Self-Selection Around Policy Recommendations: The Case of Kindergarten Entry

Michael David Ricks*

December 2024

Abstract

This paper quantifies the efficiency and equity implications of influencing behavior with policy recommendations versus requirements in the context of kindergarten entry. Because recommendations allow opting in and opting out ("self-selection"), welfare comparisons depend on how gains accrue to those who self-select ("selection on gains"). Using variation from multiple birthday-based discontinuities, I estimate these heterogeneous effects among the universe of kindergartners in Michigan public schools and find that only higher-income families positively select on test-score gains. Consequently, whereas self-selection raises average test scores of higher-income families, it reduces those of lower-income families, creating an equity-efficiency tradeoff.

*Department of Economics, University of Nebraska - Lincoln. Email: mricks4@unl.edu.

This research is the product of generous amounts of feedback and conversations with many wonderful and intelligent people including my dissertation committee, Amanda Kowalski, Brian Jacob, Jim Hines, Sara Heller; scholars like Ash Craig, Jeff Denning, John Bonney, John Friedman, Chris Walters, Rich Patterson, Joe Price, Jeff Smith, Garrett Anstreicher, Christina Weiland, Lars Lefgren, Matias Cattaneo, Gonzalo Vasquez-Bare, Peter Hull, Jason Baron, Charlie Brown, John Bound, Mike Mueller-Smith, Daphna Bassok, Julian Betts, Kevin Stange, Celeste Carruthers, Andrew Simon, Maggie Shi, Tyler Radler, Owen Kay, Jordy Berne, Stephanie Owen, Brenden Timpe, Nathan Sotherland, Neil Christy, Katie Laursen, Ryan Fraser, Jamie Fogel, Thomas Helgerman, Stacey Brockman, and other researchers at the Education Policy Initiative and Youth Policy Lab; as well as with seminar participants at the University of Michigan, Opportunity Insights, Brigham Young University, and the Nebraska Labor Summit. Thanks also to Jasmina Camo-Biogradlija, Kyle Kwaiser, and Nicole Wagner Lam who facilitated the process of data access and to Richard Lower and others at the Michigan Department of Education for their interest and feedback.

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.

1. Introduction

Many public programs involve self-selection—decisions to opt into or out of "treatment." Consider, for example, participation in needs-based services and job training; decisions about school-choice, medical providers, and utility plans; and even responses to cancer-screening and vaccination guidelines. Policy can either influence behavior using recommendations that allow self-selection (like guidelines, nudges, or defaults) or requirements (which mandate or enforce a given decision). The central trade-off is whether or not self-selection promotes social criteria for efficiency (i.e., "selection on welfare gains" Ito et al., 2021) and equity. This paper studies the nature of self-selection and its implications for policy in the context of kindergarten entry.

I consider kindergarten entry timing for four reasons. First, it is a nearly universal childhood investment decision.[1] Each year 120 million families from across the world choose between beginning public school (usually in the year children turn five) or waiting until the following year. Second, self-selection may have ambiguous effects on children. On one hand, parents have private information, so self-selection could facilitate efficient sorting (see discussion in Oster, 2022). On the other hand, as a custodial decision, self-selection could be inefficient under incomplete parental altruism or information (as in Cornelissen et al., 2018). Third, equity is often an explicit rationale for kindergarten-entry requirements (e.g., see discussion surrounding Illinois, General Assembly, 2019). Finally, the incomplete understanding of self-selection leaves core policy questions unresolved—as evidenced by the global policy dispersion and active conversations comparing recommendations and requirements.

Whether self-selection is efficient or equitable depends theoretically on the nature of effect heterogeneity. I first model parents' choices to start children in public kindergarten at age five (thus ending full-time private investments) or wait a year (and continue investments). Defining waiting as the "treatment" of interest (as in Black et al., 2011), I show that the welfare difference between recommendations and requirements is a function of the marginal treatment effects of waiting on children who opt into and out of waiting. This result reveals a critical oversight in the extensive "school entry" literature. This literature clearly documents how waiting affects children who comply with birthday-based recommendations or requirements (e.g., Bedard and Dhuey, 2006; Elder and Lubotsky, 2009; Black et al., 2011; Dhuey et al., 2019), but welfare depends on how waiting affects those who self-select. Little is known about this selection on gains or "essential heterogeneity" (Heckman et al., 2006), and the systematic differences associated with self-selection (Bassok and Reardon, 2013; Schanzenbach and Larson, 2017; Fortner and Jenkins, 2017) complicate any generalizations from existing causal research.

I empirically estimate the heterogeneous effects of waiting among the universe of kindergart-

[1]For example, roughly 90% of primary-school-age children worldwide begin public school, compared to 38% participating in early childhood education (see UNICEF (2022) and UNICEF (2024)).

ners in Michigan public schools. Identifying variation comes from policy changes that shifted birthday cutoffs and changed them from requirements to recommendations—generating multiple regression-discontinuity-style natural experiments identifying self-selection. For example, at December 1, roughly 20% of families opt into waiting, and 30-40% would opt out if allowed. Consistent with prior work (e.g., Elder and Lubotsky, 2009), the average effect of waiting at this cutoff is 0.23 standard deviations in third-grade math scores.[2] The RD effects also produce repeated evidence of positive selection on gains but because they do not identify the effect of waiting on families who opt in, they only partially inform welfare.

To fully describe self-selection, I combine the RD effects on various complier groups to identify the marginal treatment effects of waiting on third-grade math scores. This exercise is made possible by mapping the MTE framework (e.g., Heckman and Vytlacil, 2005; Mogstad et al., 2018) into an RD setting with multiple cutoffs (Cattaneo et al., 2016) under a parallel trends assumption similar to that in Cattaneo et al. (2020). The resulting MTE estimates are linear over the distribution of family unobservables; are robust to alternative functional forms, shape restrictions, and bounding methods; and are essentially identical whether based on within-year comparisons across cutoffs or same-cutoff comparisons across years.

These results reveal two insights into the economics of self-selection in kindergarten entry. First, I document substantial positive selection on gains and negative selection in levels. Positive selection on gains means that children who opt into waiting experience 2.4 times larger achievement gains on average (0.44 standard deviations) than children who opt out of waiting (0.18). Negative selection on levels means that the children who are more likely to wait would tend to be lower achieving in third grade had they started kindergarten at age five. Taken together these facts suggest that because of private information about school readiness, parents tend to know which student will (and will not) benefit from "the gift of time" and self-select accordingly. Mechanisms show that heterogeneity in gains is associated with additional private investments in the year of waiting and that effects are not related to age-rank effects (which could reflect rent seeking behavior).

Second, the data also demonstrate *differential* selection on gains between lower- and higher-income families. I find that only higher-income families positively select on gains whereas lower-income families experience much more homogeneous effects. Interpreted through the model, this differential selection suggests that higher-income families self-select in response to the expected benefits of waiting whereas lower-income families' self-selection may be driven more by the immediate costs. This interpretation is corroborated by other research showing how lower-income families respond strongly to transportation costs in early childhood education (e.g., Kline and Walters, 2016) and school choice (e.g., Walters, 2018). This differential selec-

[2]For context, a meta-analysis of 852 randomized controlled experiments in K-12 education estimates this as the 80th percentiles of intervention impacts on third-grade math scores (Kraft, 2020).

tion underpins policy trade-offs that would have been assumed away by the shape restrictions typically used to identify marginal treatment effects (Kline and Walters, 2016; Brinch et al., 2017; Mogstad et al., 2018).[3]

Combining these insights to consider policy, I find an equity-efficiency trade-off. Simulations based on self-selection shares and marginal treatment effects reveal that (relative to requirements) recommendations tend to raise test scores for children from higher-income families but lower scores for children from lower-income families. These patterns occur as higher-income families to opt in (redshirt) children who benefit substantially from waiting whereas many lower-income families to opt-out children (enter early) who would otherwise experience gains from waiting. As a result, an achievement-neutral change to a recommendation would widen income-achievement gap by 12%. Recommendations also tend to raise average test scores, suggesting although equity motivations for requirements are well founded, equity-driven policy decisions can have unanticipated efficiency costs—presenting an equity-efficiency tradeoff. Using the simulation to compare MTE estimation methods further reveals revealing that models using typical shape restrictions altogether fail to identify this tradeoff.

These results have implications for three lines of academic inquiry. First, my characterization of selection adds an important economic perspective to conversations about selection around kindergarten recommendations specifically and into public program participation more generally. Education policy research on kindergarten entry has largely relied on characterizing selection in covariates (e.g., Cameron and Wilson, 1990; Graue and DiPerna, 2000; March, 2005; Bassok and Reardon, 2013; Schanzenbach and Larson, 2017) or realized outcomes (e.g., Fortner and Jenkins, 2017). Because this work finds that children who opt into waiting have high realized test scores and covariates associated with higher achievement, one prevailing narrative predicts positive selection on levels and negative selection on gains.[4] My contributions are in presenting a model-based characterization of selection and then estimating the selection on gains. These causal estimates overturn the prevailing narrative and highlight a Roy-style comparative advantage in self-selection. This reversal highlights a need to understand selection into public programs in terms of economic primitives rather than correlates—as recently done with evaluations of the moral hazard in the Oregon Health Insurance Experiment (Kowalski, 2022) and selection on gains in the Head Start Impact Study (Kline and Walters, 2016).

Second, characterizing the economics of self-selection and resulting policy implication are important contributions to a large causal literature on school entry. This literature has built a wealth of knowledge about interesting causal effects—including on educational outcomes;[5] labor

[3] Cornelissen et al. (2016) and Kline and Walters (2019) discuss these assumptions and their implications.

[4] An exception to this trend is Fortner and Jenkins (2018) who present evidence of lower test scores among those with special education diagnoses consistent with my results of negative selection in levels.

[5] For example, Bedard and Dhuey (2006), Datar (2006), McEwan and Shapiro (2008), Elder and Lubotsky (2009), Smith (2009), Black et al. (2011), Dhuey and Lipscomb (2010), Kawaguchi (2011), Hemelt and Rosen

market outcomes;[6] medical diagnoses and mental health;[7] and even social outcomes like fertility, bullying, and crime.[8] However, the nature of economic decision making and self-selection are still not well understood. My paper studies this behavior and documents positive selection on gains by higher-income families and null selection on gains by lower-income families. A concurrent working paper, Molnar (2024), takes steps toward making a similar contribution. Using a January cutoff that bans opting in and a main birthday cutoff in June, it compares the effects of waiting across groups—focused mainly on gender gaps. Although I identify the effects of waiting over a broader support of students (including all students who opt in or out rather than those most likely to opt in but forced not to), my main innovation going beyond policy evaluation to describe the economics of selection. Furthermore, while this entire literature uses policy variation from recommendations and requirements, my paper is the first to acknowledge the importance of the marginal treatment effects for welfare inferences, to provide an econometric approach extracting those patterns from the RD estimates, or to use those primitives to answer optimal birthday-cutoff policy questions.

Finally, the differential selection on gains demonstrates the importance of effect heterogeneity in broader discussions of empirical welfare economics, public policy, and behavioral public economics. With regard to welfare economics and policy, my contribution is showing that estimating heterogeneity across *both* unobservable and observable dimensions is a first-order welfare consideration. Outside of education, strategic self-selection may lead to inefficient or inequitable allocations in decisions as varied as applying for means tested services (e.g., Finkelstein and Notowidigdo, 2019), health care provider matching (Brown et al., 2023), or selecting plans in public utility and annuity markets (Einav et al., 2010; Ito et al., 2021).[9] Given evidence that it is often disadvantaged individuals who benefit the least from self-selection (e.g., Kline and Walters, 2016; Abdulkadiroğlu et al., 2018; Chen and Pereyra, 2019), it is likely that both types of heterogeneity will be welfare relevant and that allowing for similarly heterogeneous responses is critical in optimally structuring other public programs that allow for self-selection. Relatedly, research in behavioral public economics has identified cases where more choice can result in inefficiency or inequity because of systematic mis-optimization (e.g., Allcott et al., 2022; Lockwood et al., 2024). My results highlight how choice benefits some (in this case all

(2016), Tan (2017), Attar and Cohen-Zada (2018), and Dhuey et al. (2019).

[6]For example, Dobkin and Ferreira (2010), Black et al. (2011), Mühlenweg et al. (2012), and Bedard and Dhuey (2012).

[7]For example, Elder (2010), Evans et al. (2010), Matsubayashi and Ueda (2015), Layton et al. (2018), and Dee and Sievertsen (2018)

[8] Mühlenweg (2010), McCrary and Royer (2011), Cook and Kang (2016), Depew and Eren (2016), McAdams (2016), and Tan (2017).

[9]Education examples where self-selection could lead to inefficient or inequitable outcomes could include participation in free prekindergarten (Kline and Walters, 2016; Cornelissen et al., 2018; Felfe and Lalive, 2018), school boundaries and school choice (Walters, 2018; Chen and Pereyra, 2019; Neilson, 2021), or decisions about colleges and college majors (Brand and Xie, 2010; Kirkeboen et al., 2016; Mountjoy, 2022).

parents and children from higher-income families) but harms others (children from lower-income families) likely because of the costs associated with waiting.

While not the main focus of the paper, obtaining my substantive results required a methodological innovation: using marginal treatment effects in an RD to extrapolate away from the discontinuity. These results are mainly in Appendix D, but the key insight is that marginal treatment effects apply to units away from the cutoff under a testable "parallel trends" assumption similar to those used for a multi-cutoff-RD (Cattaneo et al., 2016, 2020). This novel extrapolation method complements existing approaches based on differences in treated and untreated outcomes (Bertanha and Imbens, 2019; Cattaneo et al., 2020), conditional independence (Angrist and Rokkanen, 2015; Rokkanen, 2015), and estimated treatment effects (e.g., Dong and Lewbel, 2015; Cerulli et al., 2017). My contribution is deriving a method that neither requires that the LATE be homogeneous or "externally valid" (as in Dong and Lewbel, 2015; Bertanha and Imbens, 2019) nor that conditional potential outcomes be constant conditional on observables (as in Angrist and Rokkanen, 2015). This approach could be useful in any setting where a fuzzy RD reflects self-selection into binary treatment, but where core policy questions involve understanding effects away from the RD cutoff, such as school admissions cutoffs (Altmejd et al., 2021; Barrios-Fernández et al., 2024), workplace safety regulations (Johnson, 2020), public works and infrastructure (Asher and Novosad, 2020), and more.

The remainder of the paper includes the following sections: (2) describing the conceptual framework of self-selection and welfare; (3) estimating the effects of waiting with the regression discontinuity and assessing the extent of selection on gains under the RD assumptions; (4) characterizing the (differential) selection on gains in a marginal treatment effects framework and considering mechanisms; (6) conducting policy simulations to quantify welfare effects; and (7) containing my conclusion, discussion of results, and possible future research.

2. Background and Conceptual Framework for Self-Selection and Welfare

There are many settings in which a policymaker chooses between influencing behavior with a light-touch intervention like a recommendation or with a requirement. Applied economic research suggests three set of reasons why individuals may not self-select on welfare gains and, therefore, why requirements could be welfare-improving. First, under incomplete information, there may be harms from self-selecting around recommendations. For example, in health economics there is evidence that the individuals most likely to opt out of vaccines for COVID-19 experienced high levels of mortality (Wallace et al., 2022) whereas the individuals most likely to opt into cancer screenings are more likely to be harmed by the detection and treatment of benign tumors (Einav et al., 2020; Kowalski, 2023a). Second, decision makers may not internalize the gains of their decisions on others. For example, in education economics there is evidence

that the children who would benefit most from public services like early childhood education and charter schools are most likely to be opted out when eligible (Kline and Walters, 2016; Cornelissen et al., 2018; Felfe and Lalive, 2018; Walters, 2018). Third, inequities in revealed-preferred allocations may conflict with policy makers' distributional objectives. For example, in criminal justice requirements can restrict bias and discrimination (Kaplow, 1992; Grunwald, 2015) such as when judges deviate from algorithmic recommendations to give longer sentences to Blacks relative to Whites (Stevenson and Doleac, 2022; Angelova et al., 2023).

This section presents a framework in which decision makers respond to recommendations and requirements regarding a binary treatment decision. The framework characterizes the nature of self-selection and the implications for welfare and policy—including whether traditional estimates sufficiently describe efficiency and equity. I present the framework in the context of families' decisions about kindergarten entry, but it applies to school choice, medical screening, algorithmic sentencing recommendations, or any other setting with recommendations and self-selection.

2.1 Treatment Decision: Waiting to Enter Kindergarten

Families with school-aged children typically face two school-entry options: to start kindergarten in the year children turn five or to wait until the following year. Waiting does not mean inaction; rather it extends children's natural development and families' full-time private investments. As such this paper focuses on the binary decision to wait $w_i \in \{0, 1\}$ (as in Black et al. (2011) and Dhuey et al. (2019)).[10] Families decide whether to wait by comparing the anticipated benefits (e.g., physical, cognitive, and social development) with the costs (e.g., childcare and foregone wages). These decisions are often informed by private information, caregiver assessments, and—importantly—kindergarten entry policies.

2.2 Policy Recommendations and Requirements

Most kindergarten entry policies use birthday-based school entry cutoffs. Cutoffs generally recommend (or require) waiting ($w = 1$) for children who turn five after the cutoff date and starting ($w = 0$) for children who turn five before. Formally, a cutoff is monotonic policy function $z = f(r)$, that returns recommendations or requirements, z, given individual characteristics, r. Equation 1 illustrates potential values of $z = f(r)$ in the kindergarten context based on birthdays, r.

[10]Defining "treatment" as waiting also captures the one behaviorally-relevant decision for parents and policy makers. Although academics consider the effects of "entry age," "relative age," "testing age," (and their interactions) separately (Black et al., 2011; Cascio and Schanzenbach, 2016; Peña, 2017, e.g.,), these constructs are not separately manipulable by decision makers. Furthermore, this binary treatment avoids well-known identification issues with continuous definitions of "entry-age" treatment (e.g., Angrist and Pischke, 2008; Barua and Lang, 2016).

$$f(r): \begin{cases} z = 0 & \text{Pure control: require starting at five} \\ z = 1 & \text{Experimental control: recommend starting at five} \\ z = 2 & \text{Weak-touch intervention: recommend waiting until six} \\ z = 3 & \text{Strong intervention: require waiting until six} \end{cases} \tag{1}$$

A birthday cutoff is a birthday, r^*, at which the policy function $f(r)$ changes value. For example, under one birthday cutoff policy children who turn five on or before September 1 might be required to start and children who turn five after September 1 are required to wait. Under a different policy, children who turn five on or before July 15 are recommended to start and children who turn five after July 15 are recommended to wait. Appendix Table A.1 presents a list of current policies in 95 countries, states, and provinces, illustrating tremendous variation both in regard to cutoff dates and the use of recommendations versus requirements. Anecdotes suggest that many of these policies were set idiosyncratically rather than by optimization and that they remain due to hysteresis even in the face of dramatic changes in self-selection in the past 30 years (e.g., Deming and Dynarski, 2008). Perhaps this in part motivates growing policy conversations about optimal adjustments to birthday cutoff policies (e.g., Reeves, 2022).

Self-selection is characterized by two types of choices about starting or waiting in response to birthday cutoffs. First, when $z = 1$, some families that are recommended to start kindergarten might opt into waiting (a practice often called "academic redshirting"). A requirement, $z = 0$, would prevent these families from waiting. Second, when $z = 2$, other families that are recommended to wait might opt out and start kindergarten anyway (a practice called "early entry"). A requirement, $z = 3$, would force these families to wait.[11] The behavior of families who already follow recommendations would not be affected by a requirement.

2.3 Kindergarten Entry Recommendations and Requirements in Michigan

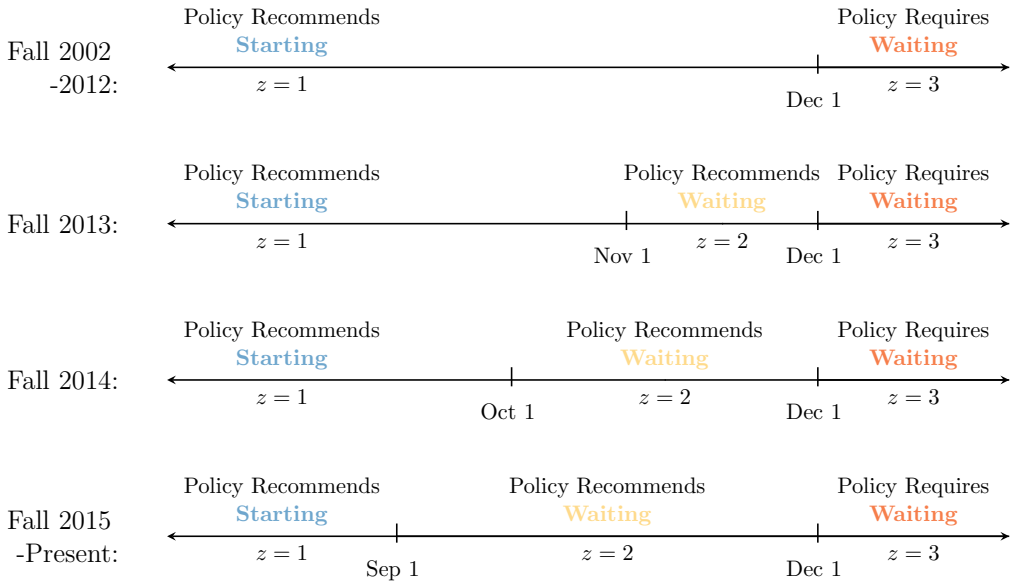
Each year nearly 100,000 Michigan children enter kindergarten—typically at age five.[12] Figure 1 visualizes 20 years of changing kindergarten entry policies changing policies. It presents recommendations and requirements as a function of student birthdays across cohorts. Before

[11]Note that this characterization of self-selection also provides a useful insight into the typical taxonomy of early entry, on-time entry, and academic redshirting used in school entry research (Bassok and Reardon, 2013; Cook and Kang, 2018; Fortner and Jenkins, 2017, see). These groups are not categorically different decisions with implicit potential outcomes (as characterized, for example, in Cook and Kang (2018) and Jenkins and Fortner (2019)); rather, they are manifestations of self-selection in to or out of waiting. In fact, defining three potential outcomes {early, on-time, redshirt} for each child would require the option of enrolling some three-year-old children who are about to turn four as well as some six-year-old children who are about to turn seven—behaviors that are both empirical anomalies and violations of most compulsory schooling laws.

[12]Although districts in Michigan are not required to offer kindergarten, if it is offered, students may not skip it. Most elementary school students (99.86%) attend kindergarten-offering districts.

2012, students who turned five on or before December 1 were recommended to start kindergarten ($z = 1$) but could opt into waiting ("redshirting").[13] Students who turned five after December 1 were required to wait and could not opt out ($z = 3$, no "early entry"). Starting in 2012, Michigan began moving the cutoff back one month each year from December 1 in 2012, to November 1 in 2013, October 1 in 2014, and then September 1 in 2015.[14] Acknowledging the effect of this change on families' decisions, Michigan allowed children who turned five after the fall cutoff but before December 1 to opt out of waiting (assigning these children $z = 2$). Children turning five after December 1, however, were still required to wait.

Figure 1: Overview of Changing Birthday Cutoff Policies in Michigan



Note: This figure depicts the changing birthday cutoff policies in Michigan from 2002-Present. Each series depicts the kindergarten entry recommendations given or requirements made based on children's birthdays.

These changes generate discontinuity-based policy experiments that affect distinct groups of compliers. In this paper, I follow the naming conventions of Mogstad et al. (2021) to refer to these children. First, children who wait no matter the recommendation (including opting in when needed) are *always takers*. Second, the recommendation cutoffs in September, October, and November (2013-Present) induced some children to wait, but not those who would rather opt out. These children would follow either a recommendation or a requirement to wait and so are *eager compliers*. Third, the December cutoffs in these years affect children who only wait if required (and who opt out when they can). These are *reluctant compliers*. Finally, in this context, the long-standing December cutoff (1900s-2012) induced both eager and reluctant

[13] This is also the model birthday cutoff policy in US States in Appendix Table A.1, albeit with varying dates.

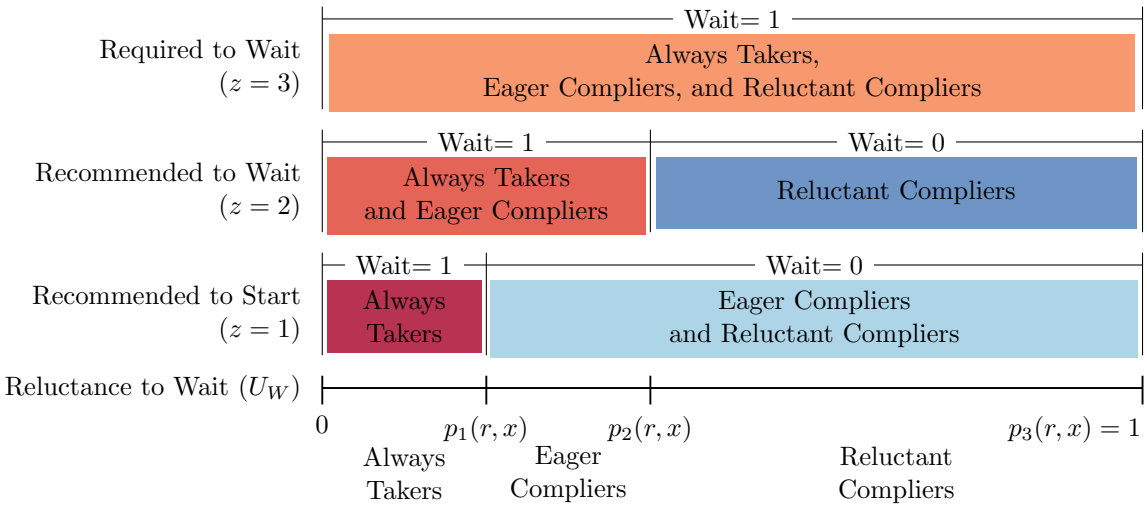
[14] Conversations with policy makers suggest that one major motivation for this change was raising standardized test scores to catch up with states using earlier cutoffs.

compliers to wait, i.e., both those who would follow recommendations and those who would rather opt out.

2.4 Self-Selection and the Effects of Waiting

To consider self-selection, this subsection overviews a model of selection and child outcomes (with formal details in Appendix B.1). Let each family i have observable characteristics x, child's birthday r, and two unobservables, γ affecting outcomes and $U_W \sim [0,1]|r,x$ reflecting the reluctance to wait.[15] Defining the share of families who wait under a given recommendation or requirement, z, as $p_z(x,r) = P(w=1|z,r,x)$ reveals how selection is determined by both the reluctance to wait and the policy recommendations or requirements. The families who wait under each policy are those with the lowest unobserved reluctance $U_W \leq p_z(x,r)$. For example, the only students who opt into waiting when recommended to start are always takers, thus always takers are the share of $p_1(x,r)$ families with the lowest unobserved reluctance to wait, U_W; similarly, because reluctant compliers opt out unless required, they are the $1-p_2(x,r)$ with the highest reluctance. This produces a strict ordering across groups that is visualized in Figure 2: reluctant compliers are more reluctant to wait than eager compliers, and eager compliers are in turn more reluctant to wait than always takers.

Figure 2: 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, $U_W \in [0,1]$. The values $p_z(r,x)$ represent the probability of waiting when facing different recommendations or requirements, z. Given birthdays r, characteristics x, and incentive z, students with values $U \leq p_z(r,x)$ will wait to enter kindergarten and others will not.

[15] Viewing U_W as the percentile of reluctance or "resistance" to treatment is without loss of generality (Mogstad et al., 2018).

To consider the effects of waiting, I assume a child's potential third-grade math scores, Y_w, depend on birthdays, r; characteristics, x; reluctance to wait, U_W; and other unobservables, $\boldsymbol{\gamma}$; but not on the policy recommendation or requirement, z.I define how the effects of waiting, $Y_1 - Y_0$, vary along the reluctance to wait, U_W:

$$\tau_{MTE}(u, r, x) \equiv \mathbb{E}[Y_1 - Y_0 | U_W = u, R = r, X = x]$$

This is the marginal treatment effect (MTE) function of Heckman and Vytlacil (2005) and of Mogstad et al. (2018) adapted to an RD setting. Given the strict ordering between always takers, eager compliers, and reluctant compliers over U_W, the MTE curve can characterize the nature of "selection on gains." Under positive selection on gains, children who are more reluctant to wait (higher U_W) would experience smaller test score improvements than those who are more likely to wait, implying τ_{MTE} is downward-sloping in u. Negative selection on gains, on the other hand, would imply an upward-sloping MTE curve.

2.5 The Efficiency and Equity of Self-Selection around Recommendations

Policy conversations about kindergarten entry focus heavily on the efficiency and equity implications of self-selection. For example, many countries and US states have requirements forcing children with birthdays after the cutoff to wait—often to raise average achievement by having older classes. Proposals requiring children with birthdays before the cutoff to start kindergarten at age five (e.g., Illinois, General Assembly, 2019) often aim to prevent self-selection by high-income families from exacerbating achievement gaps. Both efficiency and equity considerations can be understood within the theoretical framework, as can more nuanced policy considerations like moving birthday cutoffs (as in Cook and Kang, 2018) or requiring all boys to wait to enter kindergarten (Reeves, 2022).

Knowing the process of self-selection and outcome formation enables policy comparisons. Because recommendations allow self-selection and requirements do not, these two policies can be described with different policy functions (given a cutoff date, r^*):

$$f_{\text{rec}}(r) : \begin{cases} z = 1 \text{ if } r < r^* \\ z = 2 \text{ if } r \geq r^* \end{cases} \qquad f_{\text{req}}(r) : \begin{cases} z = 0 \text{ if } r < r^* \\ z = 3 \text{ if } r \geq r^* \end{cases}$$

To compare recommendations and requirements I define criteria for efficiency and equity. When comparing allocations, an allocation that is revealed preferred by families *and* results in higher average test scores for children is more efficient. Allocations with smaller achievement gaps are more equitable.[16] Appendix B.2 derives and discusses these criteria, how the (partially)

[16]The achievement gap for a characteristic x is $\big|\mathbb{E}[Y|X = x] - \mathbb{E}[Y|X \neq x]\big|$.

order allocations, and which strategic interactions and welfare functions they are robust to.

These criteria allow welfare comparisons between recommendations and requirements. Although families weakly prefer allocations generated by recommendations, the difference in average scores $\Delta\mathcal{Y} = \mathcal{Y}_{\text{rec}} - \mathcal{Y}_{\text{req}}$ is a quantitative question. Equation 2 characterizes the difference as the total effect of waiting on always takers (who are allowed to opt in) minus the total effect on reluctant compliers (who are allowed to opt out). As long as families' choices and test scores do not affect each other, these quantities are functions of the group-specific marginal treatment effects, $\tau_{MTE}(u, x)$ and self-selection shares, $p_z(r, x)$:

$$\Delta\mathcal{Y}(x) = \underbrace{\int_{\underline{r}}^{r^*} \int_0^{p_1(r,x)} \tau_{MTE}(u, r, x)\, \mathrm{d}u\, \mathrm{d}F(r|x)}_{\textbf{Total Effect on Always Takers}} - \underbrace{\int_{r^*}^{\bar{r}} \int_{p_2(r,x)}^1 \tau_{MTE}(u, r, x)\, \mathrm{d}u\, \mathrm{d}F(r|x)}_{\textbf{Total Effect on Reluctant Compliers}} \tag{2}$$

Knowing $\Delta\mathcal{Y}(x)$ identifies total effect on achievement, $\Delta\mathcal{Y} = \int_X \Delta\mathcal{Y}(x)\, \mathrm{d}F(x)$, and quantifies changes in achievement equity. Estimates of τ_{MTE} and $p(r)$ also admit policy comparisons like moving cutoff dates or relaxing requirements on only one side of a cutoff (see full discussion in Appendix B.3).

Equation 2 presents two insights about welfare comparisons. First, when the policy question involves possibly restricting self-selection, effect heterogeneity is a first-order welfare consideration. When comparing recommendations and requirements, the effects on always takers and reluctant compliers are they parameters of interest, but these effects are not identified in the rich LATE-based literature. This welfare comparison between requirements and recommendations complements similar results about *laissez faire* self-selection where the comparison between "takers" and "non-takers" is key (Ida et al., 2022). The importance of unobserved heterogeneity in Equation 2 also highlights a gap in a growing literature on targeting. Effect heterogeneity is crucial for optimal targeting (Kitagawa and Tetenov, 2018; Athey and Wager, 2021), but observables are not the only important dimension.

Second, this result also implies that the shape restrictions typically used to identify selection models can bias inferences about both efficiency and equity. Marginal treatment effects are typically identified by assuming additive separability between observables and unobservables (see Kline and Walters, 2019), but restricting $\tau_{MTE}(x, u) = \tau_x + \tau_{MTE}(u)$ implies that all groups select on gains in the same way. Appendix B.4 shows that whenever groups have different treatment effects and self-selection shares, welfare estimates based on additively separable marginal treatment effects are systematically biased by the differential selection.

Whether these analytical results are of first-order welfare importance depends on two related empirical questions. First, is the share of always takers and reluctant compliers large? Answering this question requires policy variation that separates eager and reluctant compliers. This

type of variation is rare in the kindergarten entry literature. Second, how does waiting affect always takers and reluctant compliers? Because traditional LATE estimates can only answer this question if effects are homogeneous, it is important to assess the extent of heterogeneity. Given the results about differential self-selection, this assessment must include heterogeneity on both observed and unobserved dimensions.

3. Patterns of Kindergarten Entry: Data, Descriptives, and Effect Heterogeneity

Motivated by the conceptual framework, this section presents reduced-form evidence about extent and effects of self-selection. To that end, this section describes the Michigan education data, documents substantial self-selection, and presents evidence of effect heterogeneity on observed and unobserved dimensions. Although there is much to learn from them, the RD results suggest that existing estimates insufficiently describe self-selection and cannot inform welfare comparisons between recommendations and requirements.

3.1 Data

This paper uses two data sources. The first is a longitudinal dataset of all K-12 students in Michigan public schools from 2001-02 to 2018-19 provided by the Michigan Education Data Center (MEDC). These records contain enrollment information, demographics, and students' exact dates of birth.[17] To determine the behaviorally relevant fall birthday cutoff, I assign first-time kindergartners to birth cohorts from March 1 of one year to February 28 of the next. Appendix C.1 details the data cleaning and sample selection process which drops 1.1% of first-time kindergartners, leaving me with a sample of 1,283,200 children.

My main outcome is third-grade math scores, the earliest state-wide assessment in Michigan. Note that using this outcome compares students in the same grade who test at different ages (third graders who are nine instead of eight) rather than same-aged students in different grades (nine year olds in third grade instead of fourth grade). I focus on this comparison because the within-grade effects of waiting are most salient to decision makers. For example, academic tracking, high school grades and exit exams, and college applications are all evaluated by grade rather than age.[18]

My second dataset, the Early Childhood Longitudinal Survey-Birth Cohort (ECLS-B), is a national survey following a cohort of children born in 2001 from birth to kindergarten com-

[17] Having exact date of birth is critical for the research question. In addition to the issues discussed in Attar and Cohen-Zada (2018), using month rather than date-of birth simultaneously underestimates noncompliance in the first stage (biasing IV estimates towards zero) while expanding the set of compliers (moving the target parameter in the direction of selection on gains). Thus I cannot use cohorts after 2014 due to the removal of exact date of birth from the data after 2015.

[18] And kindergarten entry decisions do affect these and similar outcomes (Ponzo and Scoppa, 2014; Hemelt and Rosen, 2016; Dhuey et al., 2019; Routon and Walker, 2020).

pletion. These ancillary data complement the Michigan data by providing outcomes observed before kindergarten and information about parental investments and attitudes about readiness. More details on sampling and cleaning can be found in Appendix C.2.

3.2 Descriptive Patterns Reveal Self-Selection and Heterogeneity

The changing kindergarten entry policies in Michigan generate substantial changes in behavior and outcomes. Comparing the children across these RD cutoffs and across years allows me to document the extent of self-selection around recommendations and generates descriptive evidence consistent with substantial heterogeneity in the effects of waiting.

The first stage identifies substantial self-selection. In Michigan over half of families choose to self-select around recommendations at the cutoffs. Figure 3 demonstrates this in context of the RD first stage. This figure depicts the share of children with a given birthday who wait to enter kindergarten in different cohorts. The 2002-2012 cohorts appear in Panel (a), the 2013 cohort in Panel (b), and the 2014 cohort in Panel (c). Recall that these self-selection shares, $p_z(r, x)$, are one of the key sets of parameters for making welfare inferences.

Note three patterns about self-selection in Figure 3. First, the share of children who wait is non-zero and increasing as birthdays approach the cutoff from the left. These children who opt into waiting despite a recommendation to start are always takers, and they make up a non-trivial share of all children (up to 19% at December 1). Second, Panels (b) and (c) reveal that most children near the cutoff opt out of waiting when possible. The recommendations at November 1 and October 1 in Panels (b) and (c) affect between 10-25% of families (eager compliers), and the requirement at December 1 affects 35-50% (reluctant compliers). This large share of reluctant compliers contrasts with kindergarten entry research showing that opting out of waiting (early entry) is fairly rare empirically.[19] The low equilibrium prevalence of opting out, however, is a function of state polices *requiring* waiting after birthday cutoffs. Directly estimating the share of reluctant compliers yields different results with first-order implications for welfare and policy. Third, because the combined share of always takers and reluctant compliers is large (50-80% of children near the cutoffs), recommendations and requirements will have substantially different policy implications.

Self-selection also varies by family characteristics. Consistent with other research, I find that opting in to waiting is much more prevalent for families with higher incomes, white families, and for families with boys (e.g., see Bassok and Reardon, 2013; Fortner and Jenkins, 2017). Opting out is somewhat more uniform across groups, with lower-income families and girls opting out slightly more (see Panel A of Appendix Table A.2 for subgroup-specific self-selection shares at December 1). These differences in self-selection shares underpin the policy concerns that opting

[19]For example, Bassok and Reardon (2013) estimate that only 2% of children enter early.

out of waiting lowers average achievement and that opting in reduces achievement equity. The differences also satisfy one of the two conditions that will bias welfare estimates based off of additively separable effects.

The reduced form suggests substantial heterogeneity. Whereas descriptive patterns form the first stage inform the relative shares of different types of children, changes in test scores contain two pieces of information about gains. Panel (a) of Figure 4 presents the reduced form of the RD: average third-grade math scores plotted over birthdays. Figure 4 presents these patterns for the 2002-2012 cohorts with analogous results for 2013 and 2014 in Appendix Figures A.1 and A.2. The reduced form depicts a large increase third grade test scores at the cutoff (about 0.2 SD). This should not be surprising given the strong first stage and repeated evidence from other research that compliers experience achievement gains from waiting. This traditional reduced form relationship, however, does not directly inform the questions of self-selection or heterogeneous gains to waiting.

To explore the connection between self-selection and achievement, Panel (b) depicts the average test scores plotted over birthdays *separately* by whether children waited to enter kindergarten. I call this relationship the "conditional reduced form." This depiction of $\mathbb{E}[y|w, r]$ is proposed in Bertanha and Imbens (2019) to consider effect heterogeneity.[20] They note that under the null of constant effects and no selection bias, we would expect the average test scores of always takers (to the left of the cutoff) to be equal to the average scores of all children who wait (to the right). As such, the discontinuities evident in Panel (b) of Figures 4, A.1, and A.2 reveal economically meaningful differences by self-selection.[21] If always takers opt into waiting because they benefit more from waiting, then a LATE based on compliers will be uninformative about the benefits of recommendations over requirements. Although these patterns suggest selection on gains, they could also arise positive selection in levels into waiting. Directly testing the extent of heterogeneity requires directly estimating and comparing effects across cutoffs.

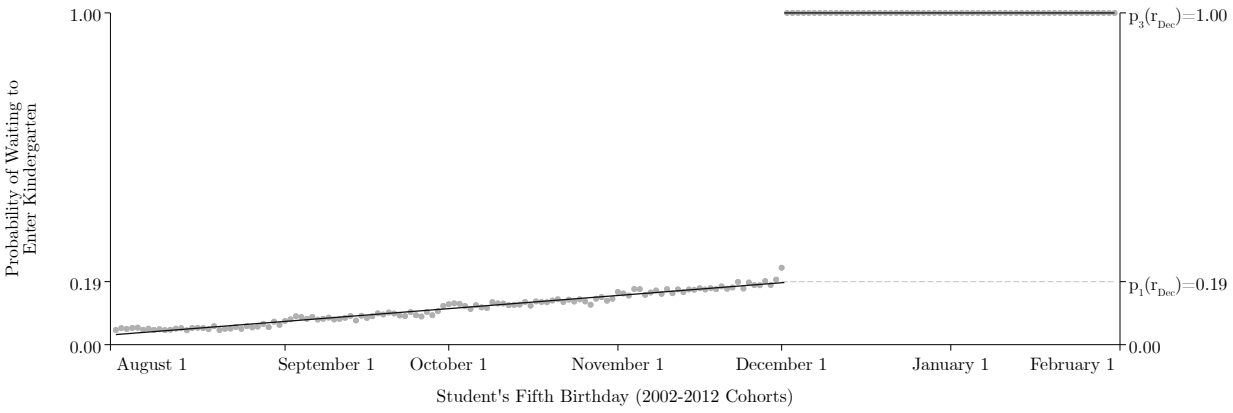
3.3 RD Estimates Suggest Heterogeneity in the Effects of Waiting

The reduced-form and first-stage relationships highlight the first-order importance of self-selection in the kindergarten entry process. To understand self-selection better, I also consider the effects of waiting and whether they vary between subgroups and between always takers, eager compliers, and reluctant compliers.

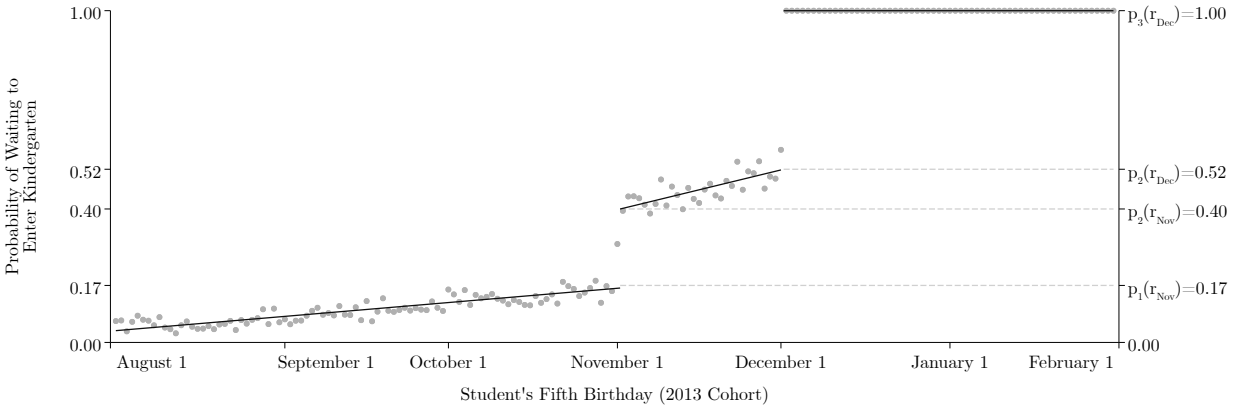
[20] Another interpretation of the conditional reduced form is a visual depiction of a control function estimator (see Wooldridge, 2015; Kline and Walters, 2019).

[21] The level difference between always takers and peers with same birthday who started kindergarten at age five is also consistent with evidence that academic redshirters tend to be higher achieving than other groups (Fortner and Jenkins, 2017).

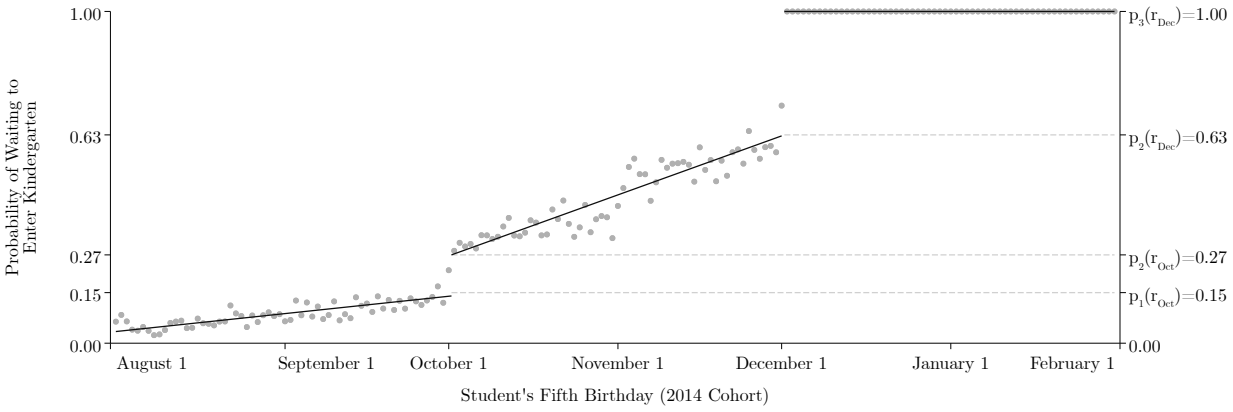
Figure 3: Multiple Cutoffs Generate Substantial Self-Selection



(a) 2002-2012 Cohorts: Only December Cutoff



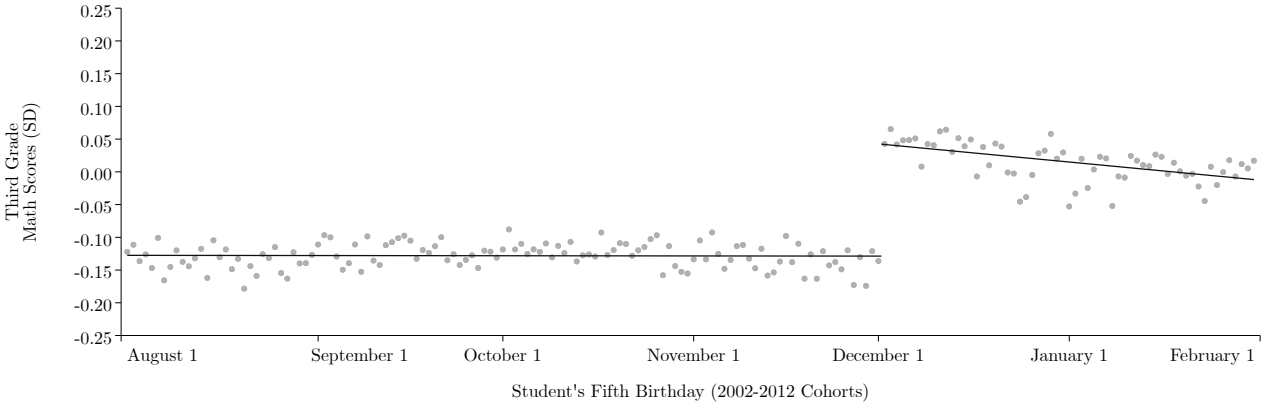
(b) 2013 Cohort: November and December Cutoffs



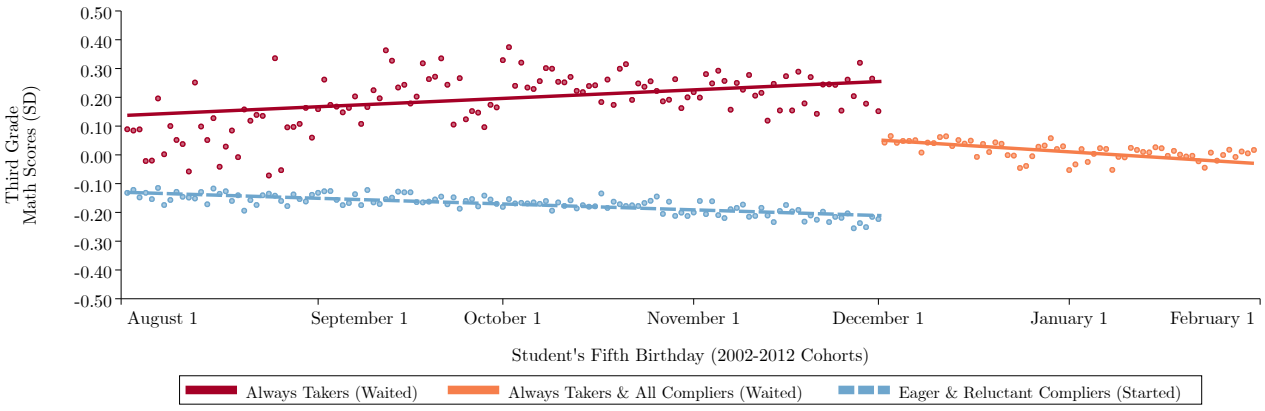
(c) 2014 Cohort: October and December Cutoffs

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 as 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 kindergartners who turned five between September 1 and December 31 2002-2014 and for whom I observe third grade test scores.

Figure 4: Self-Selection Drives Stark Patterns in Third Grade Test Scores



(a) Reduced Form



(b) Conditional Reduced Form

Note: This figure displays patterns in third-grade math achievement for students on facing different birthday-based recommendations or requirements during the original policy from 2002-2012. The third-grade test scores are measured in birth-cohort standard deviations and are plotted over student birthdays. Scores are reported separately by birthday in (Panel (a)) and by birthday and entry decision (Panel (b)) with accompanying lines of best fit. The sample is comprised of first-time kindergartners who turned five between September 1 and January 31 of 2002-2013 and for whom I observe third grade test scores.

3.3.1 Estimation and Identification

I estimate the causal effects of waiting to enter kindergarten at different birthday cutoffs and for different subgroups. Recall w_i represents child i's decision to either start kindergarten at age five ($w_i = 0$) or wait ($w_i = 1$), r_i represents the child's birthday, and y_i realized third-grade math scores. A policy vector $\boldsymbol{z}_i$ with elements $z_{i,k} \in 0, 1$ indicates whether the child's birthday is after a cutoff of a given intensity $k = 1, 2, 3$. Using birthday cutoffs as instrumental variables, these cutoffs identify three separate effects: τ_{ec}, the effect on eager compliers (using the recommendation at October 1, 2014 and November 1, 2013); τ_{rc}, the effect on reluctant compliers (using the requirement at December 1 in 2013-14); and τ_{all} the effect on both eager and reluctant compliers (using the requirement at December 1 in 2002-2012).

I estimate the following local instrumental variables regressions for $k = 1, 2, 3$

$$y_i = \tau_k w_i + \beta_k^y x_{ik} + v_{ik}$$
$$w_i = \pi_k z_{ik} + \beta_k^w x_{ik} + u_{ik}$$

Here $x_{i,k} = (r_i, r_i \cdot z_k)$ (with each cutoff, r_k^* normalized to zero). For the cohorts in 2013 and 2014, $x_{i,k} = (r_i, r_i \cdot z_k)$ includes an additional interaction to allow the slopes to possibly vary by year as seen in Figure 3. I use a rectangular kernel with a 60-day bandwidth around each cutoff.[22] Appendix Tables A.4, A.5, and A.6 show that estimates are not particularly sensitive to bandwidth, weighting, functional form, covariate inclusion, or sample criteria.

The identification of $\hat{\tau}$ is based off of the independence and exclusion assumptions in Section 2.4. Under this assumption, population potential outcomes are continuous at each cutoff.[23] This assumption will be violated if sample composition or other determinants of test scores (besides the policy recommendations and requirements) change at the birthday cutoffs. The birthday cutoff instrument has repeatedly been shown to be valid empirically, but for completeness I explore two important sets of potential violations in Appendix D.1. Birthdays are continuously distributed, family characteristics change very slightly, and dropping observations with possible exclusion-violating investments does not change any results.

Once identified, interpreting each $\hat{\tau}$ as a complier-specific LATE requires monotonicity. This assumption (stated formally in the context of the model in Appendix B.1) can be interpreted as a "no defiers" condition that no children would wait if recommended or required to start who would also start if recommended or required to wait. Under this condition each $\hat{\tau}$ reflects an average of the marginal treatment effects for the relevant compliers.

[22]The 60-day bandwidth has one exception: in 2013 because the November 1 and December 1 cutoffs are only 30 days apart, I restrict the bandwidth to only include observations between the cutoffs.

[23]Note that whereas assuming $\mathbb{E}[y|w, r]$ are continuous was an external validity argument (Bertanha and Imbens, 2019), assuming $\mathbb{E}[y_w|r]$ are continuous is an exclusion argument (see comparison to an RCT in Black et al., 2022).

Table 1: Gains from Waiting to Enter Kindergarten Are Large But Heterogeneous

Panel A: Effects by Cutoff	All Compliers (τ_{all})	Eager Compliers (τ_{ec})	Reluctant Compliers (τ_{rc}))
Effect of Waiting	0.229***	0.374***	0.172***
	(0.010)	(0.110)	(0.049)
Reduced Form	0.185***	0.059***	0.069***
	(0.008)	(0.018)	(0.020)
First Stage	0.808***	0.158***	0.402***
	(0.004)	(0.008)	(0.010)
Cutoff Month	December	October/November	December
Cutoff Years	2002-2012	2013-2014	2013-2014
Cutoff Type	Reccom. Start → Require Wait	Reccom. Start → Reccom. Wait	Reccom. Wait → Require Wait
Children within 60 days	342,642	54,801†	50,874†

Note: This table reports the effects of waiting to enter kindergarten on third-grade math test scores (measured in standard deviations) under different policies. All regression discontinuity estimates are weighted liner regressions with 60-day bandwidth around each cutoff and rectangular kernel. Standard errors allow for arbitrary variance-covariance structure within schools, but two-way clustering by birthday further shrinks standard errors. The sample is comprised of children within the bandwidth who started kindergarten in Michigan public schools in the 2002-03 to 2015-16 school years and took state math exams in third grade.
† In 2013 only children with birthdays between November 1 and December 1 are included in these subsamples so that no samples span multiple cutoffs.
$^{***}p < 0.001$.

3.3.2 Main Effects

Table 1 quantifies the behavioral changes and later effects for each cutoff policy. In the early years, the strong December cutoff induced 81% of children to wait and improved their third-grade math scores by 0.23 standard deviations (τ_{all}). On average, the recommendations at October 1 and November 1 affected 16% of children, improving their scores by 0.37 standard deviations (τ_{ec}). In these years the December 1 requirement only affected 40% of children, with a 0.17 standard deviation effect (τ_{rc}). Table 1 displays each effect with reduced-form and first-stage regressions for reference.

Consider two main messages from Table 1. First, none of these effects are small. A 0.23 standard deviation effect is comparable to replacing a third-grade teacher at the 16th percentile of math value-added with one at the 84th (see Chetty et al., 2014a). This effect aligns well

with achievement results from previous studies (e.g., Datar, 2006; Dhuey et al., 2019),[24] which should be expected as the identifying variation for the LATE for all compliers is most similar to other state policies.

Second, τ_{rc} is the first characterization (to my knowledge) showing how waiting affects "early entrants." As shown in Section 2.5, this is of first-order welfare relevance. Requiring waiting when children would rather opt out does raise scores, but the benefits seem to be smaller for reluctant compliers than to other children.[25] This pattern again suggests positive selection on gains, where children benefit less from waiting are more likely to opt out.[26]

I also find clear evidence of heterogeneity by child and school characteristics. Using the sample of children from 2002-2012, I find statistically appreciable differences in the effects of waiting on boys and girls, on children in schools with higher versus lower poverty rates, and at the intersection of school and family poverty (see Appendix Table A.2 reports results showing). The fact that effects vary based on observable characteristics that are related to self-selection underscores the importance of seriously considering effect heterogeneity in unobserved dimensions as well.

3.3.3 Gains are Correlated with Propensity to Opt Into Waiting

I directly examine the possibility of selection on gains by comparing the effects of waiting among subgroups with different rates of opting into waiting. Evidence that groups who wait more experience different gains from waiting constitutes *prima facie* evidence against effect homogeneity.

To this end, I estimate the effects of waiting across ventiles (and percentiles) of baseline likelihood to opt into waiting. The estimation has four steps. First, among children who turn five before December 1 in the 2002-2012 cohorts I regress waiting on school fixed effects and on child gender, minority status, and poverty status each interacted with ventiles of school poverty (with higher-income white boys omitted in each ventile). Second, I use these estimates to predict each child's *ex ante* probability of waiting. Third, I calculate the empirical ventiles and percentiles of this propensity. Finally, I estimate the effects for each subgroup using the RD estimator from the previous subsection. My preferred specification uses all observations and a quadratic in birthdays (rather than a linear slope with a smaller bandwidth) to increase power without under-fitting non-linearities in the wider bandwidth.

Panel (a) of Figure 5 depicts a striking pattern: Compliers from subgroups who are more

[24]Large effects in early grades typically show diminishing but persistent effects through high school and college (Elder and Lubotsky, 2009; Routon and Walker, 2020).

[25]Testing the null that all three effects are equal yields a p-value of $[p = 0.106]$.

[26]Because children who opt out of waiting tend to be higher achieving than similarly aged peers (Bassok and Reardon, 2013; Fortner and Jenkins, 2017), the fact that reluctant compliers have lower returns to waiting may be because they are already "kindergarten ready."

likely to opt into waiting experience significantly *smaller* gains. The figure depicts a scatter plot of treatment effects over the share of children opting into waiting at the cutoff (for both ventiles and percentiles). A weighted regression of the ventile effects suggests that for every 10 percentage-points increase in a subgroup's probability of opting in, compliers from that subgroup experience 0.02 SD smaller gains from waiting, with null effects at the limit. These results are robust to other specifications and quantile choices and reveal that the traditional LATE estimates cannot describe self-selection nor inform the welfare comparisons because they do not generalize well to groups who are more or less likely to wait.

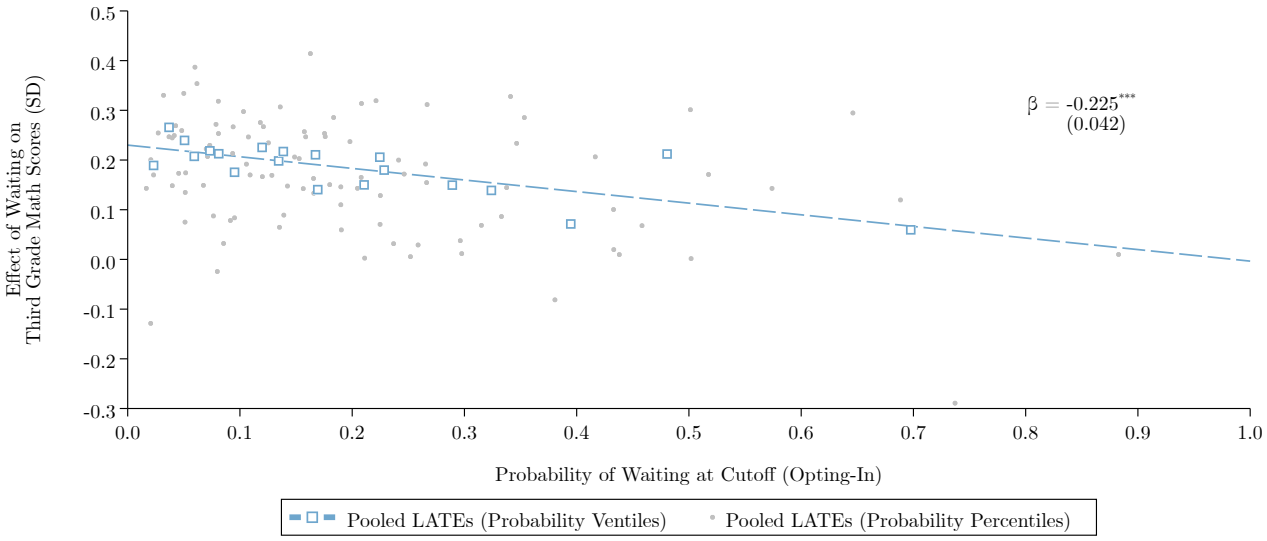
While this exercise demonstrates clear heterogeneity, it does not definitively describe the direction of selection on gains. The direction depends on whether effect heterogeneity loads on observable or unobservable dimensions. On one hand, if gains are based on observable characteristics (for example, related to a family's ability to make additional private investments in a child), then the effects on compliers and always takers from a given subgroup would be similar. In this case the pattern would suggest negative selection on gains because the groups that are most likely to be always takers—identified by the share opting in—experience the smallest gains—as identified on the compliers. On the other hand, if gains are based on unobservables (for example related to a child's specific "readiness" for kindergarten), then the effects on always takers should be much larger than for compliers from the same subgroup. In this case the negative slope would occur because, among subgroups with the most always takers, only children with smaller benefits from waiting would choose not to opt in—generating a compositional difference in compliers across subgroups. This ambiguity about the nature of self-selection echoes the theoretical welfare question about whether the effects of self-selection are additively separable in observed and unobserved components, but it could be resolved with direct evidence about the extent differential self-selection.

3.3.4 Selection on Gains Varies by Family Income

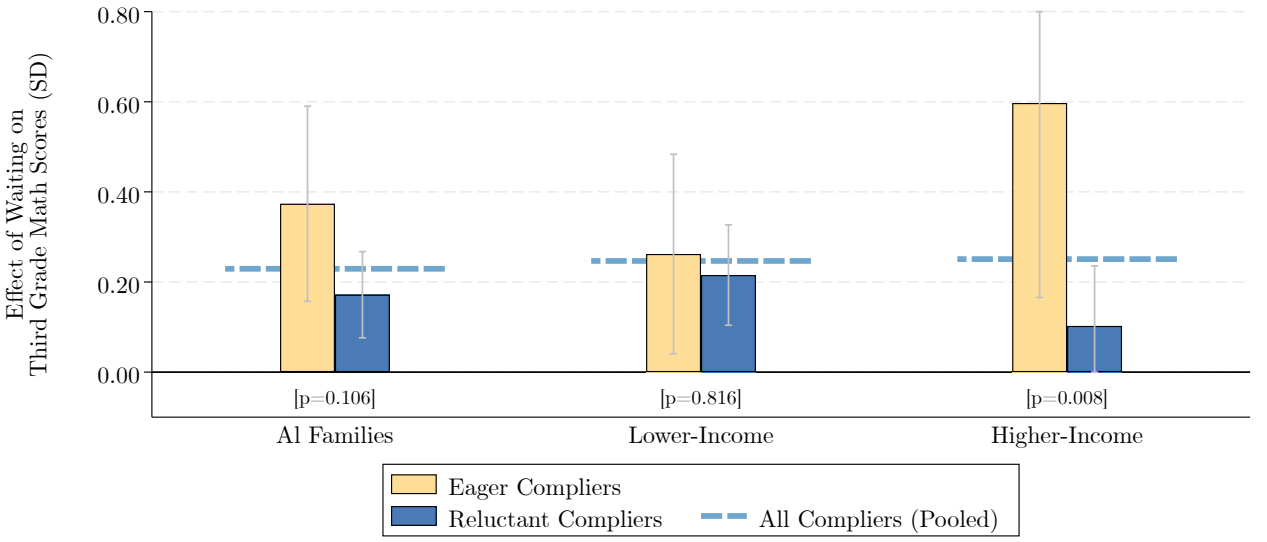
The previous analyses reveal selection on gains but can characterize neither the direction nor whether there is differential selection on gains by family type. Only a direct comparison between children who are more or less willing to wait to start kindergarten conditional on their family characteristics could identify these patterns. While this is not usually possible in an RD, the policy variation from 2013-2014 is uniquely suited to facilitate this type of comparison.

By comparing the pooled LATEs for all compliers identified in 2002-2012 with the effects on eager and reluctant compliers identified by the recommendations and requirements in 2013-2014, I can assess the nature of (differential) selection on gains. To determine whether the effects of waiting differ across different compliers in each subgroup, I estimate the RD effects from each policy in a stacked regression (by subgroup) and test the null that all three effects are

Figure 5: Patterns of Heterogeneity Are Suggestive of Selection on Gains



(a) Heterogeneity by Probability of Waiting



(b) Heterogeneity by Income and Complier Type

Note: These Figures report heterogeneous effect of waiting to enter kindergarten on third-grade math test scores (measured in standard deviations). Panel (a) reports LATEs by ventile (percentile) of propensity to wait, focusing on all children from ten cohorts from 2002-2012. It also reports the slope of the 20 ventile effects (weighted by the inverse cluster-robust standard error of the estimated effects). Panel (b) reports LATEs by family income and complier type in 60-days around each cutoff, leveraging the two cutoffs affecting the 2013 and 2014 cohorts. In Panel (b) estimates for each subgroup are estimated simultaneously in a stacked regression and 95% confidence intervals allow for arbitrary variance-covariance structure within schools but are visually truncated above at 0.8 and below at 0.0 in the figure. Reported p-values are from a test that all three effects are the same for a given group.

equal.[27] Panel (b) of Figure 5 depicts the results for each subgroup with the LATEs for eager and reluctant compliers depicted by bars over a dashed line representing the pooled effect.

Panel (b) illustrates notable heterogeneity—but only among higher-income families. As noted when discussing Table 1, the difference in the whole sample is economically large (around 0.15 standard deviations) but lacks statistical precision. Interestingly, among higher-income families the effects of waiting are larger and very heterogeneous ($p = 0.008$) whereas they are nearly homogeneous among lower-income families.The fact that eager compliers (who follow recommendations to wait) benefit more from waiting than reluctant compliers (who only wait if required) is direct evidence of positive selection on gains. But this positive selection on gains is driven exclusively by higher-income families, direct evidence of differential self-selection.

Taken together, while the RD effects demonstrated substantial self-selection around recommendations that varied by family type. Although the RD identified the effect of forcing reluctant compliers to wait, there is no analogous policy variation for always takers. Furthermore, while I do find heterogeneity at different cutoffs, the locality of RD estimates means observed differences could stem from essential heterogeneity between complier groups or other differences over birthdays or across years. Correctly characterizing the nature of self-selection and its implications for welfare will require additional structure to connect the statistical moments generated by the birthday cutoffs to deeper, policy-relevant economic objects.

4. The Marginal Effects of Waiting to Enter Kindergarten

This section uses the RD variation from Section 3 to estimate the marginal treatment effects of waiting. I first discuss the assumptions required to compare treatment effects identified at different birthday cutoffs. Estimating the marginal treatment effects for the entire sample and for children from higher- and lower-income families separately reveals selection on gains that is driven by higher-income families. I also explore the mechanisms behind these effects, presenting evidence of negative selection in levels and of compensating investments in the year children wait, but no evidence of positional externalities (e.g., being relatively older than peers).

4.1 Selection on Gains and RD Extrapolation

In an RD each LATE is local to certain set of compliers at a certain cutoff. As such, the differences identified in Section 3.3 could theoretically reflect selection on gains (heterogeneity

[27]This test compares the LATE on eager compliers who respond to a recommendation (October 1 or November 1) with the LATE on reluctant compliers required to wait (December 1 in 2013 and 2014) and with the pooled LATE on all compliers (December 1 in 2002-2012). As such the test combines within-year variation over birthdays and year-to-year variation at December 1. In Section 4.2 I show that both comparisons convey extremely similar information.

over U_W) or compositional differences (heterogeneity across years or over r[28]). To address this, I adopt a "parallel trends" assumption (inspired by Cattaneo et al., 2020) that enables me to compare effects across birthdays and a "policy invariance" assumption (from Dong and Lewbel, 2015) to compare effects across years.

Intuitively, these assumptions imply that having a slightly earlier or later birthday or turning five in a different year would not systematically change how waiting affects a child's test scores (see formal details in Appendix D.2).[29] Under these conditions, the marginal treatment effects only vary over characteristics (x) and reluctance to wait (u) implying that $\tau_{MTE}(u, r^*, x) = \tau_{MTE}(u, x)$. This simplification allows formal comparisons between effects on different cutoffs and has the added benefit of enabling extrapolation away from the cutoff, r^*, using p_z and τ_{MTE}. Appendix D.3 compares these assumptions to other approaches in a growing literature about extrapolation and external validity in RD settings.

These assumptions admit two tests to address concerns about their validity. First, the assumptions imply that marginal treatment effects do not vary over birthdays or years. In Section 4.2 a simple over-identification exercise shows that allowing heterogeneity over birthdays or years does not alter the estimates of parameterized MTE curves. Second, although I do not observe whether the *potential* third grade test scores of students are parallel in whether they started kindergarten (and tested in third grade at age eight) or waited (and tested in third grade at age nine), I do observe their *realized* test scores in both third grade (at age eight) and fourth grade (at age nine). In Appendix D.4 I use the longitudinal Michigan data to test whether math scores in third and fourth are parallel over birthdays. If anything, a Rambachan and Roth (2022)-like arguments suggests that any violations of parallel trends lead RD-estimates to slightly underestimate the selection on gains among higher-income families.

4.2 Combining RD Estimates Reveals Positive Selection on Gains

In my conceptual framework the nature of self-selection is reflected in how the effects of waiting vary over different types of families (as characterized by their reluctance to wait U_W). Recall that each birthday cutoff affects families from a specific range of unobservables and that the estimated LATEs are average marginal treatment effects over the relevant support of u. In this light, differences in treatment effects between complier groups contain capture identifying information about τ_{MTE}. Note that because eager compliers at October 1 and November 1 come from different parts of the distribution of u, I separate out these effects rather than combining them as in Section 3. This results in four main estimates: the effect on eager compliers at

[28] Either direct heterogeneity in effects over r or compositional differences in x or γ_W over r.

[29] In the model, the implication is that there are no systematic differences in unobservables along either dimension conditional on the type of child (i.e., $\gamma \perp r, t | x, U_W$). In this light, another way to interpret the policy invariance assumption is that the changing birthday cutoff policies do not alter the distribution of U_W.

October 1 (identified in the 2014 cohort), the effect on eager compliers at November 1 (identified in the 2013 cohort), the effect on reluctant compliers at December 1 (identified in the 2013-2014 cohorts), and the effect on all compliers at December 1 (identified in the 2002-2012 cohorts).

Formally, I combine the information from the four RD estimates by assuming that the MTE curve is linear (as in Brinch et al. (2017) and Kowalski (2023b)). This comparison only uses information from the nonparametrically-identified birthday-cutoff LATEs and does not rely on (often untestable) assumptions about the relationship between the outcomes of compliers and always/never to extrapolate (see Kline and Walters, 2019).[30] Instead, under my parallel trends assumption, effects are each comparable across cutoffs (even though levels are not).

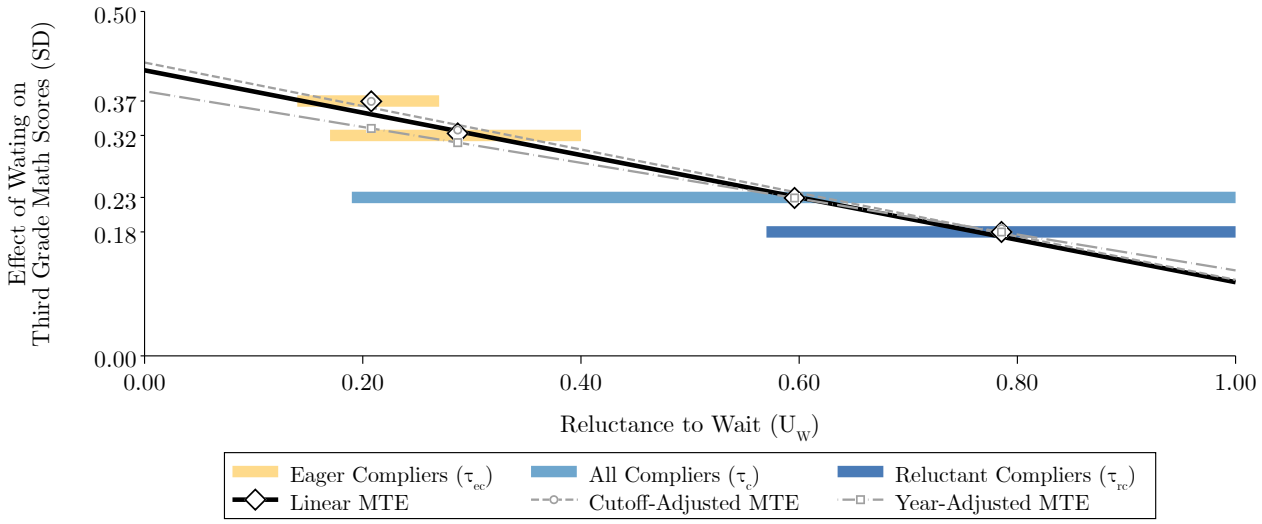
Figure 6 graphically illustrates how the information from the four LATEs identify a linear MTE curve. This figure plots estimated effects of waiting on the y axis over the relevant support of U_W, identified in the first stage (Figure 3). For example, reluctant compliers at December 1 include children with $U_W \in (0.60, 1.00]$, and these children experienced testing gains of 0.18 SD. Figure 6 contains a corresponding line in dark blue, depicting the LATE as an average of the marginal treatment effects over this range. Similar lines represent the average effect on all compliers at December 1 (0.23 over $U_W \in (0.19, 1.00]$ in light blue), eager compliers at November 1 (0.32 over $U_W \in (0.17, 0.40]$ in yellow), and eager complies at October 1 (0.37 over $U_W \in (0.15, 0.27]$ in yellow). Each line is plotted with a corresponding diamond marker at its midpoint.

The MTE curve resulting from the four LATEs confirms the pattern of selection on gains suggested in Section 3. I estimate the linear MTE curve by regressing the four average effects on the average reluctance to wait for each group of compliers (weighting by the inverse of the cluster-robust standard errors). Graphically, this is the line of best fit through the four midpoints (diamond markers). This line is downward sloping, indicating positive selection in gains. Thus in the full population children from the full distribution of U_W tend to experience test-score improvements from waiting, but the gains are about three times larger for the children who are most likely to wait (lowest quintile of U_W) than for the children who are least likely to wait (highest quintile of U_W).

This result completes the picture of economic selection suggested by the comparison between eager and reluctant compliers in Section 3. In the full sample, always takers are also positively selecting on gains into waiting relative to compliers. These students (who are the most likely to wait) experience the largest third-grade math gains from waiting. Interestingly, this result presents a counterpoint to arguments that always takers would have excelled academically even had they started (i.e., that the effects are small), and it suggests that parents tend to opt

[30]Note that using information only from the LATEs is possible because I estimate multiple more than one. In a setting with only one LATE, a linear MTE is only identified by using both the LATE *and* the outcomes of always/never takers to extrapolate (see discussions in Brinch et al. (2017) and Kline and Walters (2019)).

Figure 6: A Linear MTE Fits Well and Shows Positive Selection on Gains



Note: This figure depicts the MTE produced based off of a linear extrapolation from the estimated effects from Section 3.3. Each effect is plotted over the unobserved characteristics of the relevant compliers. The linear MTE that best fits these four points is also depicted as are corrected MTE curves for violations of parallel trends and policy invariance.

children in to waiting who stand to gain more from the investment.

Robustness. I also consider the robustness of this result in four dimensions. First, the MTE results are robust to alternative functional forms. Having four separate LATEs permits over-identification test of the linear specification. I find that the assumed specification is quite robust. An ocular inspection of Figure 6 reveals that the four nonparametrically identified points from the RD results have an almost perfectly linear relationship. This is reflected in the R^2 of the proposed MTE curve ($R^2 = 0.991$). In the over-identification test, the quadratic MTE suggests (if anything) slightly stronger selection on gains but is not appreciably different from the linear model ($p = 0.49$ on the squared term).

Second, the MTE estimates are robust to possible violations of the parallel trends and policy invariance assumptions. Under the linear parameterization of the MTE, I can include controls to demonstrate the validity of each assumption. I can assess the parallel trends assumption by adding fixed effects for the month at which each effect is estimated. This comparison estimates the MTE using only within-birthday variation at December 1 across years. Similarly, I can assess the policy invariance assumption by adding fixed effects for the policy or set of years, estimating the MTE using only within-year variation across birthday cutoffs. Both adjustments are depicted in Figure 6 with nearly indistinguishable results from the unadjusted MTE.

Third, I also find selection on gains using a bounding exercise based on the shape of $m_0(u) =$

$\mathbb{E}[Y_0|u]$. Appendix Figure D.1 visualizes the effect on always takers under the assumption proposed in Mogstad et al. (2018) that m_0 is monotonic. Using the test proposed by (Kowalski, 2023a) among the 2013 and 2014 cohorts, I reject effect homogeneity between always takers and compliers at $[p = 0.000]$ and $[p = 0.002]$ levels (see Appendix D.5 for details). These results independently confirm the economics of self-selection without imposing any parametric form on the MTE. Always takers are indeed positively selected on gains. If anything, this exercise implies that the effects on always takers are larger than the linear MTE suggests; however, I prefer the linear MTE which provides the point estimates necessary to compare recommendations and requirements in policy counterfactuals.

Finally, the entire MTE estimation relies on the assumption that U_W is scalar. Although this modeling decision is not testable, it is important to discuss. The kindergarten entry decision is complicated and involves costs and benefits to both parents and children.[31] Representing all of these as a scalar "reluctance to wait" projects other characteristics into one dimension. For example, two families might opt into waiting, one because they anticipate large academic gains in elementary school and the other because they anticipate large athletic gains in high school, both would have low U_W. Similarly, one family might follow a recommendation because they are actually indifferent between starting and waiting whereas another might follow a recommendation because they don't realize they have the option to opt out and start kindergarten at five. While the strong selection on gains identified in Figure 6 suggests that U_W is certainly relevant on average, these and other scenarios should be kept in mind for concerns about interpretation (if not identification).

4.3 Selection on Gains is Driven by Higher-Income Families

The population-level MTE estimates characterize the average selection on gains, but as shown in Section 2.5, the possibility of differential self-selection is both economically interesting and a prerequisite to making welfare comparisons.

To estimate subgroup-specific MTE curves, I split my samples by family income and re-estimate conditional effects with each of the four birthday cutoffs. To maximize power and common support, MTE estimates typically pool information from all subgroups to identify effects under an assumption of additive separability (see discussion of trade-offs in Cornelissen et al., 2016). Intuitively, these shape restrictions allow the subgroup-specific MTEs to have different intercepts while requiring them to have the same slope. Because Section 2.5 indicates that these assumptions bias welfare analyses when violated and Section 3.3 documents direct evidence against these restrictions, I estimate each subgroup-specific MTE separately.

[31] For example, differences in self-selection could reflect differences in actual gains, in perceptions of future gains, in the value placed on gains, or in the costs of waiting.

Figure 7 displays the resulting MTE curves revealing that higher-income families simultaneously gain more from waiting and select more positively on gains. As in Figure 6, this figure plots the average effects of waiting over the relevant unobservables for the full sample (diamond markers), higher-income families (circle markers), and lower-income families (square markets). There is substantial heterogeneity. Among higher-income families, the average treatment effect of waiting is 0.38 student standard deviations, and the slope of the MTE is negative, implying positive selection on gains. The implied MTE curve for lower-income families (dash-dotted line) reveals smaller, more constant effects. For lower-income families, the implied average treatment effect is 0.24 standard deviations. The slope, however, is essentially zero, implying no selection on gains whether positive or negative.[32] These patterns mean that all of the selection on gains is driven by the higher-income families whose children who are the most likely wait and that higher-income always takers benefit substantially from waiting and from self-selection.

In addition to exploring heterogeneity by family income, I use analogous variation to estimate similar marginal treatment effects by race, gender, and school poverty rates. Appendix Table A.7 reports the results. Qualitatively, it seems there are larger effects and more selection on gains among whites and in lower-poverty schools, but the effects and differences are less precise, suggesting that income is a more central driver of heterogeneous selection. Interestingly, girls also positively select on gains and may benefit more from waiting than boys.[33] If true, these differences in selection on gains may be because boys are more likely to be redshirted in anticipation of high-school athletics, something fairly unrelated to achievement gains.

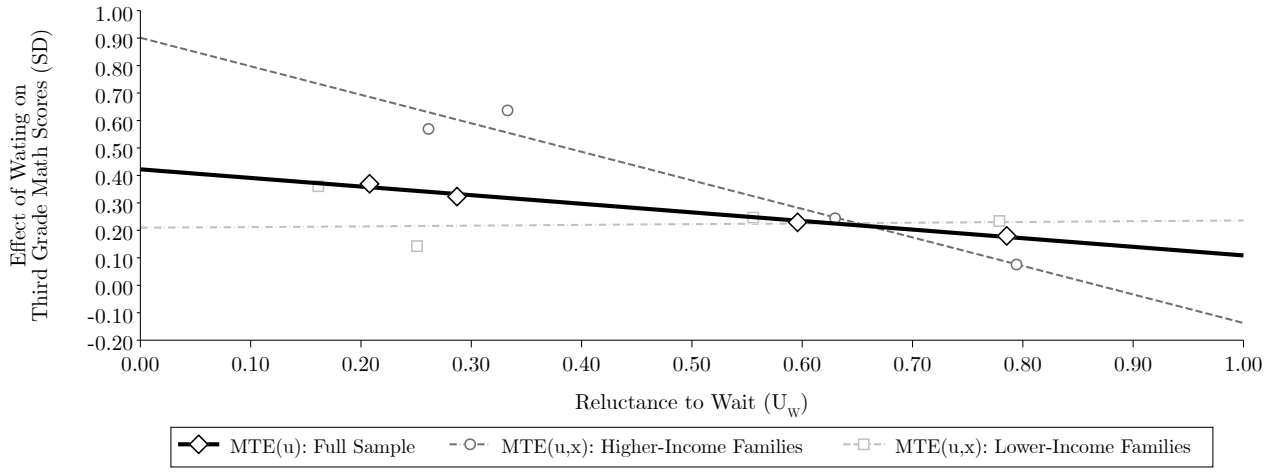
The differences in average effects and selection patterns have three important implications. First, there is substantial heterogeneity in the effects of waiting. While the heterogeneity between children from lower- and higher-income families stands out clearly, there is just as much heterogeneity between always takers and compliers among the higher-income families. Point estimates even suggest that for some children the effect of waiting may be negative. This emphasizes the individual nature of the kindergarten entry decision and the importance of private parental information about children's readiness and potential to gain from waiting. Ignoring this rich heterogeneity is one of the key limitations of any one-size-fits-all requirement. Furthermore, this evidence presents direct evidence of key human capital decisions having "wildly different effects depending on the child" (Oster, 2022).

Second, the differential selection on gains between lower- and higher-income families tells an interesting story about the economics of self-selection. Roy (1951)-style selection on gains is economically intuitive and reflects the core idea of comparative advantage. Families that benefit the most from waiting are most likely to self-select into waiting while those who benefit

[32]Block bootstrapping by school, a two-tailed test rejects the null of no selection on gains—or MTE of zero slope—for higher-income families at [$p = 0.003$] but has [$p = 0.915$] for lower-income families.

[33]The difference in MTE slopes is not statistically significant, however.

Figure 7: Positive Selection on Gains is Driven by Higher-Income Families



Note: This figure depicts the subgroup-specific MTEs produced based off of a linear extrapolation from the estimated effects from Section 3. Each effect is plotted over the unobserved characteristics of the relevant compliers in the whole population and by subgroup. The linear MTE that best fits each set of four points is also depicted. Note that the height of a linear MTE at $U_W = 0.5$ reflects the average treatment effect and that positive selection on gains implies a downward sloping MTE curve.

the least self-select out. It is striking, then, that only higher-income families select on gains. There are economic stories to justify the behavior of lower-income families, but the policy implications depend on whether they are worse at forecasting gains, respond more strongly to costs, or exhibit failures of rationality (mechanisms I will explore in section 4.4. Understanding their self-selection could affect kindergarten entry policy as well as other policy domains where lower-income families benefit less from self-selection like school choice (Abdulkadiroğlu et al., 2018) and health provider selection (Brown et al., 2023).

Finally, the nature of self-selection matters for welfare considerations. Section 2 highlighted how differential selection can have first-order implications for equity and efficiency. Because they allow self-selection, recommendations will raise the scores of children from higher-income families much more than for lower-income families (if at all). Section 5 will directly quantify these gains—gains that traditional methods would miss, underscoring the importance of relaxing such assumptions or of justifying them—whether the justification is econometric (as in Kline and Walters, 2016) or theoretical Brinch et al. (as in 2017)).

4.4 Mechanisms

Before considering welfare and counterfactual policy, considering *why* children are experiencing heterogeneous effects can enhance our understanding of self-selection. From a human capital

perspective, it is important to consider selection in levels. From a policy perspective it is important to understand the role that additional private investments play in mediating effect heterogeneity. Finally, from a welfare perspective, it is important that gains are coming from actual increase in "readiness" for school rather than from zero-sum changes in the rankings of relative age. These considerations may also shed some light on why lower-income families do not select on gains.

4.4.1 Children who Wait are Negatively Selected in Levels

One of the most straightforward ways to understand the economic story underpinning the positive selection on gains is to study the nature of selection in levels. Selection in levels relates to how children with different reluctance to wait would perform academically if they started kindergarten at age five. I formalize this concept with the marginal treatment response (MTR) function of Mogstad et al. (2018): $m_0(u, r, x) = \mathbb{E}[Y_0|R = r, X = x, U_W = u]$. This function represents the average third-grade math scores children would attain if they started kindergarten at five over U_W. Formulating selection based on potential outcomes is important because former research has largely relied on characterizing selection in covariates (e.g., Cameron and Wilson, 1990; Graue and DiPerna, 2000; March, 2005; Bassok and Reardon, 2013; Schanzenbach and Larson, 2017) or realized outcomes (e.g., Fortner and Jenkins, 2017).

Intuitively the idea of selection in levels is related to kindergarten "readiness." On one hand, under positive selection in levels, children who are more reluctant to wait (higher U_W) tend to have lower untreated test scores, implying m_0 is decreasing in u. Positive selection in levels implies that that always takers are more kindergarten ready than eager and reluctant compliers and that advantaged children (who would excel no matter what) can pull even farther ahead by waiting. This narrative is corroborated by descriptive research showing that children who wait tend to have covariates that are associated with higher achievement and by of families opting out of waiting for child-care rather than academic reasons. On the other hand, negative selection in levels means m_0 is increasing in u. Negative selection in levels implies that always takers are less kindergarten ready and that they wait in part because they would perform poorly if they did not. This narrative is reminiscent of parental concerns about readiness and (it turns out) is strongly supported by the data on potential outcomes.

This subsection uses three approaches to assess selection. First, the variation from the birthday cutoffs in 2013 and 2014 allows me to compare the math scores of eager and reluctant compliers who enter kindergarten at the same age. Second, I will compare always takers and compliers using baseline outcomes (before kindergarten) in the ECLS-B data, and third, I return to the Michigan data from 2002-2012 to compare other early-elementary school outcomes.

Eager compliers are negatively selected in levels. I begin by comparing the third-

grade math scores between eager and reluctant compliers when neither wait. To do this, I use the test proposed in Bertanha and Imbens (2019) testing whether the average scores of children who do not wait are continuous around the October 1, 2014 and November 1, 2013.[34] Conditional on waiting, math scores will only be continuous if eager and reluctant compliers are not selecting on levels (see analogous argument in Kowalski (2023b)).

I find large differences in achievement between eager and reluctant compliers who start kindergarten at five. For example, Panel (a) of Figure 8 depicts the average test scores of always takers, eager compliers, and reluctant compliers for children in the 2013 cohort. The 0.42 standard deviation difference between eager and reluctant compliers when neither wait is about the size as the difference in achievement between promoted and retained students in third grade (Jimerson, 2001).[35] Note that these differences are much larger than previous OLS estimates (like Bassok and Reardon, 2013; Fortner and Jenkins, 2017) because the birthday cutoff separates eager and reluctant compliers who are pooled in OLS estimates.[36] Appendix Tables A.8 and A.9 report the full results overall and by subgroup in 2013 and 2014 with generally similar results.

This negative selection on levels suggests a Roy-style self-selection based on comparative advantage. Eager compliers are negatively selected in levels (reluctant compliers score higher when neither group waits) but are positively selected on gains (reluctant compliers experience smaller test-score gains when both groups wait). These patterns of selection are consistent with parental use of private information about school "readiness" to self-select based on comparative advantage. Reluctant compliers opt out of waiting because they will perform well even if they start at five and because they will gain less from waiting. Eager compliers are less prepared, but among higher-income families they benefit much more from waiting and are more likely to do so. As this seems to be true for higher- and lower-income families, it seems unlikely that lower-income families' failure to select on gains is due to a lack of information about potential gains. As such, policies that permit self-selection provide a mechanism by which families who benefit less from waiting can avoid paying the costs of being forced to wait.

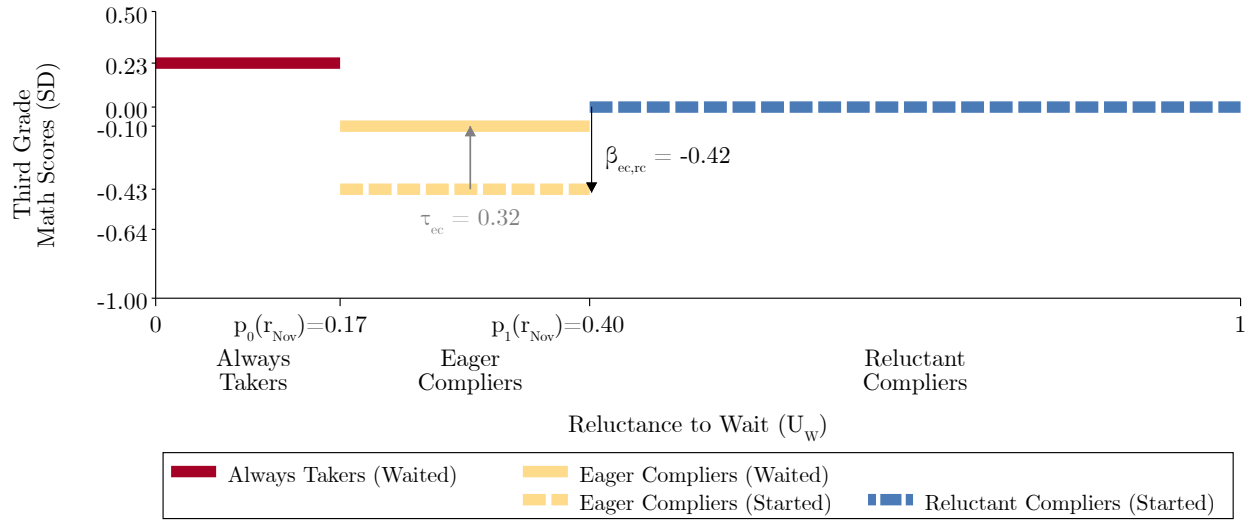
Always takers are also negatively selected. Whereas the RD variation identifies the selection between compliers, it cannot do so for always takers because no policy induces always takers to start kindergarten at age five. One narrative about selection is that always takers must be positively selected because their family covariates are associated with higher academic achievement (wealthy, white, well-educated, etc.). This assertion is at odds, however, with

[34]Under my theoretical framework this test can also be framed as a special case of a more general test proposed in Mogstad et al. (2018). Black et al. (2022) propose an identical test, also noting that rejecting the null "constitutes evidence of either selection or violation of the exclusion restriction."

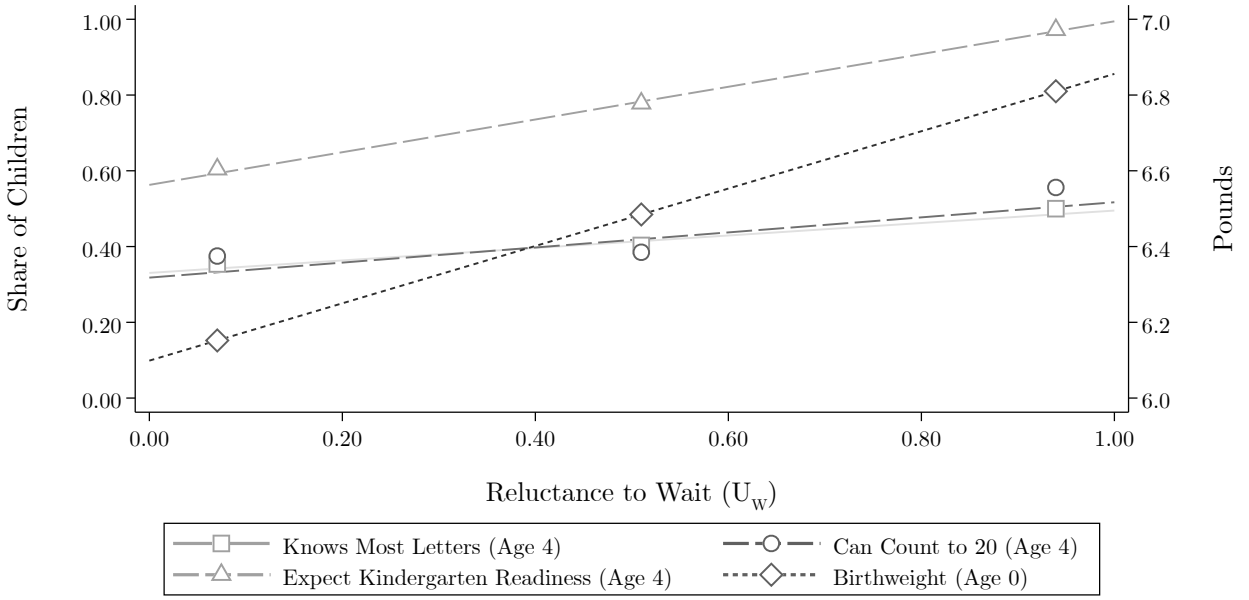
[35]This difference is the size of the discontinuity scaled by the relative shares of eager and reluctant compliers.

[36]OLS compares early entrants (all reluctant compliers after the cutoff) to other students (a mix of reluctant compliers and other children), attenuating the underlying differences.

Figure 8: Always Takers and Eager Compliers Are Negatively Selected



(a) Eager Compliers Score Lower Than Same-Age Reluctant Compliers



(b) Always Takers Are Less Prepared Than Compliers As Young Children

Note: This figure displays evidence related to selection in levels, i.e., differences in potential outcomes had children started kindergarten at age five. Panel (a) depicts average third grade test scores across complier groups and entry decisions among children facing the 2013 kindergarten cohort. The comparison relevant for selection on levels is between the control complier mean of eager and reluctant compliers, but treated outcomes are plotted for reference. Panel (b) depicts average outcomes of children from before the kindergarten entry decision (ages 0-4) across complier groups as realized at age 5 among the children in ECLS-B. This comparison is informative about selection in levels to the extent to which these early childhood outcomes are related to untreated outcomes. SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Survey-Birth Cohort (ECLS-B).

anecdotal and qualitative evidence that entry decisions are shaped by concerns about that kindergarten readiness. Waiting out of concern that children are not ready for kindergarten suggests a story of negative rather than positive selection on baseline and untreated outcomes. I provide two pieces of evidence demonstrating that always takers are in fact negatively selected in levels from baseline assessments at age four and early-elementary-school outcomes.

The first approach uses the Michigan data, focusing on the 2002-2012 cohorts to compare non-testing outcomes of always takers and compliers who wait. These results, displayed in Appendix Table A.10 show that always takers tend to have higher rates of testing accommodation, non-testing, and special education than similarly aged compliers. For these outcomes positive selection in levels can only be rationalized if waiting *increases* the prevalence of these bad outcomes while simultaneously *reducing* them for compliers (as documented in my context and in Elder, 2010; Evans et al., 2010; Layton et al., 2018; Dee and Sievertsen, 2018; Shapiro, 2023). Both of these patterns are exactly what one would expect in the presence of negative selection in levels between always takers and compliers but would be implausible under positive selection.

The second approach sacrifices power to directly observe baseline outcomes using national survey data form the Early Childhood Longitudinal Survey-Birth Cohort (ECLS-B). Panel (b) of Figure 8 compares early-childhood outcomes of future always takers, eager compliers, or reluctant compliers.[37]. I find that birthweight, standardized reading and math assessments from the year a child will turn four, and parental assessments of readiness all seem to increase in U_W. In other words, always takers look negatively selected relative to eager compliers who are negatively selected relative to reluctant compliers in these outcomes. Although, the patterns are slightly under-powered in the small survey sample, Molnar (2024) precisely documents extremely similar patterns between always takers and compliers in Hungarian administrative data.

4.4.2 Children who Wait Experience Compensating Investments

If children who are more likely to wait tend to be negatively selected, what is it about these children that results in the large achievement gains from waiting identified in Figures 6 and 7? To answer this question, I explore what happens to students who wait in the year that they are not in kindergarten using both the ECLS-B data.

The evidence I find suggests that selection on gains is mediated by heterogeneity in investments. Most of these results come from comparing the decisions of higher- and lower-income families and families with always takers and compliers. Figure 9 summarizes the results from the ECLS-B data. I find that in the year they wait children from higher-income families are 5.1

[37]I define these groups in the ECLS-B based on whether students entered kindergarten in 2006 or 2007 and the relevant state birthday cutoffs reported in Bassok and Reardon (2013).

times more likely to attend private preschool, 1.3 times more likely to have a satisfied preschool teacher, 1.9 times more likely to be read to every day, and about 1.5 times more likely to go to a library each month. These differences are consistent with the larger effects of waiting on eager compliers and always takers from higher-income families relative to lower-income families. This may not be surprising given research demonstrating that children from lower-income and less-educated families participate less in available preschool programs (see review in Shapiro et al., 2019).

I also find that *within* higher-income families, always takers may be receiving even more investments. For example, always takers from higher-income families are 31 percentage points more likely to be read to every day in the year they wait than compliers from similar families ([$p = 0.016$]). Among lower-income families, there are smaller, statistically insignificant differences (23 percentage points [$p = 0.190$]).[38] Although the small sample makes finer comparisons more imprecise, differences in other outcomes are consistent with this pattern. Given how critical early childhood investments can be in later-life outcomes (see examples in Duncan et al., 2023), these results reinforce that investments made while waiting are an important factor in the effects of waiting and a contribute to the effect heterogeneity.

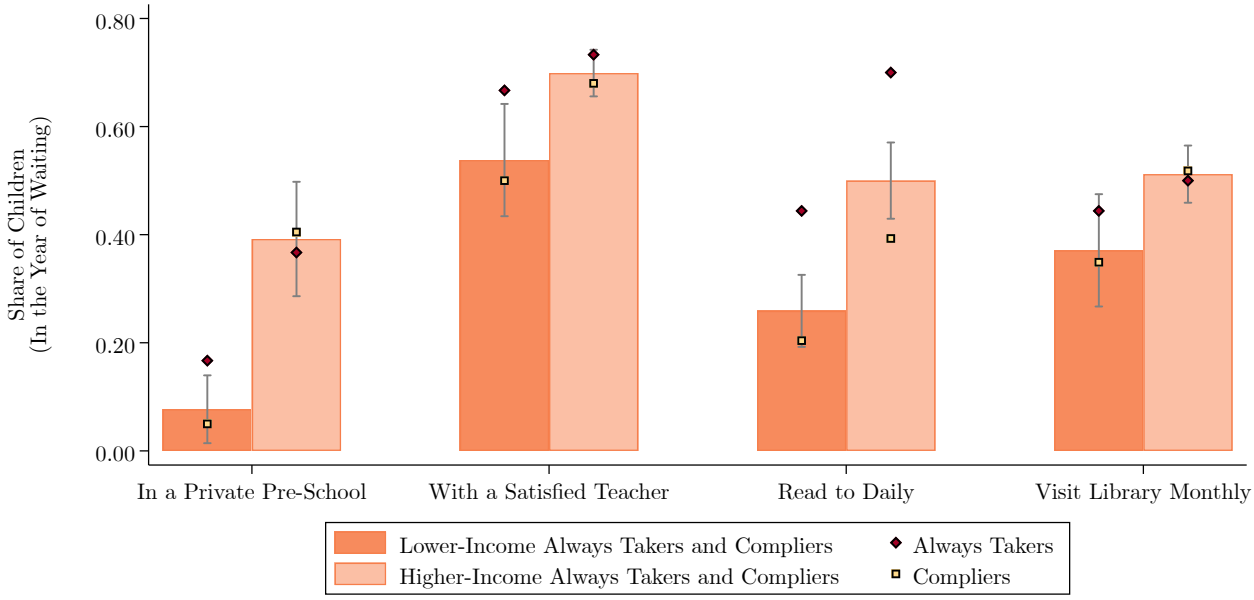
Although children from both higher- and lower-income families tend to be negatively selected in levels, only higher-income families positively select on gains. These results about investment suggest that this difference may be mediated in part by differential abilities to invest. If negatively selected children tend to require more investments, the absence of selection on gains among lower-income families reflects a core feature of the economics of selection. Rather than interpreting the zero selection on gains as a failure of parental rationality, these results suggest that lower-income families may face binding constraints imposed by the costs associated with investing while waiting.

4.4.3 Relative Age Effects and Rent Seeking Do Not Drive Results

One final mechanism to consider is whether the gains from waiting accrue because students who wait are *relatively* older than their peers. The key concern here is that if there is a zero-sum aspect to the gains from waiting, then families' decisions to wait only redistribute achievement rather than increase average scores. In this case, waiting would be costly rent seeking behavior rather than a socially optimal investment. Given the importance of investments in the year children wait, and the absence of relative age effects in other research (e.g., Cascio and Schanzenbach, 2016; Peña, 2017; Molnar, 2024), this concern seems unlikely; however, because such rent seeking behavior complicates welfare comparisons, I also consider it directly here.

[38] See Kalb and Van Ours (2014) for the causal effects of reading with parents on contemporaneous achievement.

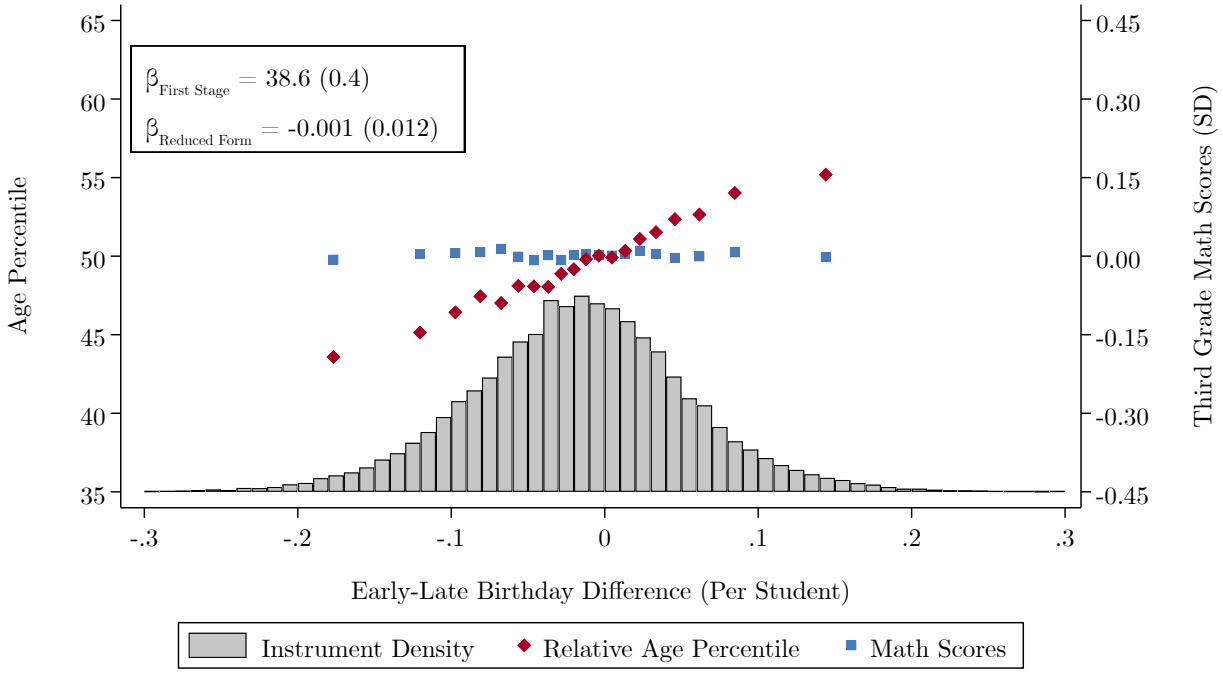
Figure 9: Heterogeneous Gains Associated with Private Investments Made While Waiting



Note: This figure depicts the share of children receiving various investments in the year they wait to enter kindergarten. The sample is split by lower and higher income families and between always takers and compliers. 95% confidence intervals are depicted for overall levels but not by complier subgroups. SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Survey-Birth Cohort (ECLS-B).

To decompose achievement gains coming from absolute and relative age, I leverage within-school variation across cohorts in the share of children's classmates who have birthdays in October-December versus January-March. The identifying intuition is that happening to have peers with birthdays after the cutoff will make eager and reluctant compliers relatively younger in their class age distribution—holding their own birthday and enrollment decision constant. For each student in the 2002-2012 cohorts, I create two leave-one-out measures: the share of peers with birthdays in October or November who started kindergarten at age five with and the share of peers with birthdays in December or January who waited. I then define the peer-birthdays instrument as the difference between these two shares divided by the total number of children. I include school fixed effects in all analyses to control for unobserved demographics possibly correlated with either birth timing or selection around recommendations.

Figure 10: Relative Age Ranks Do Not Drive Gains From Waiting



Note: This figure depicts the age percentile and achievement of children across the distribution of the peer-birthdays instrument. Having peers who happen to be born before birthday cutoffs raises a child's own percentile in the school-cohort age distribution but does not seem to affect academic achievement. The sample come from children facing the birthday cutoffs in Fall 2002-Fall 2012.

Using this cohort-to-cohort variation in peers' ages, I find no evidence of relative age effects. Figure 10 depicts both relative age and third-grade math scores over ventiles of the peer-birthdays instrument (absorbing school fixed effects). While peers' ages are strongly predictive of a child's age percentile, they are completely unrelated to third grade test scores conditional

on the child's own decision to wait. Going from the tenth to the ninetieth percentile of the instrument increases the child's relative age percentile by about 10 percentiles. The corresponding effect on achievement is negative and rules out increases greater than 0.005 student standard deviations. Although the change in relative age from waiting is larger than 10 percentiles, these results suggest that the gains from waiting likely do not come from rent-seeking behavior.[39]

5. Evaluating the Efficiency and Equity of Kindergarten Entry Policies

Combining these estimates of the subgroup-specific MTE curves with subgroup-specific self-selection shares, I can quantify the trade-offs between recommendations, requirements. In this section I also compare other possible kindergarten entry policies like one-sides requirements either prohibiting opting in and prohibiting opting out and changing birthday cutoff dates. Finally, I use the policy comparison to highlight the potential for bias from assuming additive separability or external validity when neither is true.

5.1 Recommendations Versus Requirements

I quantify the efficiency and equity implications of recommendations and requirements by plugging my empirical estimates into Equation 2. Recall that the change in average scores from a recommendation relative to a requirement has two components. The change includes the sum of the effect on all always takers who are allowed to opt into waiting minus the sum of the effects forgone by all reluctant compliers who are allowed to opt out of waiting. I calculate these values for both subgroups to compute the implications for average achievement and achievement equity using the December 1 birthday cutoff for reference.

Empirically, I calculate these objects via simulation. First, I draw students from the empirical joint distribution of birthdays and family poverty status, simulating $U_W \sim [0,1]|x,r$ for each student. Second, I use U_W to identify students who would self-select: $U_W \leq p_1(x,r)$ for always takers and $U_W > p_2(x,r)$ for reluctant compliers. This step requires using a linear extrapolation from the self-selection shares between October 1 and December 1 to identify the share of reluctant compliers between December 1 and February 1.[40] Finally, I compute the relevant effects of the recommendation for always takers ($\tau_{MTE}(u,x)$) and reluctant compliers ($-\tau_{MTE}(u,x)$) to recover the average and subgroup effects, focusing on children with birthdays between August 1 and February 1 where the self-selection occurs.

The results in Table 2 quantify the benefits and costs of recommendations in terms of achievement. Panel A reports the change in average scores among lower- and higher-income

[39] At December 1, waiting moves a complier from the 1st to the 90th percentile of relative age, implying a -0.01 relative-age effect from waiting (and ruling out a 0.045 effect, or 19% of the pooled LATE).

[40] Formally I assume that $\hat{p}(x,r)$ continues the linear slope estimated in the first stage until reaching $\hat{p} = 1.0$.

families. The fact that a recommendation allows families to opt into waiting raises test scores for both types of families, but the gains to higher-income families are seven times larger because they opt in more and show stronger selection on gains. At the same time, a recommendation also allows families to opt out of waiting which lowers test scores for both types of families. In this case, however, the decrease among lower-income families that is twenty times larger because lower-income families are more likely to opt out and because of higher-income families who do opt out tend to be forgoing smaller gains. As a result, whereas a recommendation increases average achievement for children from higher income-families, it decreases average achievement for children from lower-income families.

Table 2: Recommendations Help Higher-Income Families but Harm Lower-Income Families

Panel A: Student Subgroups	Allow Opting In	Allow Opting Out	Recommendation (Allow Both)
Lower-Income Families	0.011	-0.022	-0.012
Higher-Income Families	0.078	-0.001	0.076

Panel B: Policy Effect	Allow Opting In	Allow Opting Out	Recommendation (Allow Both)
Average Scores	0.049	-0.010	0.038
Achievement Gap	0.067	0.021	0.088

Note: This table reports the changes in average and subgroup test scores from a kindergarten entry policy featuring a recommendation relative to a requirement at December 1. Averages for students with birthdays between August 1 and February 1 are computed via simulation using the joint distribution of birthdays and poverty status for 2002-2012 combined with the estimates of self-selection and the marginal effects of waiting.

The combined implication of these results is that a policymaker choosing between a recommendation and a requirement faces an equity-efficiency trade-off. A recommendation is revealed preferred, and Panel B shows that it generates higher average test scores; however, it also causes a deterioration in achievement equity. Average scores increase by about 0.04 student standard deviations—because about 16% of students change their behavior in this window, this implies an effect of +0.24 standard deviations per affected student. On the other hand achievement equity is reduced. Relative to a requirement, the gap in average achievement between higher- and lower-income families is 0.088 standard deviations (15%) larger under a recommendation, reflecting the testing gains among higher-income families and the test score reductions among lower-income families.

In addition to quantifying this trade-off, the results in Table 2 have implications for edu-

cation policy. Regarding policy, the preceding analysis allow me to gain some traction on the long-contested question of whether "kindergarten redshirting" (i.e., opting into waiting) contributes to academic inequities. To address this question, I extend the analyses from Table 2 to the entire sample and compare the achievement gap in the status quo relative to a counterfactual policy without redshirting. The results imply that academic redshirting causes 5.5% of the third-grade income-achievement gap in Michigan. This share is larger than the percent of the gap explainable by teacher value added in North Carolina (e.g., Bates et al., forthcoming, find only small differences by income) but likely smaller than the percent of the gap explainable by school finance in Michigan (e.g., Baron et al., 2024, find that a 10% increase in funding to lower-income schools increases achievement by 12%.). While this estimate is large, it is not implausible given the large causal effects of on high-income always takers. That said, it should be interpreted with caution because despite the evidence of policy invariance, there is always the possibility that general equilibrium responses could complicate comparisons.

The fact that lower-income families are hurt on average by recommendations also highlights a central tension in the nature of self-selection. Whereas the positive selection on gains among higher-income families neatly rationalizes the self-selection observed by this group, the absence of selection on gains among lower-income families reveals the importance of costs in family decision making. If families who are revealed to have higher reluctance to wait experience similar gains, they must either value gains less differently or experience greater costs from waiting. Because research on parental preferences seems to suggest that if anything these families may value achievement gains more (Jacob and Lefgren, 2007), the fact that families choose to forgo these gains highlights the systematic costs associated with waiting. These costs which include paying for preschool, arranging for informal childcare, negotiating transportation and schedules, or reducing employment are particularly salient to lower-income families (see review in Shapiro et al., 2019). These families may not be behaving irrationally; instead they seem to be responding to very tight constraints.

5.2 Choosing Policies and Dates Imply Additional Equity-Efficiency Trade-Offs

Kindergarten entry policies not only choose between recommendations and requirements, but they also set birthday cutoffs. I expand the set of policies I compare to include four policies at cutoffs between August 1 and February 1. The first set of counterfactual policies are strict requirements—such as those in much of Eastern Asia, Western Europe, and Scandinavia. Under these policy children who turn five before the cutoff must start and children who turn five after the cutoff must wait, $(z, z')=(0, 3)$ in the notation of Section 2. The second set are recommendations, a less common policy. Under this policy, children who turn five before the cutoff would be recommended to start and children who turn five after the cutoff would be

recommended to wait, $(z, z')=(1, 2)$. The third set of policies ban opting out of waiting, like the long-standing policy in Michigan and most other US states. Under these policies, children who turn five before the cutoff would be recommended to start (but allowed to opt into waiting) whereas children who turn five after the cutoff must wait, $(z, z')=(1, 3)$. Finally, the last policies ban opting as in many Southern European countries. Under these policies, children who turn five before the cutoff must start whereas children who turn five after the cutoff are recommended to wait, but are allowed to opt out—$(z, z')=(0, 2)$. All effects are calculated by simulation based on the marginal treatment effects and self-selection shares.

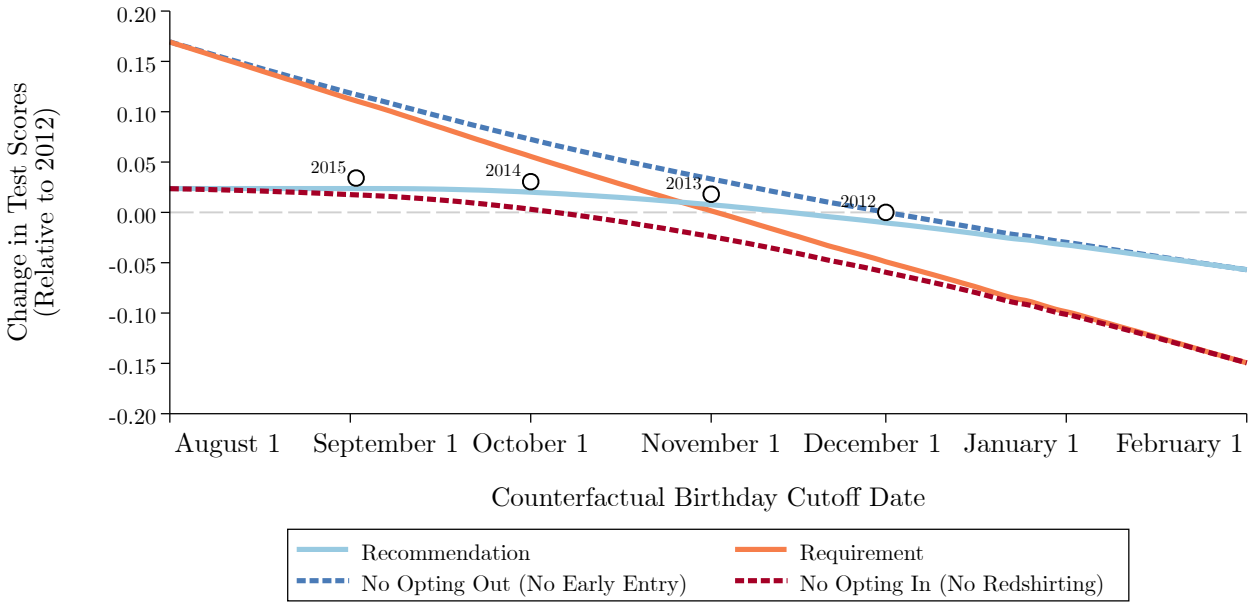
Figure 11 presents the effects of these policies relative to banning opting out after December 1 (the baseline policy in Michigan). Panel (a) presents the change in average scores, and Panel (b) presents the implications for achievement equity with higher values implying wider gaps (more inequity). Within each panel the solid lighter (cyan) line represents the effect of a recommendation based on a given cutoff date and the solid darker (orange) line represents the effect of a requirement. The darker (red) dashed line represents the effects of policies that ban opting and the lighter (blue) dashed line represents the effects of policies that ban opting out. Each panel also presents the implied effects of the actual policies in Michigan which blended a recommendation at a given cutoff with a ban on opting out for children who turned five after December 1.

These simulations identify some interesting patterns and trade-offs about raising average scores. Because waiting tends to raise average test scores, policies that induce more children to wait produce higher achievement. For example, at any cutoff date a policy that bans opting out induces the most waiting and produces the highest test scores, but a policy that bans opting in induces the least waiting and produces the lowest test scores. Similarly, when comparing policies across cutoff dates, later cutoffs induce fewer children to wait, and so all four curves slope downward over cutoff dates. Interestingly, whether recommendations or requirements produce higher scores depends on the cutoff date. For early cutoffs, recommendations produce substantial opting out, lowering scores, but for later cutoffs, recommendations produce substantial opting in, raising scores relative to a requirement.

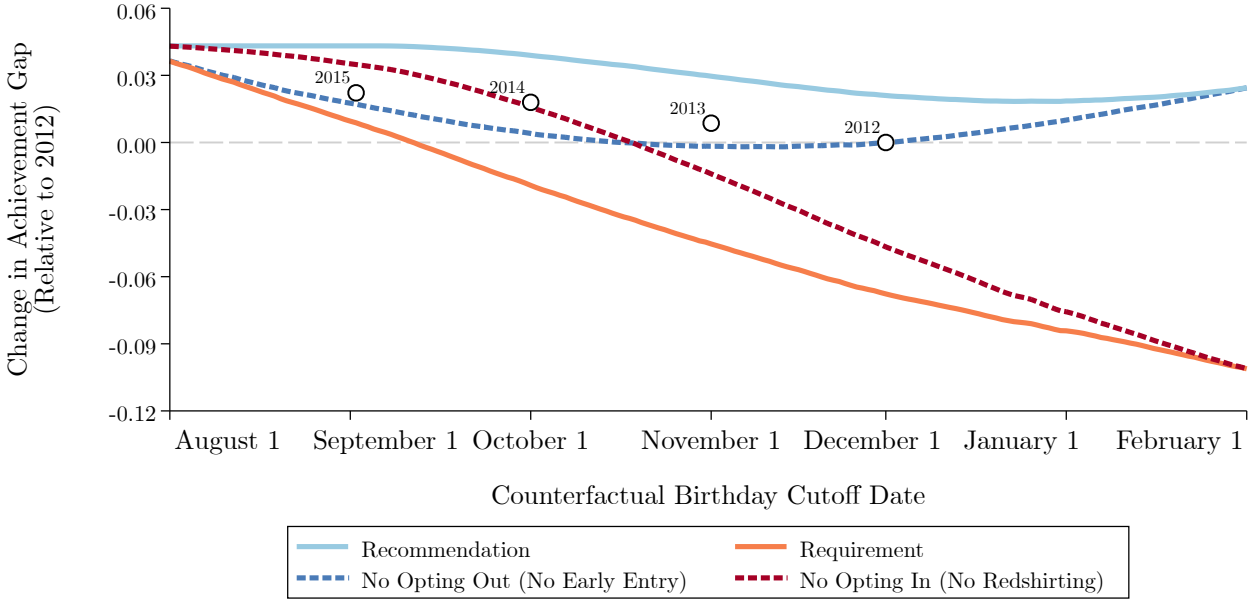
Regarding achievement gaps, these policy counterfactuals demonstrate the intuition for the equity-efficiency tradeoff and how challenging it is to reduce achievement gaps with birthday cutoff policies. Consider two patterns. First, whereas earlier cutoff dates increase test scores, they typically also widen achievement gaps. For example, under a strict requirement, moving the cutoff by from November to August could increase scores by 0.15 standard deviations, but would increase gaps by 0.08. Second, although switching from a recommendation to a requirement would raise average scores at many cutoff dates, this change would also widen the achievement gap—sometimes by as much as 0.12 standard deviations.

Interestingly, the policies implemented by the state of Michigan generated efficiency gains

Figure 11: Different Birthday Cutoff Policies Produce Widely Different Results



(a) Effect on Average Scores



(b) Effect on Achievement Gaps

Note: This figure depicts the change in average test scores and the change in achievement gap that would be implied by various kindergarten entry policies. For each cutoff date that is considered, this Figure depicts the effect of four distinct policies: recommendation, requirement, no opting in, no opting out. All estimates are based off of the subgroup-specific probabilities of waiting and marginal treatment effects reported earlier. Simulations use the distributions of student characteristics and birthdays form the 2002-2012 cohorts.

at lower equity costs than other policies might have. As can be seen by the circle markers in Figure 11, the policies implemented by Michigan raised average third-grade math scores by 0.01-0.03 standard deviations. This policy is more efficient because it was accomplished without violating any families' revealed preference because of the waiver program. The achievement gap between children from higher- and lower-income families expanded by 0.01-0.02 standard deviations. Compared a pure recommendation at these dates (also a revealed preferred policy), the observed policies created larger gains *and* better achievement equity. At the same time the flexibility introduced by the waiver system did result in lower scores and somewhat wider gaps compared to prohibiting opting out after the new cutoff dates.

5.3 Other Empirical Approaches Insufficiently Characterize Welfare

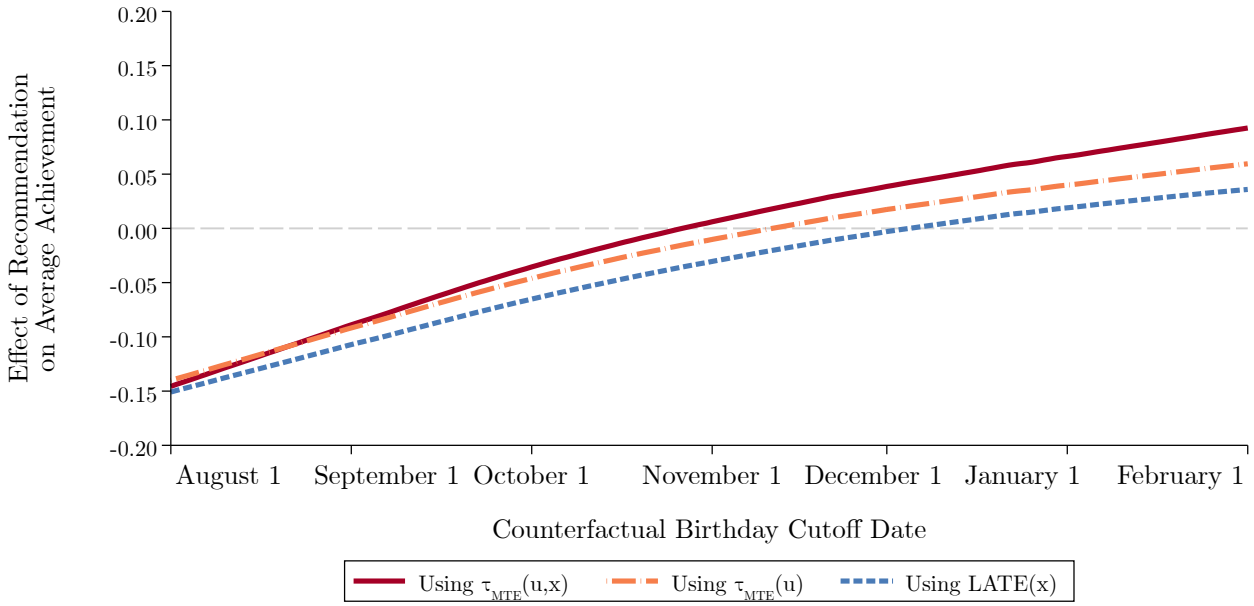
Capturing the intricate patterns underlying this equity-efficiency tradeoff is only possible because of the empirical approach I use. Because Section 2 shows that using standard LATEs or additively separable MTEs will bias welfare estimates, this subsection quantifies the size of those biases. While I find some bias in predictions for average scores, there is much more substantial bias in predicting achievement gaps. This bias arises because alternatives simultaneously understate the benefits of recommendations to higher-income families (relative to requirements) and the harms to lower-income families.

To compare the three empirical approaches that could be used to estimate my policy counterfactuals, I focus on the comparison of recommendations relative to requirements from Equation 2. For each cutoff date I calculate the difference in average achievement attained under a recommendation and requirement. I do this three times, first using my preferred estimates, second using MTE estimates from the same RD effects assuming additive separability (i.e., $\tau_{MTE}(u, x) = \beta_0(x) + \beta_1(x) \cdot u$), and third the LATE from December 1 2002-2012 (which is most similar to LATEs in existing empirical work). All three methods use the same subgroup-specific self-selection shares.

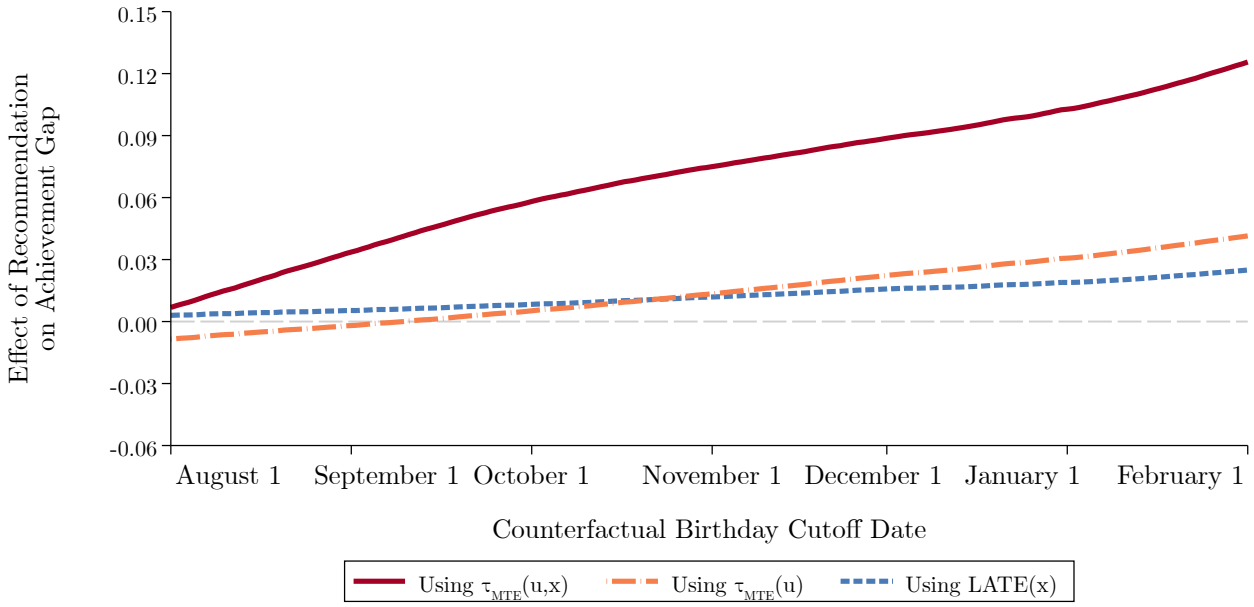
Figure 12 reports the results. Panel (a) reports the effect of having a recommendation relative to a requirement on average test scores, and Panel (b) reports the effect on achievement gaps. Within each panel the solid red line depicts the effects based off of the main estimates in the paper, the dash-dotted orange line depicts the effects based off of the estimates assuming additive separability, and the dashed blue line depicts the effects based off of estimates of subgroup-specific LATEs.

While all three estimators report fairly similar patterns in average test scores, other methods systematically underestimate the benefits of recommendations. All three lines have fairly similar upward slopes, reflecting the intuition that recommendations raise scores more when there are more always takers. Although the estimates capture changes across birthdays well, the alterative

Figure 12: Oversimplified Estimates Miss Key Welfare Trade-Offs



(a) Estimated Effect on Average Scores



(b) Estimated Effect on Achievement Gaps

Note: This figure depicts the efficiency and equity of recommendations versus requirements. Panel (a) shows how a recommendation at a given date would affect average test scores relative to a requirement. Panel (b) shows how a recommendation would affect achievement gaps relative to a requirement. In each panel, three different estimates are presented: a "true" estimate based off of the subgroup-specific marginal treatment effects, one based off of marginal treatment effects assuming additive separability, and one based off of the subgroup-specific Local Average Treatment Effects in 2002-2012. Simulations use the joint distribution of student characteristics and birthdays form the 2002-2012 cohorts.

estimators uniformly understate the gains from requirement (or overstate the harms) for any given birthday. For late cutoff dates the additively separable effects understate gains by 30-40% by ignoring differential selection on gains. The LATE ignores selection on gains altogether, and it underestimates effects by as much as 70%. This highlights the LATE's poor policy relevance in settings with strong selection on gains. Indeed, the fact that the LATE slopes upward at all is due to the self-selection shares identified by the other cutoff policies—information that would typically be unavailable in policy evaluations that would generate analogous LATE estimates.

The poor performance of alternative estimators is more stark in matters of equity. For example, my estimates suggest that, relative to a requirement, using a policy recommendation at December 1 would increase the academic achievement gaps by almost 0.09 standard deviations (15.5%) with smaller but still pronounced impacts all the way to August 1. This pattern is central to the equity efficiency trade-off, but both the LATE and the additively separable MTE remove this trade-off by implying essentially no equity costs. Both estimates suggest that recommendations and requirements have essentially identical implications for achievement equity. By ignoring the differential selection on gains both estimates simultaneously understate the benefits of recommendations to higher-income families (relative to requirements) and the harms to lower-income families. Combined, these inconsistencies lead to misleading efficiency implications and almost entirely sweep equity implications from consideration.

Taking a step back, these results underpin two main economic messages that may apply in many other contexts. First, in settings where there may be rich heterogeneity in the effects of an economic "treatment," the patters of selection in to (and out of) that treatment have first order welfare implications—both in terms of efficiency and in terms of equity. Second, the way that researchers measure that relevant heterogeneity in effects will directly affect the results of those welfare comparisons. In the kindergarten context, methods that assume away differential selection get the wrong answers about both efficiency and equity questions, and they erase trade-offs between the two. These patterns will likely matter for treatments as varied as publicly provided health or education services, financial decisions like investments or insurance, and even governmental regulation and criminal justice.

6. Conclusion

This paper proposed two main questions: (1) How does self-selection influence kindergarten entry decisions? and (2) What are the implications of self-selection for the efficiency and equity of different policies? By answering these questions this paper characterizes the economics of a ubiquitous human capital decision, identifies a key equity-efficiency trade-off underpinning kindergarten entry policies, and informs broader policy conversations about self-selection in other settings. To conclude, I discuss how my results and future research could inform parental

decision making, education policy, and program design in other policy contexts.

In the kindergarten context, self-selection is characterized by negative selection in levels and positive selection on gains. In other words, children who are more likely to wait tend to perform worse when they do not wait and tend to experience larger score gains. This pattern is true both for children who self-select into waiting (gaining the most from waiting) and who self-select out (gaining the least). These patterns are consistent with families weighing the financial and opportunity costs of waiting against the potential benefits for their children then waiting when children would benefit the most. The differential selection on gains by income suggests that higher-income families respond more to anticipated gains whereas lower-income families are more responsive to the costs of waiting.

This self-selection suggests three take-aways for families making kindergarten entry decisions. First, the results on negative selection in levels suggest that parents have private information, and if a parent is considering opting into waiting, it tends to be with good cause. This is true on average and within most subgroups. Second, gains seem to be mediated by childhood investments rather than purely by age. As such, parents may want to jointly consider both kindergarten readiness and potential care arrangements when deciding whether to wait. Finally, parents should remember that although waiting produces gains on average, there are cases where waiting may cause harm. For example, waiting may prevent early interventions with actual learning disabilities.[41] Similarly, the interaction of waiting with age-based compulsory schooling laws can reduce high school completion rates for some compliers. Because it is unknown whether this is also true for families who self-select into waiting, assessing the long-term effects on self-selected children is a promising avenue for future work.

Regarding education policy, self-selection creates a binding equity-efficiency tradeoff on two main dimensions. First, given a birthday cutoff, policy makers can choose what recommendations or requirements to make regarding kindergarten entry. Although recommendations are preferred by families and raise third-grade math scores on average, they also reduce scores for lower-income families. Similarly, given a policy recommendation (or requirement), earlier birthday cutoffs raise average scores by inducing more waiting; however, they also tend to expand achievement gaps. As more and more jurisdictions revisit (often decades-old) kindergarten entry policies, the equity-efficiency trade-offs I identify can quantitatively inform local policy making. Moreover, for policy makers considering early childhood more broadly, it may be the case that policy interventions providing opportunities for more investments (like district prekindergarten programs) may be much more effective in promoting both equity and efficiency than policies aimed at improving self-selection.[42] Given the possible role of such programs in

[41]For example, Fortner and Jenkins (2018) find that among students participating in special education, children who opt into waiting actually have the lowest average scores.

[42]This would be consistent with my results on mechanisms and selection on the costs of investments as found in other work (Kline and Walters, 2016; Cornelissen et al., 2018; Felfe and Lalive, 2018; Shapiro et al., 2019).

Michigan, future work replicating my findings in new policy environments could be of value.

Finally, I conclude by considering a few ways in which my results might inform conversations about other policy settings with self-selection. First, as with contexts as varied as college enrollment (Carneiro et al., 2011), adverse selection into insurance (Hackmann et al., 2015), and opting into voluntary regulation (Einav et al., 2022), I find appreciable selection on gains. This self-selection reinforces the economic intuition that uniform policy requirements often impose real efficiency costs. Considering these costs is likely (perhaps especially) important in cases where equity is the explicit motivation for requirements as varied as eliminating high-school honors classes (Randazzo, 2023), designing mandatory regulations (Einav et al., 2022; Ito et al., 2021), or implementing universal basic income program (Ghatak and Maniquet, 2019). Similarly, given evidence that disadvantaged individuals often benefit the least from self-selection (e.g., Kline and Walters, 2016; Abdulkadiroğlu et al., 2018; Chen and Pereyra, 2019; Brown et al., 2023), my results on differential self-selection suggest that using methods that allow researchers to consider such heterogeneity may be valuable for optimal policy considerations. Future research studying heterogeneous self-selection into other public programs may even illuminate ways whereby programs could promote both greater efficiency and equity.

References

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

Abdulkadiroğlu, A., P. A. Pathak, and C. R. Walters (2018): "Free to choose: Can school choice reduce student achievement?" American Economic Journal: Applied Economics, 10, 175–206.

Allcott, H., J. Kim, D. Taubinsky, and J. Zinman (2022): "Are high-interest loans predatory? theory and evidence from payday lending," The Review of Economic Studies, 89, 1041–1084.

Altmejd, A., A. Barrios-Fernández, M. Drlje, J. Goodman, M. Hurwitz, D. Kovac, C. Mulhern, C. Neilson, and J. Smith (2021): "O brother, where start thou? Sibling spillovers on college and major choice in four countries," The Quarterly Journal of Economics, 136, 1831–1886.

Angelova, V., W. Dobbie, and C. S. Yang (2023): "Algorithmic Recommendations and Human Discretion," NBER Working Paper.

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.

Asher, S. and P. Novosad (2020): "Rural roads and local economic development," American Economic Review, 110, 797–823.

Athey, S. and S. Wager (2021): "Policy learning with observational data," Econometrica, 89, 133–161.

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.

Baron, E. J., J. Hyman, and B. Vasquez (2024): "Public school funding, school quality, and adult crime," Review of Economics and Statistics, 1–46.

Barrios-Fernández, A., C. A. Neilson, and S. D. Zimmerman (2024): "Elite universities and the intergenerational transmission of human and social capital," Iza discussion paper.

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.

Bates, M., M. Dinerstein, A. C. Johnston, and I. Sorkin (forthcoming): "Teacher labor market policy and the theory of the second best," Quarterly Journal of Economics.

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.

Berne, J., B. Jacob, T. Musaddiq, A. Shapiro, and C. Weiland (2024): "The Effect of Early Childhood Programs on Third-Grade Test Scores: Evidence from Transitional Kindergarten in Michigan," in AEA Papers and Proceedings, American Economic Association 2014 Broadway, Suite 305, Nashville, TN 37203, vol. 114, 480–485.

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

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.

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.

Brown, Z. Y., C. Hansman, J. Keener, and A. Veiga (2023): "Information and disparities in health care quality: Evidence from GP choice in England," NBER Working Paper.

Cameron, M. B. and B. J. Wilson (1990): "The effects of chronological age, gender, and delay of entry on academic achievement and retention: Implications for academic redshirting," Psychology in the Schools, 27, 260–263.

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.

Chen, L. and J. S. Pereyra (2019): "Self-selection in school choice," Games and Economic Behavior, 117, 59–81.

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

——— (2014b): "Measuring the impacts of teachers II: Teacher value-added and student outcomes in adulthood," American Economic Review, 104, 2633–2679.

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.

Datar, A. (2006): "Does delaying kindergarten entrance give children a head start?" Economics of Education Review, 25, 43–62.

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.

Depew, B. and O. Eren (2016): "Born on the wrong day? School entry age and juvenile crime," Journal of Urban Economics, 96, 73–90.

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.

Dhuey, E. and S. Lipscomb (2010): "Disabled or young? Relative age and special education diagnoses in schools," Economics of Education Review, 29, 857–872.

Dobkin, C. and F. Ferreira (2010): "Do school entry laws affect educational attainment and labor market outcomes?" Economics of Education Review, 29, 40–54.

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.

Duncan, G., A. Kalil, M. Mogstad, and M. Rege (2023): "Investing in early childhood development in preschool and at home," Handbook of the Economics of Education, 6, 1–91.

Einav, L., A. Finkelstein, Y. Ji, and N. Mahoney (2022): "Voluntary regulation: Evidence from Medicare payment reform," The Quarterly Journal of Economics, 137, 565–618.

Einav, L., A. Finkelstein, T. Oostrom, A. Ostriker, and H. Williams (2020): "Screening and selection: The case of mammograms," American Economic Review, 110, 3836–3870.

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.

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.

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.

——— (2018): "Is delayed school entry harmful for children with disabilities?" Early Childhood Research Quarterly, 44, 170–180.

Ghatak, M. and F. Maniquet (2019): "Universal basic income: Some theoretical aspects," Annual Review of Economics, 11, 895–928.

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.

Grunwald, B. (2015): "Questioning Blackmun's Thesis: Does uniformity in sentencing entail unfairness?" Law & Society Review, 49, 499–534.

Hackmann, M. B., J. T. Kolstad, and A. E. Kowalski (2015): "Adverse selection and an individual mandate: When theory meets practice," American Economic Review, 105, 1030–1066.

Heckman, J. J., S. Urzua, and E. Vytlacil (2006): "Understanding instrumental variables in models with essential heterogeneity," The Review of Economics and Statistics, 88, 389–432.

Heckman, J. J. and E. Vytlacil (2005): "Structural equations, treatment effects, and econometric policy evaluation," Econometrica, 73, 669–738.

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.

Ida, T., T. Ishihara, K. Ito, D. Kido, T. Kitagawa, S. Sakaguchi, and S. Sasaki (2022): "Choosing who chooses: Selection-driven targeting in energy rebate programs," NBER Working Paper.

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.

International Association for the Evaluation of Educational Achievement (IEA) (2021): "National Policies on Age of School Entry and Promotion," Tech. rep., Progress in International Reading Literacy Study.

Ito, K., T. Ida, and M. Tanaka (2021): "Selection on welfare gains: Experimental evidence from electricity plan choice," NBER Working Paper.

Jacob, B. A. and L. Lefgren (2007): "What do parents value in education? An empirical investigation of parents' revealed preferences for teachers," The Quarterly Journal of Economics, 122, 1603–1637.

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, M. S. (2020): "Regulation by shaming: Deterrence effects of publicizing violations of workplace safety and health laws," American Economic Review, 110, 1866–1904.

Kalb, G. and J. C. Van Ours (2014): "Reading to young children: A head-start in life?" Economics of Education Review, 40, 1–24.

Kaplow, L. (1992): "Rules versus standards: An economic analysis," Duke Law Journal, 42, 557.

Kawaguchi, D. (2011): "Actual age at school entry, educational outcomes, and earnings," Journal of the Japanese and International Economies, 25, 64–80.

Kirkeboen, L. J., E. Leuven, and M. Mogstad (2016): "Field of study, earnings, and self-selection," The Quarterly Journal of Economics, 131, 1057–1111.

Kitagawa, T. and A. Tetenov (2018): "Who should be treated? empirical welfare maximization methods for treatment choice," Econometrica, 86, 591–616.

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.

Kowalski, A. E. (2022): "How to examine external validity within an experiment," Journal of Economics & Management Strategy, 1–19.

——— (2023a): "Behaviour within a clinical trial and implications for mammography guidelines," The Review of economic studies, 90, 432–462.

——— (2023b): "Reconciling seemingly contradictory results from the Oregon health insurance experiment and the Massachusetts health reform," Review of Economics and Statistics, 105, 646–664.

Kraft, M. A. (2020): "Interpreting effect sizes of education interventions," Educational Researcher, 49, 241–253.

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.

Lockwood, B. B., H. Allcott, D. Taubinsky, and A. Sial (2024): "What drives demand for state-run lotteries? evidence and welfare implications," Review of Economic Studies, rdae086.

March, C. (2005): "Academic redshirting: Does withholding a child from school entrance for one year increase academic success," Issues in Educational Research, 15, 69–85.

Matsubayashi, T. and M. Ueda (2015): "Relative age in school and suicide among young individuals in Japan: a regression discontinuity approach," PloS One, 10.

McAdams, J. M. (2016): "The effect of school starting age policy on crime: Evidence from US microdata," Economics of Education Review, 54, 227–241.

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.

McEwan, P. J. and J. S. Shapiro (2008): "The benefits of delayed primary school enrollment discontinuity estimates using exact birth dates," Journal of Human Resources, 43, 1–29.

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 (2021): "The causal interpretation of two-stage least squares with multiple instrumental variables," American Economic Review, 111, 3663–3698.

Molnar, T. L. (2024): "Can Academic Redshirting Shrink the Education Gender Gap? Causal Evidence on Student Achievement and Mental Health," Working paper.

Mountjoy, J. (2022): "Community colleges and upward mobility," American Economic Review, 112, 2580–2630.

Mühlenweg, A., D. Blomeyer, H. Stichnoth, and M. Laucht (2012): "Effects of age at school entry (ASE) on the development of non-cognitive skills: Evidence from psychometric data," Economics of Education Review, 31, 68–76.

Mühlenweg, A. M. (2010): "Young and innocent: international evidence on age effects within grades on victimization in elementary school," Economics Letters, 109, 157–160.

National Center for Education Statistics (NCES) (2020): "Types of state and district requirements for kindergarten entrance and attendance," Tech. rep., State Education Practices.

Neilson, C. (2021): "Targeted vouchers, competition among schools and the academic achievement of poor students," Working Paper.

Oster, E. (2022): The Family Firm: A Data-driven Guide to Better Decision Making in the Early School Years, vol. 3, Penguin.

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.

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.

Randazzo, S. (2023): "To Increase Equity, School Districts Eliminate Honors Classes," The Wall Street Journal.

Reeves, R. (2022): Of boys and men: Why the modern male is struggling, why it matters, and what to do about it, Brookings Institution Press.

Rokkanen, M. A. (2015): "Exam schools, ability, and the effects of affirmative action: Latent factor extrapolation in the regression discontinuity design," .

Routon, P. and J. K. Walker (2020): "Older and wiser? Relative age and success in high school and college," SSRN Working Paper.

Roy, A. D. (1951): "Some thoughts on the distribution of earnings," Oxford Economic Papers, 3, 135–146.

Schanzenbach, D. W. and S. H. Larson (2017): "Is your child ready for kindergarten?" Education Next, 17.

Shapiro, A. (2023): "Over diagnosed or over looked? The effect of age at time of school entry on students receiving special education services," Exceptional Children, 89, 161–177.

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.

Smith, J. (2009): "Can regression discontinuity help answer an age-old question in education? The effect of age on elementary and secondary school achievement," The BE Journal of Economic Analysis & Policy, 9.

Stevenson, M. T. and P. J. L. Doleac (2022): "Algorithmic risk assessment in the hands of humans," Tech. rep., SSRN Working Paper.

Tan, P. L. (2017): "The impact of school entry laws on female education and teenage fertility," Journal of Population Economics, 30, 503–536.

UNICEF (2022): "Primary education," UNICEF Data.

——— (2024): "Early childhood education," UNICEF Data.

Vytlacil, E. (2002): "Independence, monotonicity, and latent index models: An equivalence result," Econometrica, 70, 331–341.

Wallace, J., P. Goldsmith-Pinkham, and J. L. Schwartz (2022): "Excess death rates for Republicans and Democrats during the COVID-19 pandemic," NBER Working Paper.

Walters, C. R. (2018): "The demand for effective charter schools," Journal of Political Economy, 126, 2179–2223.

Wooldridge, J. M. (2015): "Control function methods in applied econometrics," Journal of Human Resources, 50, 420–445.

Online Appendix

A Appendix Tables and Figures **55**

B Theoretical Appendix **69**

B.1 A Model of Self-Selection and Kindergarten Entry 69

B.2 Defining Efficiency and Equity 70

B.3 Test Score Comparisons with Self-Selection 73

B.4 Equity, Efficiency, and Estimators 75

C Data Appendix **79**

C.1 Michigan Data 79

C.2 ECLS-B Data 80

D Econometric Appendix **81**

D.1 RD Robustness 81

D.2 Extrapolation Between Cutoffs 81

D.3 Comparison to Other Approaches to RD Extrapolation 83

D.4 Testing Parallel Trends 84

D.5 Bounding Effects on Always Takers 87

A. Appendix Tables and Figures

Table A.1: Kindergarten Entry Policies Vary Immensely Across Jurisdictions

	Policy Before Cutoff: Require Starting	Require Starting (district exceptions)	Recommend Starting
Policy After Cutoff:			
Require Waiting	Egypt (9/30), France (1/1), Japan (4/1), Jordan (1/1), Kazakhstan (1/1), Saudi Arabia (8/15), Singapore (1/1), South Africa (1/1), Spain (1/1), Hawaii (7/31)	China (8/31), Czech Republic (9/1), United Arab Emirates (8/31 and 3/31)	Albania (8/31), Iran (9/21), Israel (1/1), Mexico (12/31), South Korea (1/1), New South Wales (Australia), Alabama (9/1), Alaska (9/1), Arkansas (8/1), Florida (9/1), Georgia (9/1), Idaho (9/1), Iowa (9/15), Kansas (8/31), Kentucky (8/1), Maine (10/15), Mississippi (9/1), Nevada (8/1), North Carolina(8/31), Oklahoma (9/1), South Carolina* (9/1), South Dakota* (9/1) Utah (9/1)
Require Waiting (district exceptions)	Brazil (1/1), Bulgaria (1/1), Hungary (8/31)	Austria (8/31), Belgium (1/1), Finland (1/1), Finland (1/1), Norway (1/1), Poland (1/1), Serbia (3/1), Western Australia (6/30)	England (9/1), California (9/1–after fifth birthday), Wyoming (8/1), Arizona (8/31), Delaware (8/31), Illinois (9/1), Louisiana (9/30), Maryland (9/1–through 10/15), Minnesota (9/1), Missouri (7/31), Montana (9/10), North Dakota (8/1), Tennessee (8/15), Texas (9/1), Virginia* (9/30), Washington (8/31), West Virginia (7/1), Wisconsin (9/1)
Recommend Waiting	Portugal (9/15), Italy* (1/1), Ontario Canada (9/1–through 12/31)	Denmark* (1/1–through 10/1), Turkey (1/1–through 4/1)	Morocco (3/31), Peru (3/31–through 7/31), Russia (3/1), Quebec Canada (9/30), Victoria Australia (1/1), Michigan (9/1–through 12/1), Connecticut (9/1–through 1/1), Indiana (8/1–through 10/1), Nebraska (7/31–through 10/15)
Locally Defined At Fifth Birthday	Germany, Colorado, Massachusetts, New Hampshire, New Jersey, New York, Ohio, Oregon, Pennsylvania Rhode Island, Vermont Netherlands, New Zealand, South Australia		

Note: This table presents the kindergarten entry policies of 95 jurisdictions including many countries, states, and provinces. Almost universally kindergarten entry policies are based on birthday cutoffs. This table reports what recommendation or requirement is given to a child with a birthday before or after the relevant cutoff. Policies are split into three categories. Strict requirements for which there is essentially no scope for opting in or opting out, requirements that allow selection at the school districts' discretion, and recommendations that allow selection at families' discretion. For each country, state, or province the relevant cutoff date (or dates) are included in the text. Jurisdictions denoted with an * require a waiver for self-selection around a recommendation. Most policies come from NCES (2020) and IEA (2021) with some additions and updates for 2024.

Table A.2: Subgroup Heterogeneity in self-selection Shares and Effects of Waiting

Panel A: Self-Selection Shares (at December 1)	All Students	Lower-Income	Higher-Income	Girls	Boys
At All Schools:					
Opt In	0.192	0.111	0.253	0.156	0.226
Opt Out	0.406	0.425	0.383	0.447	0.366
At Wealthiest Schools					
Opt In	0.288	0.163	0.309	0.237	0.338
Opt Out	0.394	0.458	0.374	0.439	0.346
At Middle-Income Schools					
Opt In	0.200	0.152	0.231	0.157	0.240
Opt Out	0.392	0.388	0.398	0.443	0.343
At Poorest Schools					
Opt In	0.095	0.082	0.133	0.079	0.111
Opt Out	0.433	0.441	0.388	0.458	0.407

Panel B: Effects of Waiting (on All Compliers)	All Students	Lower-Income	Higher-Income	Girls	Boys
At All Schools	0.227***	0.247***	0.248***	0.255***	0.197***
	(0.010)	(0.011)	(0.014)	(0.012)	(0.014)
At Wealthiest Schools	0.212***	0.224***	0.225***	0.234***	0.186***
	(0.019)	(0.034)	(0.021)	(0.023)	(0.029)
At Middle-Income Schools	0.183***	0.193***	0.212***	0.254***	0.110***
	(0.016)	(0.020)	(0.021)	(0.020)	(0.021)
At Poorest Schools	0.260***	0.260***	0.317***	0.268***	0.252***
	(0.013)	(0.013)	(0.026)	(0.016)	(0.018)
Subgroup Children within 60 days					
2002-2012 cohorts	342,642	151,554	191,088	168,914	173,728
2013† and 2014 cohorts	50,818	27,983	22,835	25,000	25,818

Note: This table reports the selection into and out of waiting and effects of waiting to enter kindergarten on third-grade math test scores (measured in standard deviations) for different subgroups. Standard errors allow for arbitrary variance-covariance structure within schools. Panel A reports self-selection shares. The share opting in to waiting at December 1 is the constant of the first-stage regression on children from the 2002-2012 cohorts. The share opting out of waiting at December 1 is the jump in the probability of waiting in the first stage regression for the 2013 and 2014 cohorts. Panel B reports effects of waiting for students in the 2002-2012 cohorts. Unless otherwise noted, all estimates use a sample of children within the 60-day bandwidth and subgroup who started kindergarten in Michigan public schools and took state math exams in third grade.
$^{***}p < 0.001$.

Table A.3: Covariates Are Fairly Stable Across Cutoffs

	October 2014	1-Nov 2013	1-Dec 2002-2012	1-Dec 2013-2014
Demographic:				
Lower-Income	-0.016	0.012	0.026***	-0.002
	(0.015)	(0.011)	(0.004)	(0.009)
Female	-0.003	0.002	0.004	-0.011
	(0.014)	(0.012)	(0.003)	(0.009)
White	0.018	0.002	0.019***	0.011
	(0.013)	(0.011)	(0.004)	(0.009)
Black	-0.004	-0.001	-0.014***	-0.001
	(0.012)	(0.009)	(0.003)	(0.008)
English Learner	-0.017*	-0.004	-0.003	-0.013*
	(0.008)	(0.007)	(0.002)	(0.005)
Neighborhood:				
Median Income	740	-570	180	-830
	(790)	(620)	(230)	(550)
Percent College Educated	0.003	-0.003	0.001	-0.008*
	(0.005)	(0.004)	(0.001)	(0.003)
School:				
Developmental K	0.006	0.005	-0.017**	-0.020*
	(0.013)	(0.010)	(0.006)	(0.009)
Lowest Income Tercile	0.003	0.006	-0.009*	0.010
	(0.014)	(0.011)	(0.004)	(0.009)
Middle Income Tercile	0.007	0.003	0.009*	-0.002
	(0.014)	(0.011)	(0.004)	(0.009)

Note: This table reports reduced-form regressions of the characteristics of children on either side of the four main cutoff dates. This is a placebo test to see whether it is likely that potential outcomes are continuous at the cutoffs. $*p < 0.05$, $**p < 0.01$, $***p < 0.001$.

Table A.4: Treatment Effects Are Constant Over Alternative Specifications

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
December 1, 2002-2012	0.244***	0.245***	0.227***	0.213***	0.238***	0.235***	0.231***	0.223***	0.255***
	(0.018)	(0.013)	(0.010)	(0.009)	(0.011)	(0.011)	(0.014)	(0.010)	(0.014)
Observations	84,177	169,854	342,642	521,061	342,642	342,642	38,643	84,177	342,642
November 1, 2013	0.422+	0.335*	0.333**	0.309**	0.347**	0.346**	0.323*	0.216*	0.330+
	(0.218)	(0.146)	(0.121)	(0.113)	(0.128)	(0.125)	(0.136)	(0.086)	(0.176)
Observations	7,535	15,066	23,689	31,775	23,689	23,689	3,669	7,535	23,689
October 1, 2014	0.598	0.416+	0.355+	0.408**	0.375+	0.366+	0.383+	0.239+	0.407
	(0.389)	(0.219)	(0.183)	(0.154)	(0.208)	(0.203)	(0.218)	(0.122)	(0.300)
Observations	8,042	15,853	31,112	39,761	31,112	31,112	3,805	8,042	31,112
December 1, 2013-2014	0.251**	0.244***	0.180***	0.187***	0.200***	0.192***	0.190**	0.169***	0.216**
	(0.085)	(0.058)	(0.048)	(0.045)	(0.052)	(0.050)	(0.062)	(0.039)	(0.074)
Observations	14,479	29,069	50,818	65,331	50,818	50,818	6,455	14,479	50,818
Bandwidth	15 day	30 day	60 day	90 day	60 day	60 day	7 day	15 day	60 day
Kernel	Uniform	Uniform	Uniform	Uniform	Triangular	Epanichokov	Uniform	Uniform	Uniform
Polynomial	Linear	Linear	Linear	Linear	Linear	Linear	Level	Level	Quadratic

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 2002-03 or 2015-16 school years; turning five within the bandwidth of the relevant cutoff; and taking state math exams in third grade. Overall, the first stage, reduced form, and fuzzy RD relationships are not sensitive to the regression specification. Column (1) reports the main specification from the paper. Columns (2), (3), and (4) 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 (5) and (6) show that the uniform kernel gives similar results as a triangular and Epanichokov kernels. Columns (7) through (9) 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.

Table A.5: Treatment Effects Are Fairly Constant With Controls

	All Families			Lower-Income Families			Higher-Income Families		
December 1, 2002-2012	0.227***	0.229***	0.227***	0.247***	0.229***	0.228***	0.248***	0.229***	0.225***
	(0.010)	(0.009)	(0.008)	(0.011)	(0.010)	(0.010)	(0.014)	(0.013)	(0.013)
November 1, 2013	0.333**	0.294**	0.263*	0.151	0.159	0.140	0.650*	0.685*	0.499+
	(0.121)	(0.109)	(0.108)	(0.115)	(0.110)	(0.110)	(0.284)	(0.286)	(0.302)
October 1, 2014	0.355+	0.440**	0.441**	0.350+	0.405*	0.374*	0.549+	0.505+	0.480
	(0.183)	(0.164)	(0.166)	(0.200)	(0.192)	(0.186)	(0.301)	(0.294)	(0.313)
December 1, 2013-2014	0.180***	0.186***	0.219***	0.220***	0.228***	0.252***	0.115+	0.126+	0.178*
	(0.048)	(0.043)	(0.044)	(0.056)	(0.054)	(0.056)	(0.067)	(0.065)	(0.070)
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 2002-03 or 2015-16 school years; turning five within 60 days of the birthday cutoff; and taking state math exams in third grade. Since all of the regressions in the paper are performed without controls, Table A.5 shows that the first stage, reduced form, and fuzzy RD relationships are not sensitive to the regression specification. The table focuses on the Lower-income Higher-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 Bachelor's or more), and an interaction of lower-income and sex. Including these controls does not alter the effects. Including (kindergarten) school fixed effects may reduce the eager complier LATE for higher-income families, but the differences are imprecise.

Table A.6: Treatment Effects Constant Dropping Possible Problems

	All Families			Lower-Income Families			Higher-Income Families		
December 1, 2002-2012	0.224***	0.231***	0.262***	0.244***	0.235***	0.264***	0.240***	0.132**	0.291***
	(0.011)	(0.034)	(0.013)	(0.012)	(0.039)	(0.014)	(0.015)	(0.050)	(0.018)
Observations	268,499	24,222	203,563	117,775	12,498	91,737	150,724	11,724	111,826
November 1, 2013	0.363**	0.427*	0.281*	0.166	0.014	0.142	0.731*	0.838*	0.434
	(0.134)	(0.182)	(0.142)	(0.123)	(0.173)	(0.141)	(0.335)	(0.373)	(0.297)
Observations	18,664	18,402	16,341	10,285	9,251	9,668	8,379	9,151	6,673
October 1, 2014	0.322	0.819**	0.339	0.377	0.620+	0.400	0.359	0.786+	0.370
	(0.204)	(0.269)	(0.212)	(0.231)	(0.318)	(0.249)	(0.310)	(0.410)	(0.322)
Observations	24,592	23,147	21,299	13,189	10,936	12,215	11,403	12,211	9,084
December 1, 2013-2014	0.137*	0.187***	0.260***	0.174**	0.147*	0.257***	0.088	0.134*	0.213*
	(0.055)	(0.050)	(0.060)	(0.064)	(0.058)	(0.066)	(0.076)	(0.067)	(0.098)
Observations	40,226	36,299	35,541	21,971	17,195	20,477	18,255	19,104	15,064
Drop Weekends	✓			✓			✓		
Drop GSRP		✓			✓			✓	
Drop DK Schools			✓			✓			✓

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 2002-03 or 2015-16 school years; turning five within 60 days of either cutoff; taking state math exams in third grade; and the additional restrictions for each column. Appendix D.1 explains the motivations for considering each of these restrictions.

Table A.7: Subgroup MTE Estimates

	Average Effect of Waiting	MTE Slope	Test of No Selection on Gains
All Children	0.261	-0.318	
	(0.030)	(0.202)	[p=0.125]
Lower-Income	0.241	-0.030	
	(0.024)	(0.183)	[p=0.915]
Higher-Income	0.377	-0.967	
	(0.067)	(0.384)	[p=0.003]
White	0.257	-0.447	
	(0.040)	(0.226)	[p=0.065]
Non-White	0.244	-0.063	
	(0.030)	(0.301)	[p=0.745]
Girls	0.296	-0.392	
	(0.033)	(0.260)	[p=0.105]
Boys	0.228	-0.286	
	(0.047)	(0.277)	[p=0.285]
At Poorest Schools	0.289	-0.220	
	(0.030)	(0.285)	[p=0.465]
At Middle-Income Schools	0.208	-0.230	
	(0.046)	(0.310)	[p=0.465]
At Wealthiest Schools	0.310	-0.592	
	(0.801)	(0.429)	[p=0.125]

Note: This table reports the estimates of the average and marginal effect of waiting to enter kindergarten on third-grade math test scores (measured in standard deviations). Estimates come from extrapolating RD effects following Section 4. Standard errors and hypothesis test are based off of a nonparametric block bootstrap by kindergarten school with 1,000 draws for effects on all children and by income and with 200 for all other subgroups. The sample is comprised of children within the 60-day bandwidth who started kindergarten in Michigan public schools in the 2002-03 to 2015-16 school years and took state math exams in third grade.

Table A.8: Eager Compliers Are Negatively Selected on Third-Grade Math Scores (2013 Cohort)

	Scores After Starting Kindergarten at Five		
	Eager Compliers $\mu_{0,ec}(r_{Nov})$	Reluctant Compliers $\mu_{0,rc}(r_{Nov})$	Test of No Selection in Levels $\mu_{0,ec}(r_{Nov}) = \mu_{0,rc}(r_{Nov})$
All Students	-0.428	0.002	$[p = 0.000]$
N= 23,689	(0.097)	(0.034)	
Higher-Income	-0.230	0.438	$[p = 0.006]$
N= 10,477	(0.225)	(0.040)	
Lower-Income	-0.455	-0.370	$[p = 0.473]$
N= 13,212	(0.089)	(0.039)	
Black	-0.850	-0.621	$[p = 0.293]$
N= 5,171	(0.170)	(0.055)	
White	-0.321	0.253	$[p = 0.000]$
N= 15,436	(0.122)	(0.037)	
Girls	-0.505	-0.050	$[p = 0.016]$
N= 11,587	(0.161)	(0.042)	
Boys	-0.365	0.062	$[p = 0.005]$
N= 12,083	(0.120)	(0.048)	

Note: This table compares the third-grade math scores of students from different unobserved groups. Expected complier outcomes are calculated by regression using the regression approach of Abdulkadiroğlu et al. (2018) from the results of Imbens and Rubin (1997) and Abadie (2002). The final column reports the p-values of the Bertanha and Imbens (2019) test of outcomes between groups. Both regressions are weighted liner regressions with 60-day bandwidth around each cutoff and rectangular kernel. The sample for the November cutoff come from 23,689 students who start kindergarten in Michigan public schools in the 2013-14 or 2014-15 school years, who turning five between September 1 and December 1, 2013, and who take state math exams in third grade.

Table A.9: Eager Compliers Are Negatively Selected on Third-Grade Math Scores (2014 Cohort)

	Scores After Starting Kindergarten at Five		
	Eager Compliers $\mu_{0,ec}(r_{Oct})$	Reluctant Compliers $\mu_{0,rc}(r_{Oct})$	Test of No Selection in Levels $\mu_{0,ec}(r_{Oct}) = \mu_{0,rc}(r_{Oct})$
All Students	-0.364	-0.047	$[p = 0.093]$
N = 31,112	(0.163)	(0.024)	
Higher-Income	-0.182	0.356	$[p = 0.065]$
N = 14,262	(0.262)	(0.030)	
Lower-Income	-0.563	-0.355	$[p = 0.296]$
N = 16,850	(0.177)	(0.027)	
Black	-1.091	-0.633	$[p = 0.393]$
N = 6,798	(0.486)	(0.040)	
White	-0.253	0.173	$[p = 0.033]$
N = 20,191	(0.171)	(0.027)	
Girls	-0.385	-0.102	$[p = 0.373]$
N = 15,340	(0.262)	(0.029)	
Boys	-0.355	0.017	$[p = 0.107]$
N = 15,750	(0.199)	(0.035)	

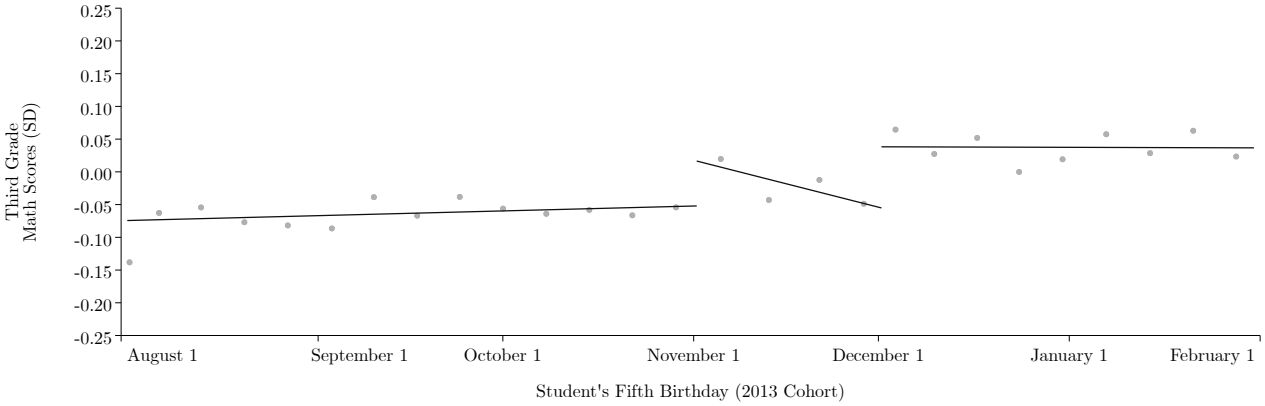
Note: This table compares the third-grade math scores of students from different unobserved groups. Expected complier outcomes are calculated by regression using the regression approach of Abdulkadiroğlu et al. (2018) from the results of Imbens and Rubin (1997) and Abadie (2002). The final column reports the p-values of the Bertanha and Imbens (2019) test of outcomes between groups. Both regressions are weighted liner regressions with 60-day bandwidth around each cutoff and rectangular kernel. The sample for the September cutoff come from 31,112 students who start kindergarten in Michigan public schools in the 2014-15 or 2015-16 school years, who turning five between August 1 and December 1, 2014, and who take state math exams in third grade.

Table A.10: Early Elementary School Outcomes Suggest that Always Takers Are Negatively Selected

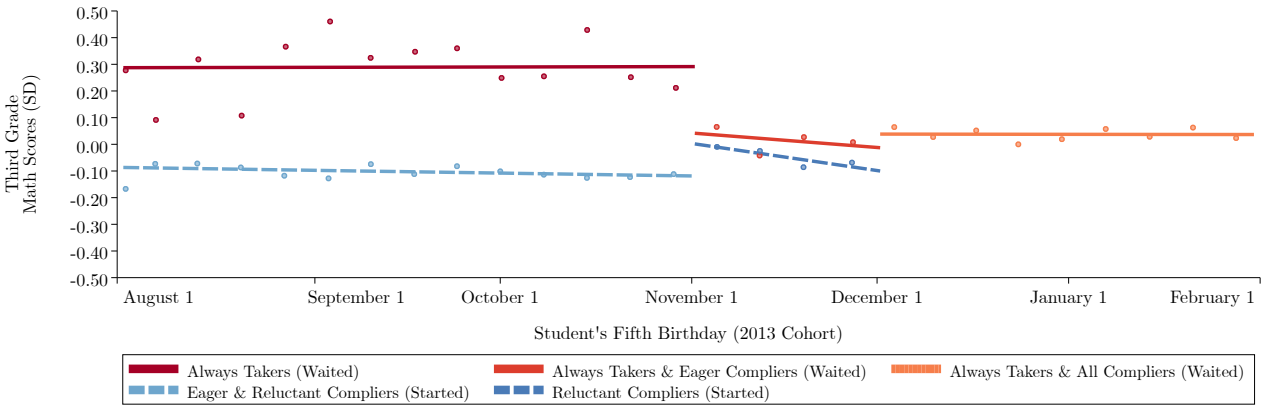
	Sample Mean	Always Takers	Compliers Wait	Compliers Start	Difference	LATE
Sample Shares		20.3%	77.3%			
Testing Outcomes:						
No Third Grade Math Test	0.110	0.154	0.083	0.116	0.071***	-0.033***
	(0.001)	(0.006)	(0.003)	(0.003)	(0.008)	(0.004)
Alternative Third Grade Math Test	0.008	0.013	0.006	0.006	0.008***	0.000
	(0.000)	(0.002)	(0.001)	(0.001)	(0.002)	(0.001)
Kindergarten Outcomes:						
Attendance Rate in Kindergarten	0.940	0.946	0.942	0.936	0.004***	0.006***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.002)	(0.001)
Enrolled in Kindergarten Twice	0.165	0.020	0.017	0.361	0.003	-0.344***
	(0.004)	(0.002)	(0.001)	(0.010)	(0.003)	(0.010)
Special Education in 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)
Elementary School Outcomes:						
Repeated Grade (Other than Kindergarten)	0.070	0.030	0.040	0.119	-0.010***	-0.080***
	(0.002)	(0.003)	(0.002)	(0.003)	(0.004)	(0.004)
Ever Skipped a Grade	0.008	0.012	0.008	0.005	0.004	0.004***
	(0.000)	(0.002)	(0.001)	(0.001)	(0.002)	(0.001)
Detailed Special Education Outcomes:						
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 and all compliers. The sample is comprised of students who meet the following criteria: enter kindergarten in Michigan public schools in the fall of 2002-2009 or 2011-2012; turn five within thirty days of December 1, 2013. 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.

Figure A.1: Average Test Scores Change at Each Birthday Cutoffs (2013 Cohort)



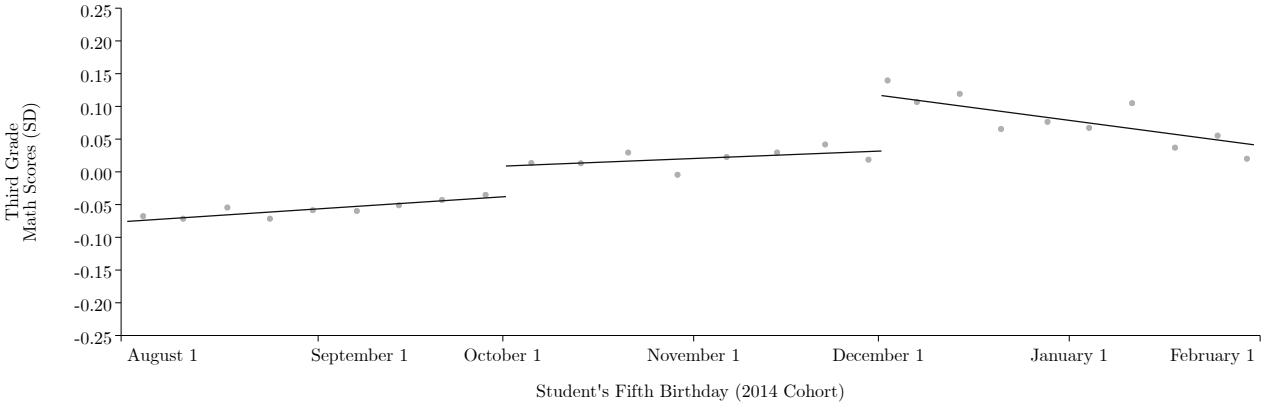
(a) Reduced Form



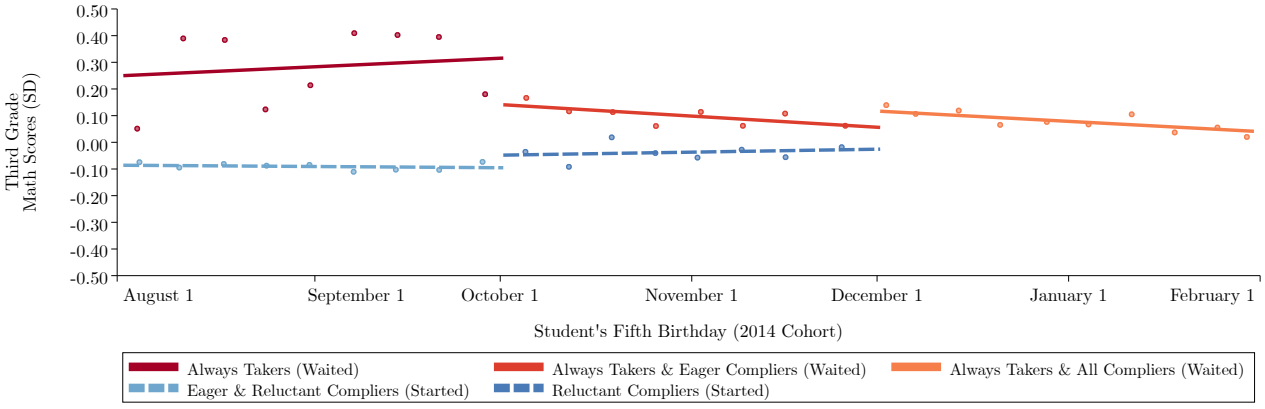
(b) Conditional Reduced Form

Note: This figure displays patterns in third-grade math achievement for students on facing different birthday-based recommendations or requirements. The third-grade test score measured in standard deviations is plotted over student birthdays. Reported scores are reported separately by birthday in seven-day averages with accompanying lines of best fit (with eight-day averages between November 1 and December 1). The sample is comprised of first-time kindergartners who turned five between August 1, 2013 and January 31, 2014 and for whom I observe third grade test scores.

Figure A.2: Average Test Scores Change at Each Birthday Cutoffs (2014 Cohort)



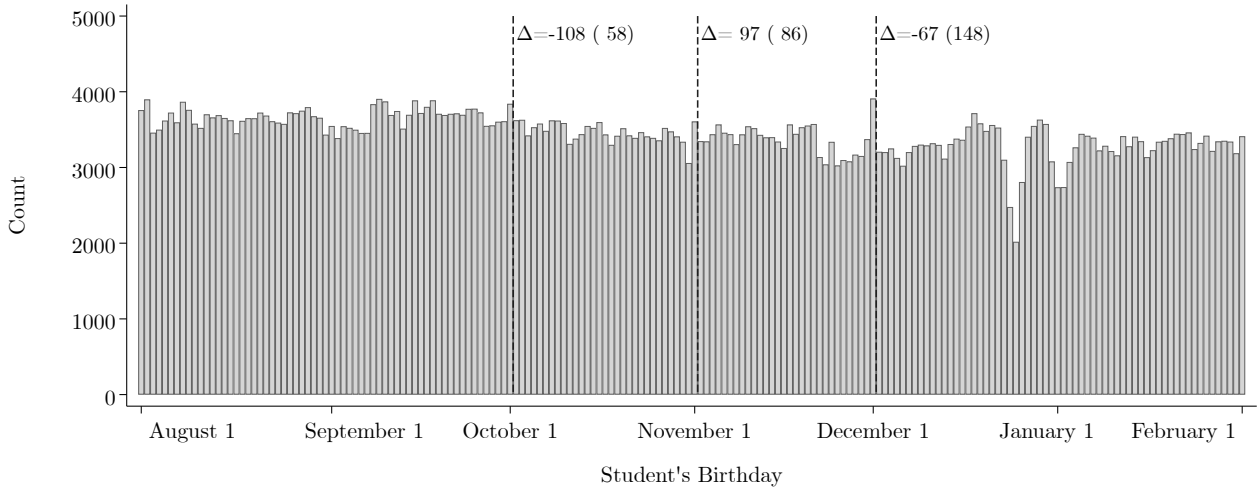
(a) Reduced Form



(b) Conditional Reduced Form

Note: This figure displays patterns in third-grade math achievement for students on facing different birthday-based recommendations or requirements. The third-grade test score measured in standard deviations is plotted over student birthdays. Reported scores are reported separately by birthday in seven-day averages with accompanying lines of best fit (with eight-day averages between October 1 and December 1). The sample is comprised of first-time kindergartners who turned five between August 1, 2014 and January 31, 2015 and for whom I observe third grade test scores.

Figure A.3: Birthdays Are Stable around the Cutoffs



Note: This figure reports the count of students with each birthday. The sample includes students who entered kindergarten in Michigan public schools from fall 2002-2015, turned five between August and February, and took state math exams in third grade. Changes in daily counts are reported using a McCrary test with a bandwidth of 30 days.

B. Theoretical Appendix

This appendix formalizes family decision-making, selection, and the implications for efficiency and equity. In the model the marginal treatment effects of waiting which together with the rates of self-selection characterize the welfare resulting from different policy recommendations or requirements.

B.1 A Model of Self-Selection and Kindergarten Entry

B.1.1 First Stage: Waiting to Start Kindergarten

Let each family i be characterized by their child's birthday, r, and other observable characteristics, x, and by their reluctance to wait, $U_W \sim [0,1]|r,x$, and other unobservables, γ. Families decide whether to wait $w \in \{0,1\}$ as influenced by birthday-based recommendations or requirements $f(r) = z$. A family will wait if the unobserved cost is less than a relevant threshold:

$$W = \mathbb{1}[U_W \leq p_z(r,x)] \qquad \text{where } p_z(r,x) = P(W=1|z,r,x) \tag{3}$$

Note that this threshold crossing assumption is equivalent to the LATE monotonicity assumption assuming "no-defiers" (see Imbens and Angrist (1994) and Vytlacil (2002)).

Figure 2 in the main text illustrates how families' innate reluctance to wait U_W characterizes the partial-compliance groups discussed in Section 2. It presents the unit interval for U_W and depicts whether or not always takers, eager compliers, and reluctant compliers will wait under different policy recommendations or requirements. Among families who are recommended to start ($z = 1$), only always takers opt in and wait, as such they are revealed to have the lowest reluctance to wait ($U_W \leq p_1(r,x)$). Facing the same incentives eager compliers and reluctant compliers do not wait because they are more reluctant ($U_W > p_1(r,x)$). In contrast, when recommended to wait ($z = 2$), the always takers and eager compliers both wait ($U_W \leq p_2(r,x)$) while reluctant compilers start; finally, everyone waits when required to wait.[43]

B.1.2 Second Stage: The Effects of Waiting

Assume that potential test scores, Y_w, depend on birthdays r, characteristics x, reluctance to wait U_W, and other unobservables $\boldsymbol{\gamma}$ but not on recommendations or requirements, z,: $Y_w = g_w(r,x,u,\gamma_w)$. Three functions from the marginal treatment effects (MTE) literature allow me to characterize the nature of self-selection. The first and third are the marginal treatment response (MTR) functions of Mogstad et al. (2018), and the second is the MTE function of

[43] Whether there are no true never takers is an empirical question. In the data there fewer than 100 students identified as never takers in each cohort (less than 0.1%). Although I drop these students for my analyses because some seem to reflect data or measurement error, including them does not change any results.

Heckman and Vytlacil (2005).

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

The MTR functions capture differences in average test scores over U_W if all children started kindergarten at age five (for m_0) or waited (for m_1). The first, m_0, is a generalization of selection bias (Kowalski, 2023b) that I use to conceptualize "selection in levels." Under positive selection in levels, children who are more likely to wait (lower U_W) have higher test scores than those who are less likely to wait, implying m_0 is decreasing in u. In the presence of negative selection in levels, m_0 would be increasing in u. The second, m_1, represents the sum of baseline differences in scores and differential gains from waiting, so its slope is less interpretable.

The MTE function τ_{MTE} captures differences in the effects of waiting among families with varying reluctance to wait. This is the function of Heckman and Vytlacil (2005) applied to my RD setting. I use this function to conceptualize "selection on gains." Under positive selection on gains, children who are more likely to wait (lower U_W) experience larger test score improvements than those who are less likely to wait, implying τ_{MTE} is decreasing in u. In the presence of negative selection on gains, τ_{MTE} would be increasing in u.

Because the MTR and MTE functions are not nonparametrically identified by birthday cutoffs, I also define aggregated statistics for always takers, eager compliers, and reluctant compliers, denoting the average potential outcomes for children from each group $g \in \mathcal{P}(\{at, ec, rc\})$:

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

These are weighted averages of the underlying MTR functions $m_w(u,r,x)$.[44] At any birthday r the average outcomes for two groups will be identified (those with $U_W \leq p_z(r,x)$ and those with $U_W > p_z(r,x)$). Differences between these means are the average marginal treatment effects for each group, which capture the LATEs identified by birthday cutoffs.

B.2 Defining Efficiency and Equity

I define the realized welfare from any allocation as $\mathcal{W} = \sum_i \phi_i U_i^c + \psi_i U_i^p$, where ϕ_i and ψ_i are welfare weights for the lifetime utility of children and parents from family i. Given this formulation, denote the expected welfare from allocation j given the observed behaviors and realized outcomes as $\widehat{\mathcal{W}}_j = \mathbb{E}[\mathcal{W}|\boldsymbol{Y}_j, \boldsymbol{w}_j]$. To make this tractable I make the following assumption:

[44] Specifically, $\mu_{w,g}(r,x) = \int_0^1 \omega_g(r,x) m_w(u,r,x)\,\mathrm{d}u$ 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.

Assumption 1. Assume (1) that there is no interdependence in test scores, waiting decisions, or utilities between families; (2) that third-grade math scores sufficiently characterize expected utility of children; and (3) that the social planner is indifferent between utility gains among different children and among different parents (i.e., $\phi_i = \phi$ and $\psi_i = \psi \; \forall$).

Because parental and child utility both feature in the welfare function, even under Assumption 1, the natural equilibrium may not be socially optimal. In this setting, the decision makers for each family (parents) typically pay the costs of waiting but do not experience the gains (children do). If the benefits to children are not fully internalized by parents, there will be under-provision of paternal investments. While allowing parents to choose which children start kindergarten (or only intervening with non-binding recommendations) will generate revealed preferred allocations, these allocations may not, in fact, be socially optimal. Instead I present a set of two criteria that together can (partially) order allocations to consider efficiency.

Proposition 1. Under Assumption 1, if allocation j is revealed preferred by families to allocation j' and allocation j also produces higher average test scores, then allocation j is expected to produce higher welfare than allocation j'.

Proof. Consider the difference between allocation j and j', then denoting $\Delta\mathcal{Y}$ as the difference average test scores, and $\Delta\hat{U}_i^p$ the difference in parental utility we have

$$\begin{aligned}
\widehat{\mathcal{W}}_j - \widehat{\mathcal{W}}_{j'} &= \mathbb{E}[\mathcal{W}|\boldsymbol{Y}_j, \boldsymbol{w}_j,] > \mathbb{E}[\mathcal{W}|\boldsymbol{Y}_{j'}, \boldsymbol{w}_{j'},] \\
&= \sum_{i=1}^{N} \mathbb{E}[\phi U_{j,i}^c | Y_{j,i}, w_{j,i}] + \mathbb{E}[\psi U_{j,i}^p | Y_{j,i}, w_{j,i}] - \sum_{i=1}^{N} \mathbb{E}[\phi U_{j',i}^c | Y_{j',i}, w_{j',i}] - \mathbb{E}[\psi U_{j',i}^p | Y_{j',i}, w_{j',i}] \\
&= \sum_{i=1}^{N} \phi(Y_{j,i} - Y_{j',i}) + \psi \left(\mathbb{E}[U_{j,i}^p | Y_{j,i}, w_{j,i}] \mathbb{E}[U_{j',i}^p | Y_{j',i}, w_{j',i}] \right) \\
&\propto \Delta\mathcal{Y} + \frac{\psi}{\phi} \frac{1}{N} \sum_{i=1}^{N} \Delta\hat{U}_i^p > 0
\end{aligned}$$

where the first equation is the definition of $\widehat{\mathcal{W}}_j$, the second comes from the no-inter-family-spillovers assumption and utilitarian welfare weights among parents and among children, and the third comes from the sufficiency of test scores. Because the expected welfare is ordinal, division by the welfare weights and number of families preserves the ordering in the fourth line.

If j is revealed preferred to j', then $\Delta\hat{U}_i^p > \forall i$; therefore, if $\Delta\mathcal{Y}$ is also positive, $\widehat{\mathcal{W}}_j$ must be greater than $\widehat{\mathcal{W}}_{j'}$. □

Proposition 1 gives a set of criteria that are sufficient for efficiency. Note that if both of these conditions are met, then the more efficient outcome will be preferred by a social planner with any relative weights on students and parents. That said, while an allocation that meets these criteria will be revealed preferred by all families, it may not be a Pareto improvement because some children may not prefer it. Finally note that because these criteria form a sufficient but not necessary condition for efficiency, they only partially order allocations whenever parents do not fully internalize the test-score benefits to their children.[45]

Whether Proposition 1 is informative depends on how plausible Assumption 1 is. Consider each component in turn. First, I assume that there are no spillovers across families in test scores, waiting decisions, or utility. This assumption rules out cases where parents compete for scarce positional rents by having their children wait to enter kindergarten. Section 4.4.3 presents direct evidence that the gains from waiting are not driven by relative age effects, as has been found in other settings (Cascio and Schanzenbach, 2016; Peña, 2017; Molnar, 2024). This assumption also assumes that families make decisions based on private information about their child not other families' entry decisions. The patterns in Figure 3 suggest that on the margin families to not respond to other families' decisions to wait. The Figure shows that when the cutoff date moves to earlier in the year, the share of students waiting before both cutoffs does not change despite the implicit changes in peer age compositions. Similar patterns have been found in other contexts (e.g., Cook and Kang, 2018).

Second, I assume that third grade test scores are welfare relevant. This assumption seems plausible given stark evidence of achievement gains on lifetime earnings (e.g. Chetty et al., 2014b); however, the evidence of earnings gains because of waiting is somewhat mixed.[46] That said, it is clear that waiting increases test scores, positive tracking, AP enrollment, and college persistence, suggesting that the achievement effects from waiting do capture welfare-relevant information for some children (see Ponzo and Scoppa, 2014; Hemelt and Rosen, 2016; Dhuey et al., 2019; Routon and Walker, 2020). Assuming test scores are sufficient for welfare rules out cases where children who waited regret forgone earnings (leading to overestimates of welfare gains from waiting) or where they experience life-long returns from other noncognitive gains from waiting beyond what was captured in test scores (leading to underestimates of welfare gains from waiting). Ruling these cases out seems plausible given the existing evidence on earnings trajectories (Black et al., 2011) and the relatively smaller effects from waiting on non-cognitive outcomes a in the long run (McCrary and Royer, 2011; Cook and Kang, 2016).

Possibly the least plausible part of Assumption 1 is that the social planner values gains

[45]For example, it could not compare an allocation that is revealed preferred by parents but generates low average scores with one that is not preferred but generates higher scores.

[46]For example, in Norway waiting slightly decreases earnings measured at a given age (Black et al., 2011). But effects seem to occur because of the timing of labor-force entry and the losses more than fade out by age 35, suggesting difference in trajectories.

to all children equally. Rather than taking a stand on which dimensions the social planner does or does not consider in the welfare weights, I address this problem by defining a separate criterion for equity. Because social planners and policy makers often care about the outcomes of different groups, a utilitarian concept of efficiency may be insufficient for discussing welfare. I define achievement-equity for students of type $X = x$ as the average differences in realized test scores $|\mathbb{E}[Y|X = x] - \mathbb{E}[Y|X \neq x]|$. Allocations can also be compared using this metric by measuring the change in achievement gaps in different allocations. 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 but does limit the extent of formal welfare consideration of groups.

B.3 Test Score Comparisons with Self-Selection

Given the criteria for equity and efficiency, it is now time to operationalize these criteria in terms of self-selection and treatment effects.

Definition 1. Under Assumption 1, for any policy function with a cutoff r^* such that $f(r) = z$ when $r < r^*$ and $f(r) = z'$ when $r \geq r^*$, the average test scores can be found by integrating the treated and untreated MTR functions over r and u (by x or in the full sample):

$$\begin{aligned}\mathcal{Y}(x) = \int_r \mathbb{E}[Y|X = x]\,\mathrm{d}F(r|x) &= \int_r P(W = 0|R = r, X = x)\mathbb{E}[Y|W = 0, R = r, X = x] \\ &\quad + P(W = 1|R = r, X = x)\mathbb{E}[Y|W = 1, R = r, X = x]\mathrm{d}F(r|x) \\ &= \int_{\underline{r}}^{r^*} \int_0^{p_z(r,x)} m_1(u, r, x)\,\mathrm{d}u + \int_{p_z(r,x)}^1 m_0(u, r, x)\,\mathrm{d}u\,\mathrm{d}F(r|x) \\ &\quad + \int_{r^*}^{\bar{r}} \int_0^{p_{z'}(r,x)} m_1(u, r, x)\,\mathrm{d}u + \int_{p_{z'}(r,x)}^1 m_0(u, r, x)\,\mathrm{d}u\,\mathrm{d}F(r|x)\end{aligned}$$

Consider a recommendation and requirement policy based off of a given birthday cutoff r^*:

$$f_{\text{rec}}(r) : \begin{cases} z = 1 \text{ if } r < r^* \\ z = 2 \text{ if } r \geq r^* \end{cases} \qquad f_{\text{req}}(r) : \begin{cases} z = 0 \text{ if } r < r^* \\ z = 3 \text{ if } r \geq r^* \end{cases}$$

We can compare the recommendation and requirement using the criteria for efficiency defined above. If the difference in average scores $\mathcal{Y}_{\text{rec}} - \mathcal{Y}_{\text{req}} \equiv \Delta\mathcal{Y}$ is positive, the allocation generated by the recommendation will be more efficient because families have (weak) revealed preference for allocations generated by recommendations. Similarly, any difference in achievement-equity can be calculated using the $\Delta\mathcal{Y}(x)$ change differences for each subgroup. These differences are a function of the self-selection shares and marginal treatment effects:

Proposition 2. Under Assumption 1, if allocation j is generated by a recommendation based

off of cutoff r^* and allocation j' is generated by a requirement at the same cutoff, then for any subgroup x (or the full sample)

$$\Delta\mathcal{Y}(x) = \underbrace{\int_{\underline{r}}^{r^*}\int_0^{p_1(r,x)} \tau_{MTE}(r,x,u)\,\mathrm{d}u\,\mathrm{d}F(r|x)}_{\textbf{Total Effect Always Takers}} - \underbrace{\int_{r^*}^{\bar{r}}\int_{p_2(r,x)}^{1} \tau_{MTE}(r,x,u)\,\mathrm{d}u\,\mathrm{d}F(r|x)}_{\textbf{Total Effect on Reluctant Compliers}}$$

Proof. Beginning with the difference of $\mathcal{Y}$ based on Definition 1,

$$\begin{aligned}
\Delta\mathcal{Y}(x) &= \int_{r}^{r^*}\int_0^{p_1(r,x)} m_1(u,r,x)\,\mathrm{d}u + \int_{p_1(r,x)}^{1} m_0(u,r,x)\,\mathrm{d}u\,\mathrm{d}F(r|x) \\
&+ \int_{r^*}^{\bar{r}}\int_0^{p_2(r,x)} m_1(u,r,x)\,\mathrm{d}u + \int_{p_2(r,x)}^{1} m_0(u,r,x)\,\mathrm{d}u\,\mathrm{d}F(r|x) \\
&- \int_{r}^{r^*}\int_0^{p_0(r,x)} m_1(u,r,x)\,\mathrm{d}u + \int_{p_0(r,x)}^{1} m_0(u,r,x)\,\mathrm{d}u\,\mathrm{d}F(r|x) \\
&+ \int_{r^*}^{\bar{r}}\int_0^{p_3(r,x)} m_1(u,r,x)\,\mathrm{d}u + \int_{p_3(r,x)}^{1} m_0(u,r,x)\,\mathrm{d}u\,\mathrm{d}F(r|x) \\
&= \int_{\underline{r}}^{r^*}\left[\int_0^{p_1(r,x)} m_1(u,r,x)\,\mathrm{d}u + \int_{p_1(r,x)}^{1} m_0(u,r,x)\,\mathrm{d}u\right] - \int_0^1 m_0(u,r,x)\,\mathrm{d}u\,\mathrm{d}F(r|x) \\
&+ \int_{r^*}^{\bar{r}}\left[\int_0^{p_2(r,x)} m_1(u,r,x)\,\mathrm{d}u + \int_{p_2(r,x)}^{1} m_0(u,r,x)\,\mathrm{d}u\right] - \int_0^1 m_1(u,r,x)\,\mathrm{d}u\,\mathrm{d}F(r|x) \\
&= \int_{\underline{r}}^{r^*}\left[\int_0^{p_1(r,x)} m_1(u,r,x)\,\mathrm{d}u\right] - \int_0^{p_1(r,x)} m_0(u,r,x)\,\mathrm{d}u\,\mathrm{d}F(r|x) \\
&+ \int_{r^*}^{\bar{r}}\left[\int_{p_2(r,x)}^{1} m_0(u,r,x)\,\mathrm{d}u\right] - \int_{p_2(r,x)}^{1} m_1(u,r,x)\,\mathrm{d}u\,\mathrm{d}F(r|x) \\
&= \int_{\underline{r}}^{r^*}\int_0^{p_z(r,x)} m_1(u,r,x) - m_0(u,r,x)\,\mathrm{d}u\,\mathrm{d}F(r|x) - \int_{r^*}^{\bar{r}}\int_{p_{z'}(r,x)}^{1} m_1(u,r,x) - m_0(u,r,x)\,\mathrm{d}u\,\mathrm{d}F(r|x) \\
&= \underbrace{\int_{\underline{r}}^{r^*}\int_0^{p_1(r,x)} \tau_{MTE}(r,x,u)\,\mathrm{d}u\,\mathrm{d}F(r|x)}_{\textbf{Total Effect Always Takers}} - \underbrace{\int_{r^*}^{\bar{r}}\int_{p_2(r,x)}^{1} \tau_{MTE}(r,x,u)\,\mathrm{d}u\,\mathrm{d}F(r|x)}_{\textbf{Total Effect on Reluctant Compliers}}
\end{aligned}$$

where the second equality comes from the fact that under a requirement $p_0(r) = 0$ and $p_3(r) = 1$, the third comes from noting that $\int_a^c f(u)\,\mathrm{d}u = \int_a^b f(u)\,\mathrm{d}u + \int_b^c f(u)\,\mathrm{d}u$ and canceling like terms, and the final equality is the definition of the marginal treatment effect from above.

□

Also note three immediate corollaries. First, the change in average test scores is thus $\Delta\mathcal{Y} = \sum_x \rho_x x\mathcal{Y}(x)$, where ρ_x are the population shares. Second, achievement-equity can be determined by the differences in $\Delta\mathcal{Y}(x=1) - \Delta\mathcal{Y}(x=0)$ by subgroup. Finally, other policy functions at a given r^* (such as a recommendation before but requirement after) can also be

compared if the self-selection shares and marginal treatment effects are known.

B.4 Equity, Efficiency, and Estimators

In practice treatment effects and self-selection shares are not known and must be estimated. Consider four estimation approaches.

1. **Traditional LATE.** Using one birthday cutoff as an instrument for the effect of waiting can identify the (subgroup-specific) self-selection shares but cannot identify the marginal treatment effects unless $\tau_{MTE}(r, x, u) = LATE(x)$ for all children.

2. **Population-Average MTE.** If there are more moments than the one birthday cutoff (or under stronger assumptions), parameterized versions of τ_{MTE} may be identified. Since we are concerned with average effects, it seems intuitive to use $\tau_{MTE}(r, x, u) = \tau_{MTE}(r, u)$ to consider average scores.

3. **Additively Separable.** Often τ_{MTE} is identified by assuming $\tau_{MTE}(r, x, u) = \tau_x(r) + \tau_{MTE}(u)$. Under this assumption the direction and magnitude of selection must be homogeneous (on average) for all subgroups. This approach will allow for differences in self-selection shares but not effects. This assumption is not implausible, but it is also not guaranteed, and it depends on whether heterogeneity loads on observed or unobserved dimensions.

4. **Subgroup-Specific MTE.** In a setting where there is enough information to identify τ_{MTE} without additive separability, subgroup-specific MTEs, $\tau_{MTE}(r, x, u)$ can also be identified. This would allow the consideration of efficiency and equity considerations and to directly test the plausibility of additive separability.

This subsection considers and compares the welfare implications of using these estimators.

Proposition 3. Under Assumption 1, using the population MTE, $\tau_{MTE}(u)$, or additively separable versions of the MTE will give inconsistent welfare estimates whenever there are differences in both self-selection shares and effects across groups. Furthermore, the biases in $\Delta\hat{\mathcal{Y}}$ at any r^* can be characterized as:

$$\mathcal{B}\text{ias}_{average} = \underbrace{\int_{p_1(r^*,1)}^{p_1(r^*)} \rho_1 \tau_{MTE}(r^*,1,u)\,\mathrm{d}u - \int_{p_1(r^*)}^{p_1(r^*,2)} \rho_2 \tau_{MTE}(r^*,2,u)\,\mathrm{d}u}_{\textbf{Error from Always Takers}}$$
$$+ \underbrace{\int_{p_2(r^*,1)}^{p_2(r^*)} \rho_1 \tau_{MTE}(r^*,1,u)\,\mathrm{d}u - \int_{p_2(r^*)}^{p_2(r^*,2)} \rho_2 \tau_{MTE}(r^*,2,u)\,\mathrm{d}u}_{\textbf{Error from Reluctant Compliers}}$$

$$\mathcal{B}\text{ias}_{separable} = \rho_1\rho_2 \Bigg[\underbrace{\int_{p_1(r^*,1)}^{p_1(r^*,2)} \tau_{MTE}(r^*,1,u) - \tau_{MTE}(r^*,2,u)\,\mathrm{d}u}_{\textbf{Differential Self-Selection among Always Takers}}$$
$$+ \underbrace{\int_{p_2(r^*,1,u)}^{p_2(r^*,2)} \tau_{MTE}(r^*,1) - \tau_{MTE}(r^*,2,u)\,\mathrm{d}u}_{\textbf{Differential Self-Selection among Reluctant Compliers}}$$
$$+ \underbrace{(\tau_{ATE}(r^*,1) - \tau_{ATE}(r^*,2)) \left[(p_1(r^*,2) - p_1(r^*,1) + (p_2(r^*,2) - p_2(r^*,1))\right]}_{\textbf{Selection Share Bias}} \Bigg]$$

Proof. Under Assumption 1 Proposition 2 characterizes the true $\Delta\mathcal{Y}$ across $x = 1$ and $x = 2$ at any r as

$$\Delta\mathcal{Y} = \rho_1 \int_0^{p_1(r,1)} \tau_{MTE}(r,1,u)\,\mathrm{d}u - \int_{p_2(r,1)}^{1} \tau_{MTE}(r,1,u)\,\mathrm{d}u$$
$$+ \rho_2 \int_0^{p_1(r,2)} \tau_{MTE}(r,2,u)\,\mathrm{d}u - \int_{p_2(r,2)}^{1} \tau_{MTE}(r,2,u)\,\mathrm{d}u$$

Similarly, we can write the approximations of $\Delta\mathcal{Y}$ using the average MTE ($\tau_{MTE}(r,x,u) = \tau_{MTE}(r,u)$ and an additively separable MTE ($\tau_{MTE}(r,x,u) = \tau_x(r) + \tau_{MTE}(u)$) as follows

$$\Delta\hat{\mathcal{Y}}_{average} = \int_0^{p_1(r)} \tau_{MTE}(r,u)\,\mathrm{d}u - \int_{p_2(r)}^{1} \tau_{MTE}(r,u)\,\mathrm{d}u$$

$$\Delta\hat{\mathcal{Y}}_{separable} = \int_0^{p_1(r,1)} \tau_1 + \tau_{MTE}(r,u)\,\mathrm{d}u - \int_{p_2(r,1)}^{1} \tau_1 + \tau_{MTE}(r,u)\,\mathrm{d}u$$
$$+ \int_0^{p_1(r,2)} \tau_2 + \tau_{MTE}(r,u)\,\mathrm{d}u - \int_{p_2(r,2)}^{1} \tau_2 + \tau_{MTE}(r,u)\,\mathrm{d}u$$

Let the bias $\mathcal{B}\text{ias}_j = \Delta\widehat{\mathcal{Y}}_j - \Delta\mathcal{Y}$. Choosing group 1 such that $p_1(r,1) \leq p_1(r) \leq p_1(r,2)$ (without loss of generality) consider the case when $p_2(r,1) \leq p_2(r) \leq p_2(r,2)$, then

$$
\begin{aligned}
\mathcal{B}\text{ias}_{average} &= \int_0^{p_1(r)} \tau_{MTE}(r,u)\,\mathrm{d}u - \int_{p_2(r)}^1 \tau_{MTE}(r,u)\,\mathrm{d}u \\
&- \rho_1 \int_0^{p_1(r,1)} \tau_{MTE}(r,1,u)\,\mathrm{d}u - \int_{p_2(r,1)}^1 \tau_{MTE}(r,1,u)\,\mathrm{d}u \\
&- \rho_2 \int_0^{p_1(r,2)} \tau_{MTE}(r,2,u)\,\mathrm{d}u - \int_{p_2(r,2)}^1 \tau_{MTE}(r,2,u)\,\mathrm{d}u \\
&= \int_0^{p_1(r,1)} \tau_{MTE}(r,u) - \rho_1\tau_{MTE}(r,1,u) - \rho_2\tau_{MTE}(r,2,u)\,\mathrm{d}u \\
&+ \int_{p_1(r,1)}^{p_1(r)} \tau_{MTE}(r,u) - \rho_2\tau_{MTE}(r,2,u)\,\mathrm{d}u - \int_{p_1(r)}^{p_1(r,2)} \rho_2\tau_{MTE}(r,2,u)\,\mathrm{d}u \\
&+ \int_{p_2(r,1)}^{p_2(r)} \rho_1\tau_{MTE}(r,1,u)\,\mathrm{d}u - \int_{p_2(r)}^{p_1(r,2)} \tau_{MTE}(r,u) - \rho_1\tau_{MTE}(r,1,u)\,\mathrm{d}u \\
&+ \int_{p_2(r,2)}^1 \tau_{MTE}(r,u) - \rho_1\tau_{MTE}(r,1,u) - \rho_2\tau_{MTE}(r,2,u)\,\mathrm{d}u \\
&= \int_{p_1(r^*,1)}^{p_1(r^*)} \rho_1\tau_{MTE}(r^*,1,u)\,\mathrm{d}u - \int_{p_1(r^*)}^{p_1(r^*,2)} \rho_2\tau_{MTE}(r^*,2,u)\,\mathrm{d}u \\
&+ \int_{p_2(r^*,1)}^{p_2(r^*)} \rho_1\tau_{MTE}(r^*,1,u)\,\mathrm{d}u - \int_{p_2(r^*)}^{p_2(r^*,2)} \rho_2\tau_{MTE}(r^*,2,u)\,\mathrm{d}
\end{aligned}
$$

Where the third equality follows from the fact that the population MTE is the average of the subgroup MTEs in the first, second, fifth, and sixth terms. The analytic expression is equivalent in the case when $p_2(r,1) \geq p_2(r) \geq p_2(r,2)$, but the signs on the reluctant complier terms would be reversed by inverting the integration bounds.

$$
\begin{aligned}
\mathcal{B}\text{ias}_{separable} &= \rho_1 \int_0^{p_1(r,1)} \tau_1 + \tau_{MTE}(u)\,\mathrm{d}u - \rho_1 \int_{p_2(r,1)}^1 \tau_1 + \tau_{MTE}(u)\,\mathrm{d}u \\
&+ \rho_2 \int_0^{p_1(r,2)} \tau_2 + \tau_{MTE}(u)\,\mathrm{d}u - \rho_2 \int_{p_2(r,2)}^1 \tau_2 \tau_{MTE}(u)\,\mathrm{d}u \\
&- \rho_1 \int_0^{p_1(r,1)} \tau_{MTE}(r,1,u)\,\mathrm{d}u - \rho_1 \int_{p_2(r,1)}^1 \tau_{MTE}(r,1,u)\,\mathrm{d}u \\
&- \rho_2 \int_0^{p_1(r,2)} \tau_{MTE}(r,2,u)\,\mathrm{d}u - \rho_2 \int_{p_2(r,2)}^1 \tau_{MTE}(r,2,u)\,\mathrm{d}u \\
&= \rho_2 \int_{p_1(r,1)}^{p_1(r,2)} \tau_2(r) + \rho_1(\tau_{MTE}(r,1,u) - \tau_{MTE}(r,2,u))\,\mathrm{d}u \\
&- \rho_1 \int_{p_2(r,1)}^{p_2(r,2)} \tau_1(r) + \rho_2(\tau_{MTE}(r,2,u) - \tau_{MTE}(r,1,u))\,\mathrm{d}u \\
&= \rho_2 \int_{p_1(r,1)}^{p_1(r,2)} \rho_1 \left[\tau_{ATE}(r,1) - \tau_{ATE}(r,2) + \tau_{MTE}(r,1,u) - \tau_{MTE}(r,2,u)\right] \mathrm{d}u \\
&- \rho_1 \int_{p_2(r,1)}^{p_2(r,2)} \rho_2 \left[\tau_{ATE}(r,2,u) - \tau_{ATE}(r,1) - \tau_{MTE}(r,2,u) - \tau_{MTE}(r,1,u)\right] \mathrm{d}u \\
&= \rho_1\rho_2 \Bigg[\int_{p_1(r,1)}^{p_1(r,2)} \tau_{MTE}(r,1,u) - \tau_{MTE}(r,2,u)\,\mathrm{d}u \\
&\qquad + \int_{p_2(r,1)}^{p_2(r^*,2)} \tau_{MTE}(r,1) - \tau_{MTE}(r,2,u)\,\mathrm{d}u \\
&\qquad + (\tau_{ATE}(r,1) - \tau_{ATE}(r,2)) \left[(p_1(r,2) - p_1(r,1) + (p_2(r,2) - p_2(r,1))\right] \Bigg]
\end{aligned}
$$

Where the second equality follows from the fact that the average of the two additively separable MTEs is equal to the average of the subgroup MTEs, and the third from fact that for the intercept terms, $\tau_x(r) = \tau_{ATE}(r) - \tau_{ATE}(r,x) = \rho_{x'}(\tau_{ATE}(r,x') - \tau_{ATE}(r,x))$—where $\tau_{ATE}(r,x) = \int_0^1 \tau_{MTE}(r,x,u)\,\mathrm{d}u$.

□

A corollary of Proposition 3 is that achievement equity faces a similar bias when estimated assuming additive separability. Whereas the efficiency results rely on weighted averages ($\Delta\mathcal{Y} = \sum_x \rho_x \Delta\mathcal{Y}(x)$, achievement equity focus on differences in levels. As a result, equity effects have an additional bias increasing in the difference $|\rho_x - \frac{1}{2}|$.

C. Data Appendix

C.1 Michigan Data

Data for this project largely come from the Michigan Education Data Center (MEDC) which houses administrative education data that is 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 (in the early years this corresponded to free or reduced price lunch status until SNAP or TANF recipient, homeless, migrant, or in foster care were included), Census block group, and—importantly for this paper—date of birth. They also include relevant scholastic variables like grade, school and district attended, assigned district, attendance rate, English language proficiency and language spoken at home, and detailed special education service receipt. Unfortunately, the grade variable has difficulty separating out developmental and traditional kindergarten in the majority of the sample period, for the most recent years it does indicate whether students were in a separate developmental kindergarten classroom.

I sample 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 turn five on March 1, 2013 November 1, 2013 and January 1, 2014 are all choosing between entering in 2013 and waiting to enter in 2014. This assumption implies that all students face one relevant cutoff. For example, the November 1 2013 cutoff is relevant to students who turn five between March 1, 2013 and February 28, 2014. If academic redshirting is so prolific that 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 it would violate this assumption. 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 any case, estimation focuses on students born in closer to the cutoff dates, where this concern is a non-issue.

I then restrict my sample. I keep students facing the December 1, 2002 cutoff through the October 1, 2014 cutoff (i.e., turning five between March 1 2002 and February 28 2015) because

these are the cutoffs for which I observe all the relevant students and scores and birthdates. 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 relevant cutoff. While these students are recorded as entering enter in a year that wouldn't be allowed based on their birthdate, it is unclear how much of this is due to true (and in many cases prohibited) choices, accommodations made for students moving into the state mid-year, or measurement error in birthdates. Of the children dropped, 69% are recorded as starting at five even though they turned five after December 1 in the early years of the sample (2002-2010).[47] This restriction leaves me with 1,283,200 students who enter kindergarten between 2002 and 2015.

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. I standardize the scaled math scores (by grade and year) among the full set of first-time kindergartners before dropping any students.

C.2 ECLS-B Data

The Early Childhood Longitudinal Survey-Birth Cohort (ECLS-B), is a nationally representative survey maintained by the National Center for Education Statistics. The survey follows a cohort of 10,700 children born in 2001 from birth through kindergarten completion. The survey waves aimed at collecting data at ages nine-months, two-years, four years (preschool), five years (kindergarten eligible), and six years (for children in kindergarten in 2007-2008). Each wave includes some standardized assessments and education information; parent, teacher, and caregiver surveys; month of birth; and state of residence.

I sample children with survey responses in all years and who attended public kindergarten in fall 2006 or 2007 (ages five or six). I characterize families' kindergarten entry decisions based on the list of state birthday cutoffs in Bassok and Reardon (2013), dropping children living in states with district-determined cutoffs. The main variables of interest include whether the child started kindergarten at five or at six, measures of kindergarten readiness at age four (parental perceptions, math and language skills, etc.), and measures of human capital investments (preschool attendance, preschool quality, reading, going to the library, etc.).

[47] Professionals at the Michigan Department of Education pointed out that in these early years, districts with part-day kindergarten programs could enroll some early entrants due to cross subsidization from other students—who still earned districts full-day-equivalent funding. This insight is backed up empirically. Between 2008 and 2011 the state transitioned to paying based on the portion of full-day instruction received, and the rates of un-reimbursed early entry quickly drop toward zero.

D. Econometric Appendix

D.1 RD Robustness

This section considers two sets of conditions that could violate the exclusion restriction that nothing besides the birthday cutoff policy is changing at the cutoff.

The first set of potential exclusion violations come from the distribution of families around the cutoff. The RD approach will not identify effects if the population is not comparable before and after the cutoff. Appendix Figures A.3 shows the density of birthdays is not statistically discontinuous at any cutoff, and Appendix Table A.3 shows that most and family characteristics are constant too. There are some statistically significant but economically small differences at one of the December 1 cutoffs. Students with birthdays after December 1 are slightly more likely to come from lower-income families, but they are also less likely to come from low-income schools. Including these variables as controls does not change the estimated effects, but this is an additional motivation for estimating the main effects separately by subgroup. Within a given year, estimates also be biased because fewer children are born on weekends (with a higher share from low-income families). Appendix Table A.6 show this is not a concern, dropping affected observations and obtaining similar results. Finally, the constant density of birthdays is also evidence that there is not selective attrition into private schools to avoid requirements to wait after December 1.

The second set of potential violations stem from differential childhood investments influenced by other policies. For examples, the state's public preschool option (GSRP) also uses a December 1 cutoff to determine eligibility. Bias could arise if take-up is discontinuous at any cutoff and GSRP affects third-grade math scores. Interestingly, I do not find jumps in take-up at December 1—likely because the GSRP does not allow re-enrollment for families who opt into waiting—but I do find slight differences at November 1. Bias could also arise if certain districts that use birthday cutoffs to require developmental kindergarten for children who opt out of waiting, generating extra investments.[48] There is no comprehensive record of whether districts had such requirements, but this would only be possible in schools that offer developmental kindergarten. Table A.6 shows that dropping GSRP participants or dropping students at schools with developmental kindergarten produces statistically indistinguishable effects.

D.2 Extrapolation Between Cutoffs

To comparing the marginal treatment effects of waiting on different groups, I leverage different birthday cutoffs that affect different types of compliers. Because estimates are local to set of compliers at the cutoff ($U_W \in [p_z(r^*, x), p_{z'}(r^*, x)]$ and $r = r^*$), differences between cutoffs could

[48] Developmental kindergarten programs are designed to give children two years of kindergarten before entering first grade, see (Berne et al., 2024).

be driven by true selection on gains or differences in composition. Identifying the patterns of self-selection requires extrapolation across cutoffs (or years).

To extrapolate, I adopt a "parallel trends" assumption and a "policy invariance" assumption. First, I assume that children's' average potential test scores ($\mathbb{E}[Y_0]$ and $\mathbb{E}[Y_1]$) evolve in parallel in r over some window $[\underline{r}, \bar{r}]$ around a cutoff r^*, conditional on x and u. Recalling the notation $m_0(u, r, x) = \mathbb{E}[Y_0|U_W = u, R = r, X = x]$ and $m_1(u, r, x) = \mathbb{E}[Y_1|U_W = u, R = r, X = x]$ I assume:

$$\begin{aligned} \forall\, \tilde{r} \in [\underline{r}, \bar{r}]: \quad & m_1(u, r^*, x) - m_1(u, \tilde{r}, x) = b_{1,u,x} \\ & m_0(u, r^*, x) - m_0(u, \tilde{r}, x) = b_{0,u,x} \end{aligned}$$

where $b_{1,u,x}$ and $b_{0,u,x}$ are level differences. This assumption is inspired by Cattaneo et al. (2020) to discipline the possible effect heterogeneity over r. Note that whereas I assume that treated and untreated outcomes both evolve in parallel for units with a given (u, x) (what Cattaneo et al. (2020) would call facing a given cutoff), Cattaneo et al. (2020) only assume that untreated outcomes evolve in parallel, but assume that the difference is the same for all u and x (units facing different cutoff) and that there are no always takers.[49] If a child happened to have a slightly earlier or later birthday, this assumption implies that the effect of waiting would not change systematically (even though both potential outcomes may). In the model, this assumption is equivalent with assuming that conditional on x and U_W the unobserved determinants of scores, $\boldsymbol{\gamma}$ are uniformly distributed within $[\underline{r}, \bar{r}]$.

Second, I assume that conditional on a student's birthday, the cutoff policy is unrelated to the expected treatment effect. Denoting students who face different cutoffs (whether different dates in one year or the same date but different years) c, I assume

$$\mathbb{E}[Y_1 - Y_0|r, u, x, C = c] = \mathbb{E}[Y_1 - Y_0|r, u, x]$$

This assumption is equivalent to the "policy invariance" assumption of Dong and Lewbel (2015) (see theorem in Cattaneo et al., 2016) and disciplines effect heterogeneity over time. One intuitive way to think of this assumption is that if a child had been born in a different year, the effect of waiting would not change systematically.

Together these assumptions imply that there is no appreciable heterogeneity in treatment effects over r or across cutoffs (conditional on U_W and x), i.e., $\tau_{MTE}(u, r^*, x) = \tau_{MTE}(u, \tilde{r}, x) \;\; \forall \tilde{r} \in$

[49] To make the comparison concrete, in 2014 the October 1 recommendation is the relevant cutoff for eager compliers, but the December 1 requirement is relevant to reluctant compliers. In my model which cutoff a student "faces" is a function of (u, x), but I only need parallel trends within each group. This means the average need not be parallel if there are compositional changes over r (for example,in the share of lower-income families across quarters of birth).

$[\underline{r}, \bar{r}]$. This simplification enables economically meaningful comparisons between effects estimated at different cutoffs. Appendix D.3 compares my approach to other approaches for extrapolation and Appendix D.4 presents an empirical test for whether these assumptions are likely to hold in my context leveraging the longitudinal nature of the Michigan data. Section 4 directly tests both of these assumptions using the over-identification provided by parameterizing the MTE curve.

D.3 Comparison to Other Approaches to RD Extrapolation

The assumptions in Appendix D.2 permit a unique approach to extrapolation in an RD based in the estimation of heterogeneity. The intuition is that if the economically meaningful heterogeneity occurs over U_W and x, I can use that information extrapolate effects over the running variable 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.

- **Ignorability.** My parallel trends assumption implies that x and u are the main drivers of heterogeneity, without the strong assumption that r is ignorable conditional on certain characteristics—observed or latent (Angrist and Rokkanen, 2015; Rokkanen, 2015). This approach to extrapolation requires that changes in effects over r only operate through compositional changes in x (or the predicted latent type). Assuming r cannot directly affect outcomes, however, may only be plausible in cases with rich x and predictable r like test score cutoffs. There are many common RD setting, however, where r is not (conditionally) excluded such as close elections, income thresholds, or birthdays. In a fuzzy RD, this approach also requires that take-up is also not directly affected by r (i.e., $p(x, r) = p(x)$), which may also often be implausible. My approach also assumes that individual characteristics are the main driver of heterogeneity; however, the parallel trends assumption allows both the probability of treatment and potential outcomes to vary over r (even conditional on x and u)—similar to how identification in a difference in differences may be more plausible than under ignorability.

- **Multiple Cutoffs.** Using information about self-selection into treatment enables extrapolation with multiple cutoffs in fuzzy RD settings. When subpopulations face different cutoffs, a parallel trends assumption on untreated outcomes similar to mine enables extrapolation between cutoffs (see Cattaneo et al., 2020).[50] 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

[50] Including treated outcomes in the assumption as I do would allow for extrapolation beyond the cutoffs.

by making assumptions about treated outcomes, I can extrapolate in fuzzy RD settings with two-sided noncompliance, can extrapolate beyond the first and last cutoffs in r, and can still extrapolate when the relevant cutoff to each unit is unobserved.

- **Using LATEs.** My approach is also similar in concept to those that use the LATE to extrapolate, but I estimate richer heterogeneity. One method to "extrapolate" using the LATE is to assume the LATE is externally valid—no selection and no effect heterogeneity (Bertanha and Imbens, 2019).[51] If r is the main driver of heterogeneity, extrapolation is easy: the effect is $\mathbb{E}[Y|W=1,r]-\mathbb{E}[Y|W=0,r]$ for any r. Unfortunately, this external validity assumption often may not hold. In this case, changes in the LATE at the cutoff can be extrapolated by the mean value theorem (Dong and Lewbel, 2015).[52] My approach uses the LATE for extrapolation, but rather than assume homogeneity, I directly estimate heterogeneity. The limitation of my approach is the need for multiple cutoffs or additional ancillary assumptions (as in Mogstad et al., 2018) to identify this heterogeneity.

Relative to these 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 to compliers, and then those parameters can be the basis of extrapolation away from r^* because of 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 policy incentives or the placement of the cutoff itself.

D.4 Testing Parallel Trends

Conveniently, the assumption necessary to identify and interpret these effects have a testable implication made possible by the longitudinal assessment data. Under parallel trends and policy invariance, the slope of test scores over birthdays is the same for students had they started or waited. Although I 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), I do observe their *realized* 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., after December 1), I can determine how close to

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

[52] 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), but it could be used to extrapolate.

parallel the trends are. Failure to reject does not imply parallel trends in potential outcomes but does build credibility.

To estimate whether there are differences in the slopes of test scores, I test the 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, 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 eighth grade test scores).

Table D.1 demonstrates that the trends are close to parallel. The first four rows of Table D.1 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 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

Table D.1: 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. $^{+}p < 0.1$, $*p < 0.05$, $**p < 0.01$, $***p < 0.001$.

the largest plausible deviation from parallel trends.

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.[53] 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.

[53]This is assuming a negative relationship as would be expected over birthdays. 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.

D.5 Bounding Effects on Always Takers

The effects on always takers are not identified by the RD because they always wait. That said, knowing more about the average counterfactual scores of always takers would allow me to estimate effects of waiting affects them. One alternative to data on these counterfactual outcomes, is making additional assumptions. Mogstad et al. (2018) proposes assuming that untreated outcomes are monotonic, i.e., $m_0(x,r,u) = \mathbb{E}[Y_0|x,r,u] < m_0(x,r,u') = \mathbb{E}[Y_0|x,r,u'] \; \forall u > u'$.[54] Assuming monotonicity does not point identify untreated outcomes, but it provides a bound which can be used to bound the effects on always takers.

Figure D.1 graphically illustrates how this assumption can bound the effects on always takers with Panel (a) for the 2013 cohort and Panel (b) for the 2014 cohort. Each panel reports the average outcomes of always takers, treated eager compliers, untreated eager compliers, and untreated reluctant compliers. The difference between the untreated outcomes of eager and reluctant compliers is the direction of selection on levels and of the assumed monotonicity. Under this assumption the average untreated outcomes of always takers must be at least as low as those of eager compliers, depicted in the dashed line. This level in turn bounds the possible effect of waiting on always takers ($\tau_a t \geq 0.66$ in 2013 and $\tau_a t \geq 0.65$ in 2014).[55] The figure also depicts the treated and untreated outcomes of reluctant compliers at December 1 to visualize the size of the different treatment effects.

To test the null of effect homogeneity between always takers and eager compliers under monotonicity, I use a test for heterogeneity from in Kowalski (2023a). This test determines whether the eager-complier LATE, τ_{ec}, falls within the range bounded by monotonicity. The one-tailed test statistic returns a 1 if it rejects treatment effect homogeneity (i.e., whether τ_{ec} is outside of the bounds from $\tilde{\tau}_{at}$) and a 0 otherwise:

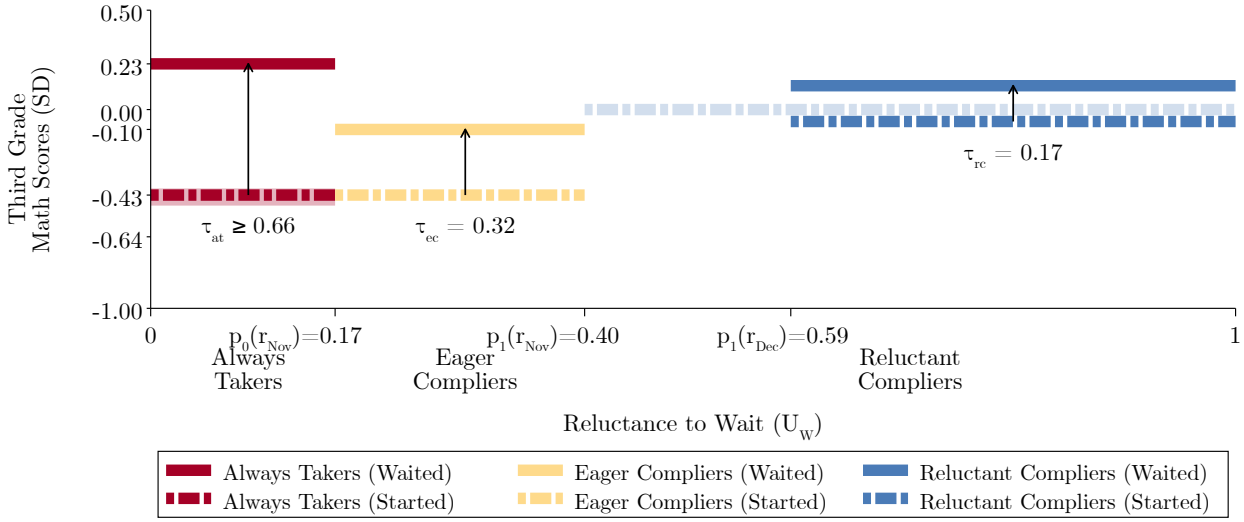
$$\mathbb{1}\left[(\tilde{\tau}_{at} - \tau_{ec}) \cdot \mathcal{B}_{0,ec,rc} < 0\right]$$

For inference I use a 200 replications of a nonparametric block bootstrap, blocking by school. I reject homogeneity at the $p = 0.000$ level in 2013 and $p = 0.02$ level in 2014.

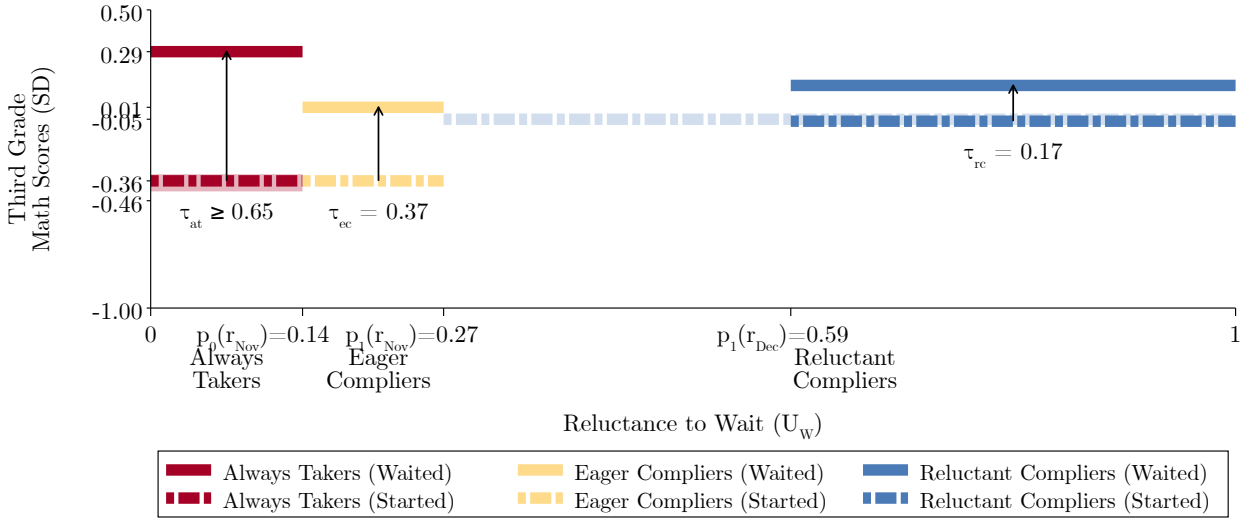
[54] Technically monotonicity could be positive or negative, but the data reveal negative selection in levels between eager and reluctant compliers (see Section 4.4.2).

[55] Note that it is not guaranteed that the bound implied by a monotonicity assumption is informative (see discussion in Kowalski, 2022), but negative selection in levels does provide information in this case.

Figure D.1: Always Takers Gain Much More from Waiting than Compliers



(a) Selection on Gains in 2013



(b) Selection on Gains in 2014

Note: This figure illustrates the selection on gains between always takers and compliers. It presents the effects on each complier group (comparable under the parallel trends assumption) and the bound on the effects on always takers under monotonic selection in levels for both 2013 and 2014.