renewable energy support," Journal of Public Economics, 176, 94–117.

Achen, C. H. and L. M. Bartels (2017): "Democracy for realists," in Democracy for Realists, Princeton University Press.

Adhvaryu, A., S. Bednar, T. Molina, Q. Nguyen, and A. Nyshadham (2020): "When it rains it pours: The long-run economic impacts of salt iodization in the United States," Review of Economics and Statistics, 102, 395–407.

Aldy, J. E., T. D. Gerarden, and R. L. Sweeney (2021): "Investment versus output subsidies: Implications of alternative incentives for wind energy," Working paper.

Alesina, A., M. F. Ferroni, and S. Stantcheva (2021): "Perceptions of racial gaps, their causes, and ways to reduce them," NBER Working Paper w29245.

Alesina, A., A. Miano, and S. Stantcheva (2018): "Immigration and redistribution," NBER Working Paper w24733.

——— (2020): "The polarization of reality," in AEA Papers and Proceedings, vol. 110, 324–28.

Althaus, S. L. (1998): "Information effects in collective preferences," American Political Science Review, 92, 545–558.

Ambec, S. and C. Crampes (2019): "Decarbonizing electricity generation with intermittent sources of energy," Journal of the Association of Environmental and Resource Economists, 6, 1105–1134.

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

Angrist, J. D. and M. Rokkanen (2015): "Wanna get away? Regression discontinuity estimation of exam school effects away from the cutoff," Journal of the American Statistical Association, 110, 1331–1344.

Attar, I. and D. Cohen-Zada (2018): "The effect of school entrance age on educational outcomes: Evidence using multiple cutoff dates and exact date of birth," Journal of Economic Behavior & Organization, 153, 38–57.

Barber, M. and J. C. Pope (2019): "Does party trump ideology? Disentangling party and ideology in America," American Political Science Review, 113, 38–54.

Barua, R. and K. Lang (2016): "School entry, educational attainment, and quarter of birth: A cautionary tale of a local average treatment effect," Journal of Human Capital, 10, 347–376.

Bassok, D. and S. F. Reardon (2013): "'Academic redshirting' in kindergarten: Prevalence, patterns, and implications," Educational Evaluation and Policy Analysis, 35, 283–297.

Bedard, K. and E. Dhuey (2006): "The persistence of early childhood maturity: International evidence of long-run age effects," The Quarterly Journal of Economics, 121, 1437–1472.

——— (2012): "School-entry policies and skill accumulation across directly and indirectly affected individuals," Journal of Human Resources, 47, 643–683.

Bertanha, M. and G. W. Imbens (2019): "External validity in fuzzy regression discontinuity designs," Journal of Business & Economic Statistics, 1–39.

Besley, T. and A. Case (1995): "Incumbent behavior: Vote-seeking, tax-setting, and yardstick competition," American Economic Review, 85, 25.

Black, D., J. Joo, R. LaLonde, J. A. Smith, and E. Taylor (2022): "Simple tests for selection: Learning more from instrumental variables," NBER Working Paper.

Black, S. E., P. J. Devereux, and K. G. Salvanes (2011): "Too young to leave the nest? The effects of school starting age," The Review of Economics and Statistics, 93, 455–467.

Bound, J., B. Braga, G. Khanna, and S. Turner (2019): "Public universities: The supply side of building a skilled workforce," RSF: The Russell Sage Foundation Journal of the Social Sciences, 5, 43–66.

Bound, J., M. F. Lovenheim, and S. Turner (2010): "Why have college completion rates declined? An analysis of changing student preparation and collegiate resources," American Economic Journal: Applied Economics, 2, 129–57.

Bound, J. and A. Simon (2021): "College choice, private options, and the incidence of public investment in higher education," NBER Working Paper.

Bound, J. and S. Turner (2007): "Cohort crowding: How resources affect collegiate attainment," Journal of Public Economics, 91, 877–899.

Bovenberg, A. L. and F. van der Ploeg (1994): "Environmental policy, public finance and the labour market in a second-best world," Journal of Public Economics, 55, 349–390.

Brand, J. E. and Y. Xie (2010): "Who benefits most from college? Evidence for negative selection in heterogeneous economic returns to higher education," American Sociological Review, 75, 273–302.

Brinch, C. N., M. Mogstad, and M. Wiswall (2017): "Beyond LATE with a discrete instrument," Journal of Political Economy, 125, 985–1039.

Burr, C. (2016): "Subsidies and Investments in the Solar Power Market," Working Paper.

Cai, Y. and T. S. Lontzek (2019): "The Social Cost of Carbon with Economic and Climate Risks," Journal of Political Economy, 127, 2684–2734.

Callaway, D. S., M. Fowlie, and G. McCormick (2018): "Location, location, location: The variable value of renewable energy and demand-side efficiency resources," Journal of the Association of Environmental and Resource Economists, 5, 39–75.

Carneiro, P., J. J. Heckman, and E. J. Vytlacil (2011): "Estimating marginal returns to education," American Economic Review, 101, 2754–81.

Cascio, E. U. and D. W. Schanzenbach (2016): "First in the class? Age and the education production function," Education Finance and Policy, 11, 225–250.

Cattaneo, M. D., L. Keele, R. Titiunik, and G. Vazquez-Bare (2020): "Extrapolating treatment effects in multi-cutoff regression discontinuity designs," Journal of the American Statistical Association, 1–12.

Cattaneo, M. D., R. Titiunik, G. Vazquez-Bare, and L. Keele (2016): "Interpreting regression discontinuity designs with multiple cutoffs," The Journal of Politics, 78, 1229–1248.

Cerulli, G., Y. Dong, A. Lewbel, and A. Poulsen (2017): "Testing stability of regression discontinuity models," in Regression Discontinuity Designs, Emerald Publishing Limited.

Chatterji, A. K., J. Kim, and R. C. McDevitt (2018): "School spirit: Legislator school ties and state funding for higher-education," Journal of Public Economics, 164, 254–269.

Conzelmann, J. G., S. W. Hemelt, B. Hershbein, S. M. Martin, A. Simon, and K. M. Stange (2022): "Grads on the go: Measuring college-specific labor markets for graduates," NBER Working Paper.

Cook, J. A. and C.-Y. C. L. Lawell (2020): "Wind turbine shutdowns and upgrades in Denmark: Timing decisions and the impact of government policy," The Energy Journal, 41.

Cook, P. J. and S. Kang (2016): "Birthdays, schooling, and crime: Regression-discontinuity analysis of school performance, delinquency, dropout, and crime initiation," American Economic Journal: Applied Economics, 8, 33–57.

——— (2018): "The school-entry-age rule affects redshirting patterns and resulting disparities in achievement," NBER Working Paper.

Cornelissen, T., C. Dustmann, A. Raute, and U. Schönberg (2016): "From LATE to MTE: Alternative methods for the evaluation of policy interventions," Labour Economics, 41, 47–60.

——— (2018): "Who benefits from universal child care? Estimating marginal returns to early child care attendance," Journal of Political Economy, 126, 2356–2409.

Cremer, H., F. Gahvari, and N. Ladoux (1998): "Externalities and optimal taxation," Journal of Public Economics, 70, 343–364.

Cruces, G., R. Perez-Truglia, and M. Tetaz (2013): "Biased perceptions of income distribution and preferences for redistribution: Evidence from a survey experiment," Journal of Public Economics, 98, 100–112.

Cullen, J. (2013): "Measuring the environmental benefits of wind-generated electricity," American Economic Journal: Economic Policy, 5, 107–33.

De Groote, O. and F. Verboven (2019): "Subsidies and time discounting in new technology adoption: Evidence from solar photovoltaic systems," American Economic Review, 109, 2137–72.

Dee, T. S. and H. H. Sievertsen (2018): "The gift of time? School starting age and mental health," Health Economics, 27, 781–802.

Deming, D. and S. Dynarski (2008): "The lengthening of childhood," Journal of Economic Perspectives, 22, 71–92.

Deming, D. J. and C. R. Walters (2017): "The impact of price caps and spending cuts on US postsecondary attainment," Nber working paper.

Dharmapala, D., J. Slemrod, and J. D. Wilson (2011): "Tax policy and the missing middle: Optimal tax remittance with firm-level administrative costs," Journal of Public Economics, 95, 1036–1047.

Dhuey, E., D. Figlio, K. Karbownik, and J. Roth (2019): "School starting age and cognitive development," Journal of Policy Analysis and Management, 38, 538–578.

Diamond, P. A. and J. A. Mirrlees (1971): "Optimal taxation and public production I: Production efficiency," American Economic Review, 61, 8–27.

Dixit, A. (1985): "Tax policy in open economies," in Handbook of Public Economics, Elsevier, vol. 1, 313–374.

Dong, Y. and A. Lewbel (2015): "Identifying the effect of changing the policy threshold in regression discontinuity models," Review of Economics and Statistics, 97, 1081–1092.

Doyle, W. R. (2007): "Public opinion, partisan identification, and higher education policy," The Journal of Higher Education, 78, 369–401.

——— (2010): "US senator's ideal points for higher education: Documenting partisanship, 1965–2004," The Journal of Higher Education, 81, 619–644.

Dunne, T., S. D. Klimek, M. J. Roberts, and D. Y. Xu (2013): "Entry, exit, and the determinants of market structure," The RAND Journal of Economics, 44, 462–487.

Ehrenberg, R. G. (2008): What's happening to public higher education?: The shifting financial burden, JHU Press.

Einav, L., A. Finkelstein, and P. Schrimpf (2010): "Optimal mandates and the welfare cost of asymmetric information: Evidence from the UK annuity market," Econometrica, 78, 1031–1092.

Elder, T. E. (2010): "The importance of relative standards in ADHD diagnoses: evidence based on exact birth dates," Journal of Health Economics, 29, 641–656.

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

Epley, N. and T. Gilovich (2016): "The mechanics of motivated reasoning," Journal of Economic perspectives, 30, 133–40.

Evans, W. N., M. S. Morrill, and S. T. Parente (2010): "Measuring inappropriate medical diagnosis and treatment in survey data: The case of ADHD among school-age children," Journal of Health Economics, 29, 657–673.

Felfe, C. and R. Lalive (2018): "Does early child care affect children's development?" Journal of Public Economics, 159, 33–53.

Fell, H., D. T. Kaffine, and K. Novan (2021): "Emissions, transmission, and the environmental value of renewable energy," American Economic Journal: Economic Policy, 13, 241–72.

Ferrario, B. and S. Stantcheva (2022): "Eliciting people's first-order concerns: Text analysis of open-ended survey questions," NBER Working Paper.

Finkelstein, A. and M. J. Notowidigdo (2019): "Take-up and targeting: Experimental evidence from SNAP," The Quarterly Journal of Economics, 134, 1505–1556.

Fortner, C. K. and J. M. Jenkins (2017): "Kindergarten redshirting: Motivations and spillovers using census-level data," Early Childhood Research Quarterly, 38, 44–56.

FRYER, R. G. (2017): "The production of human capital in developed countries: Evidence from 196 randomized field experiments," in Handbook of Economic Field Experiments, Elsevier, vol. 2, 95–322.

GAINES, B. J., J. H. KUKLINSKI, P. J. QUIRK, B. PEYTON, AND J. VERKUILEN (2007): "Same facts, different interpretations: Partisan motivation and opinion on Iraq," The Journal of Politics, 69, 957–974.

GIACCOBASSO, M., B. C. NATHAN, R. PEREZ-TRUGLIA, AND A. ZENTNER (2022): "Where do my tax dollars go? Tax morale effects of perceived government spending," NBER Working Paper.

GILENS, M. (2001): "Political ignorance and collective policy preferences," The American Political Science Review, 95, 379–396.

GRAUE, M. E. AND J. DIPERNA (2000): "Redshirting and early retention: Who gets the 'gift of time' and what are its outcomes?" American Educational Research Journal, 37, 509–534.

GREENSTONE, M. AND I. NATH (2020): "Do renewable portfolio standards deliver cost-effective carbon abatement?" Working Paper, Energy Policy Institute at the University of Chicago.

GRIGORIEFF, A., C. ROTH, AND D. UBFAL (2018): "Does information change attitudes towards immigrants? representative evidence from survey experiments," Working paper.

GROBMAN, J. H. AND J. M. CAREY (2002): "The effect of policy uncertainty on wind-power investment," The Journal of Energy and Development, 28, 1–14.

GUENTHER, G. (2018): "The Section 179 and Section 168(k) expensing allowances: Current law and economic effects," Tech. rep., Congressional Research Service.

HAALAND, I. AND C. ROTH (2020): "Labor market concerns and support for immigration," Journal of Public Economics, 191, 104256.

——— (2021): "Beliefs about racial discrimination and support for pro-black policies," The Review of Economics and Statistics, 1–38.

HAALAND, I., C. ROTH, AND J. WOHLFART (2020): "Designing information provision experiments," SSRN Scholarly Paper ID 3644820, Social Science Research Network, Rochester, NY.

HAMILTON, S. D., D. MILLSTEIN, M. BOLINGER, R. WISER, AND S. JEONG (2020): "How does wind project performance change with age in the United States?" Joule, 4, 1004–1020.

HARTIG, H. (2020): "Democrats overwhelmingly favor free college tuition, while Republicans are divided by age, education," .

Heckman, J. J. and E. Vytlacil (2005): "Structural equations, treatment effects, and econometric policy evaluation," Econometrica, 73, 669–738.

Helm, C. and M. Mier (2021): "Steering the energy transition in a world of intermittent electricity supply: Optimal subsidies and taxes for renewables and storage," Journal of Environmental Economics and Management, 109, 102497.

Hemelt, S. W. and R. B. Rosen (2016): "School entry, compulsory schooling, and human capital accumulation: evidence from Michigan," The B.E. Journal of Economic Analysis & Policy, 16.

Holland, S. P., E. T. Mansur, N. Z. Muller, and A. J. Yates (2016): "Are there environmental benefits from driving electric vehicles? The importance of local factors," American Economic Review, 106, 3700–3729.

Hollingsworth, A. and I. Rudik (2019): "External impacts of local energy policy: The case of renewable portfolio standards," Journal of the Association of Environmental and Resource Economists, 6, 187–213.

Holz, J., R. Jiménez Durán, and E. Laguna-Müggenburg (2020): "Quantifying repugnance to price gouging with an incentivized reporting experiment," Working paper.

Howell, W. G. and M. R. West (2009): "Educating the public: How information affects Americans' support for school spending and charter schools," Education Next, 9, 40–48.

Hurst, R., T. Tidwell, and D. Hawkins (2017): "Down the rathole? Public support for US foreign aid," International Studies Quarterly, 61, 442–454.

Illinois, General Assembly (2019): SB2075: An Act Considering Education.

Imbens, G. W. and J. D. Angrist (1994): "Identification and estimation of local average treatment effects," Econometrica, 62, 467–475.

Imbens, G. W. and D. B. Rubin (1997): "Estimating outcome distributions for compliers in instrumental variables models," The Review of Economic Studies, 64, 555–574.

Imlay, S. J. (2021): "Tuition, targeting, and tradeoffs: A conjoint analysis of Americans' preferences over the design of higher education subsidies," The Journal of Higher Education, 1–32.

Ito, K., T. Ida, and M. Tanaka (2021): "Selection on welfare gains: Experimental evidence from electricity plan choice," NBER Working Paper.

Jacobs, B. (2018): "The marginal cost of public funds is one at the optimal tax system," International Tax and Public Finance, 25, 883–912.

Jarvis, S. (2021): "The Economic Costs of NIMBYism-Evidence from Renewable Energy Projects," Working Paper.

Jenkins, J. M. and C. K. Fortner (2019): "Forced to redshirt: Quasi-Experimental impacts of delayed kindergarten entry," Ed Working Papers.

Jimerson, S. R. (2001): "Meta-analysis of grade retention research: Implications for practice in the 21st century," School Psychology Review, 30, 420–437.

Johnson, R. C. and C. K. Jackson (2019): "Reducing inequality through dynamic complementarity: Evidence from Head Start and public school spending," American Economic Journal: Economic Policy, 11, 310–49.

Johnston, S. (2019): "Nonrefundable tax credits versus grants: The impact of subsidy form on the effectiveness of subsidies for renewable energy," Journal of the Association of Environmental and Resource Economists, 6, 433–460.

Kaffine, D. T., B. J. McBee, and S. J. Ericson (2020): "Intermittency and CO_2 reductions from wind energy," The Energy Journal, 41.

Kane, T. J., P. R. Orszag, E. Apostolov, R. P. Inman, and A. Reschovsky (2005): "Higher education appropriations and public universities: Role of Medicaid and the business cycle," Brookings-Wharton papers on urban affairs, 99–146.

Karaduman, Ö. (2021): "Large-scale wind power investment's impact on wholesale electricity markets," Working paper.

Keen, M. and J. Slemrod (2017): "Optimal tax administration," Journal of Public Economics, 152, 133–142.

Kline, P. and C. R. Walters (2016): "Evaluating public programs with close substitutes: The case of Head Start," The Quarterly Journal of Economics, 131, 1795–1848.

——— (2019): "On Heckits, LATE, and numerical equivalence," Econometrica, 87, 677–696.

Kopczuk, W. (2003): "A note on optimal taxation in the presence of externalities," Economics Letters, 80, 81–86.

Kowalski, A. E. (2022a): "Behavior within a clinical trial and implications for mammography guidelines," Review of Economic Studies, forthcoming.

——— (2022b): "Reconciling seemingly contradictory results from the Oregon health insurance experiment and the Massachusetts health reform," Review of Economics and Statistics, forthcoming.

Kraft, M. A. (2020): "Interpreting effect sizes of education interventions," Educational Researcher, 49, 241–253.

Kuziemko, I., M. I. Norton, E. Saez, and S. Stantcheva (2015): "How elastic are preferences for redistribution? Evidence from randomized survey experiments," American Economic Review, 105, 1478–1508.

Laderman, S. and K. Heckert (2021): "SHEF: State Higher Education Finance, FY 2020." State Higher Education Executive Officers.

LaLonde, R. J. (1986): "Evaluating the econometric evaluations of training programs with experimental data," American Economic Review, 604–620.

Layton, T. J., M. L. Barnett, T. R. Hicks, and A. B. Jena (2018): "Attention deficit-hyperactivity disorder and month of school enrollment," New England Journal of Medicine, 379, 2122–2130.

Lee, D. S. and T. Lemieux (2010): "Regression discontinuity designs in economics," Journal of Economic Literature, 48, 281–355.

Lee, J. M. and G. Howard (2021): "The impact of technical efficiency, innovation, and climate policy on the economic viability of renewable electricity generation," Energy Economics, 100, 105357.

Lergetporer, P., G. Schwerdt, K. Werner, M. R. West, and L. Woessmann (2018): "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, P. and B. Smarzynska Javorcik (2019): "The political economy of higher education finance: How information and design affect public preferences for tuition," CESifo Working Paper.

Lergetporer, P., K. Werner, and L. Woessmann (2020): "Educational inequality and public policy preferences: Evidence from representative survey experiments," Journal of Public Economics, 188, 104226.

Liski, M. and I. Vehviläinen (2020): "Gone with the wind? An empirical analysis of the equilibrium impact of renewable energy," Journal of the Association of Environmental and Resource Economists, 7, 873–900.

List, J. A. (2001): "Do explicit warnings eliminate the hypothetical bias in elicitation procedures? Evidence from field auctions for sportscards," American Economic Review, 91, 1498–1507.

List, J. A. and C. A. Gallet (2001): "What experimental protocol influence disparities between actual and hypothetical stated values?" Environmental and Resource Economics, 20, 241–254.

List, J. A., P. Sinha, and M. H. Taylor (2006): "Using choice experiments to value non-market goods and services: Evidence from field experiments," Advances in Economic Analysis & Policy, 5.

Lohawala, N. (2022): "Roadblock or accelerator? The effect of electric vehicle subsidy elimination," Working paper.

Mankiw, N. G. (2020): Principles of Economics, Cengage Learning, ninth ed.

McCrary, J. and H. Royer (2011): "The effect of female education on fertility and infant health: Evidence from school entry policies using exact date of birth," American Economic Review, 101, 158–95.

Mettler, S. (2014): Degrees of inequality: How the politics of higher education sabotaged the American dream, Basic Books.

Mogstad, M., A. Santos, and A. Torgovitsky (2018): "Using instrumental variables for inference about policy relevant treatment parameters," Econometrica, 86, 1589–1619.

Mogstad, M., A. Torgovitsky, and C. R. Walters (2020): "The causal interpretation of two-stage least squares with multiple instrumental variables," NBER Working Paper.

Molnar, T. L. (2020): "The impact of academic redshirting on student outcomes and mental health: Evidence from Hungary," Working paper.

Moretti, E. (2004): "Human capital externalities in cities," in Handbook of Regional and Urban Economics, Elsevier, vol. 4, 2243–2291.

Morrison, F. J., D. M. Alberts, and E. M. Griffith (1997): "Nature-nurture in the classroom: Entrance age, school readiness, and learning in children," Developmental Psychology, 33, 254.

Ng, Y.-K. (1980): "Optimal corrective taxes or subsidies when revenue raising imposes an excess burden," American Economic Review, 70, 744–751.

Novan, K. (2015): "Valuing the wind: Renewable energy policies and air pollution avoided," American Economic Journal: Economic Policy, 7, 291–326.

Nyhan, B. and J. Reifler (2010): "When corrections fail: The persistence of political misperceptions," Political Behavior, 32, 303–330.

Oates, W. E. (1972): Fiscal federalism, The Harbrace series in business and economics, New York: Harcourt Brace Jovanovich.

——— (1999): "An essay on fiscal federalism," Journal of Economic Literature, 37, 1120–1149.

Parish, R. M. and K. R. McLaren (1982): "Relative cost-effectiveness of input and output subsidies," Australian Journal of Agricultural Economics, 26, 1–13.

Parker, K. (2019): "The growing partisan divide in views of higher education," Pew Research Center.

Peña, P. A. (2017): "Creating winners and losers: Date of birth, relative age in school, and outcomes in childhood and adulthood," Economics of Education Review, 56, 152–176.

Petersen, C., M. Reguant, and L. Segura (2021): "Wind power and intermittency: The impact of subsidy design," Working paper.

Peyton, K. (2020): "Does trust in government increase support for redistribution? Evidence from randomized survey experiments," American Political Science Review, 114, 596–602, publisher: Cambridge University Press.

Ponzo, M. and V. Scoppa (2014): "The long-lasting effects of school entry age: Evidence from Italian students," Journal of Policy Modeling, 36, 578–599.

Rambachan, A. and J. Roth (2022): "A more credible approach to parallel trends," Working paper.

Rokkanen, M. A. (2015): "Exam schools, ability, and the effects of affirmative action: Latent factor extrapolation in the regression discontinuity design," Working paper.

Roth, C., S. Settele, and J. Wohlfart (2021): "Beliefs about public debt and the demand for government spending," Journal of Econometrics.

Routon, P. and J. K. Walker (2020): "Older and wiser? Relative age and success in high school and college," SSRN Working Paper.

Sandmo, A. (1975): "Optimal taxation in the presence of externalities," The Swedish Journal of Economics, 86–98.

Schanzenbach, D. W. and S. H. Larson (2017): "Is your child ready for kindergarten?" Education Next, 17.

Schmalensee, R. (2012): "Evaluating policies to increase electricity generation from renewable energy," Review of Environmental Economics and Policy.

Settele, S. (2019): "How do beliefs about the gender wag gap affect the demand for public policy?" SSRN Scholarly Paper ID 3382325, Social Science Research Network, Rochester, NY.

Sevenans, J. (2021): "How public opinion information changes politicians' opinions and behavior," Political Behavior.

Shapiro, A., E. Martin, C. Weiland, and R. Unterman (2019): "If you offer it, will they come? Patterns of application and enrollment behavior in a universal prekindergarten context," AERA Open.

Sharpiro, A. (2020): "Age at time of kindergarten entry and special education service receipt," Working paper.

Shavell, S. (2011): "Corrective taxation versus liability," American Economic Review, 101, 273–76.

Snyder, T. D., C. De Brey, and S. A. Dillow (2019): "Digest of education statistics 2017, NCES 2018-070." National Center for Education Statistics.

Stantcheva, S. (2020): "Understanding tax policy: How do people reason?" NBER Working Paper w27699.

Stevenson, B. and J. Wolfers (2020): Principles of Economics, Vital Source, first ed.

Su, S. (2022): "Updating politicized beliefs: How motivated reasoning contributes to polarization," Journal of Behavioral and Experimental Economics, 96, 101799.

Taylor, B. J., B. Cantwell, K. Watts, and O. Wood (2020): "Partisanship, white racial resentment, and state support for higher education," The Journal of Higher Education, 91, 858–887.

United States Congress (1991): "Congressional record - Senate; February 21, 1991," Congressional record permanent digital collection.

United States Energy Information Administration (2022): "How much electricity does an American home use?" Website.

United States Environmental Protection Agency (2022): "Greenhouse gases equivalencies calculator - Calculations and references," Website.

Vytlacil, E. (2002): "Independence, monotonicity, and latent index models: An equivalence result," Econometrica, 70, 331–341.

Walters, C. R. (2018): "The demand for effective charter schools," Journal of Political Economy, 126, 2179–2223.

Wiser, R. and M. Bolinger (2021): "Land-Based Wind Market Report: 2021 Edition," Tech. rep., Lawrence Berkeley National Laboratory.

Yi, F., C.-Y. L. Lawell, and K. Thome (2018): "A dynamic model of subsidies: Theory and application to ethanol industry," Working paper.